UNICA

UNIVERSITÀ
DEGLI STUDI
DI CAGLIARI

Ph.D. DEGREE IN
ELECTRONIC AND COMPUTER ENGINEERING

Cycle XXXVIII

TITLE OF THE Ph.D. THESIS

Learning Based Objective QoE Models Across Interactive and Immersive Media

Scientific Disciplinary Sector(s)

ING-INF/03

Ph.D. Student	MohammadAli Hamidi
Supervisor	Prof. Luigi Atzori
Co-Supervisor	Dr. Alessandro Floris

Final exam. Academic Year 2024/2025
Thesis defence session: February 2026

UNIVERSITÀ DEGLI STUDI
DI CAGLIARI

HTTPS://EN.UNICA.IT/EN

First edition, December 17, 2025

Acknowledgements

I would first like to express my deepest gratitude to my family, whose unconditional love, patience, and encouragement have been my greatest source of strength throughout this journey. Their belief in me made every challenge possible to overcome.

My special thanks go to Parisa, for her endless support, kindness, and inspiration. Her presence and understanding have brought balance, motivation, and meaning to every step of this path.

I am sincerely grateful to my supervisor, Prof. Luigi Atzori, for his invaluable guidance, insight, and trust. His scientific rigor, vision, and encouragement have profoundly shaped both my research and my academic mindset.

I also extend my appreciation to my co-supervisor, Dr. Alessandro Floris, and to Dr. Simone Porcu for their continuous support, insightful discussions, and constructive collaboration during my Ph.D and throughout this work.

My warm thanks go to Dr. Hadi Amirpour, whose mentorship, collaboration, and genuine enthusiasm for research have been both inspiring and motivating. Working with him has been an important and memorable part of this Ph.D. journey.

Abstract

Delivering high-quality multimedia experiences at scale requires objective Quality of Experience (QoE) models that reliably approximate human perception across heterogeneous contents, devices, and network conditions. Subjective user studies remain the reference standard for assessing perceived quality, but they are costly, time-consuming, and impractical for continuous or large-scale monitoring. Consequently, service providers and system designers increasingly rely on objective, learning-based models to predict and manage QoE in real time. Yet, existing approaches often fail to generalize across content types, devices, and operating conditions, leaving a persistent gap between measurable system parameters and what users actually experience.

This dissertation addresses the general problem of defining, training, and deploying learning-based QoE models. It develops a two-layer conceptual framework encompassing both service-level QoE, which models user experience as a function of network and application behavior, and content-level QoE, which assesses the perceptual quality of visual media itself. Across these complementary layers, the thesis demonstrates how data-driven methods can extract task-relevant features, learn perceptual dependencies, and achieve resource-efficient deployment. Together, these studies outline a methodological paradigm for learning-based QoE estimation applicable to interactive and immersive media services across diverse multimedia contexts, including conversational, streaming, immersive, and biometric applications. The proposed methodology demonstrates how QoE modeling principles can be adapted to different media types, learning paradigms, and deployment settings. Each study corresponds to a peer-reviewed publication, collectively forming a systematic progression of research that advances toward a consistent methodology for perception-aware QoE prediction.

The proposed framework is built around five pillars: (i) signal design, identifying task-relevant cues from application, network, and content while removing noise and redundancy; (ii) temporal modeling, capturing sequential dependencies and dynamic effects that shape perceived quality over time; (iii) multi-view fusion, enabling learning from integrating heterogeneous or partially shared data sources; (iv) multi-projection fusion, enabling learning from complementary feature spaces including multi-projection representations for 3D volumetric media; and (v) computational efficiency, ensuring the resulting models are lightweight, scalable, and suitable for real-time or edge deployment.

Using this framework, the dissertation develops several specialized models: an application-telemetry-driven QoE predictor for real-time audiovisual Web Real-Time Communication (WebRTC) conversations; a transformer-based estimator for adaptive video streaming that encodes start-up delays, quality switches, and stalls; a collaborative multi-view learning approach that supports privacy-preserving QoE modeling under distributed data constraints; and a no-reference point cloud quality assessment (NR-PCQA) model that employs multi-projection features and adaptive view weighting to evaluate volumetric content without pristine references. A final case study on face image quality assessment (FIQA) demonstrates the adaptability of these design principles, highlighting the generality and deployability of the framework.

Across these chapters, the models show strong correlation with subjective judgments, robustness to variations in content and devices, and favorable computational profiles for real-time

operation. They establish consistent evaluation protocols and feature analysis procedures that promote interpretability, reproducibility, and cross-context reliability. Collectively, the contributions provide a general recipe for objective QoE modeling to design the right signals, model temporal dependencies, fuse complementary and multi-projection views, and optimize for efficiency validated across diverse datasets and multimedia modalities.

Overall, this dissertation bridges the gap between QoE theory and practical implementation, offering a comprehensive, data-driven framework for predicting human-perceived quality across interactive, streaming, and immersive media.

Contents

List of Figures

List of Tables

1

Introduction

1.1 | Motivation

The rapid growth of real-time communication, adaptive video streaming, and emerging volumetric/Extended Reality (XR) media has intensified the need for objective predictors of human-perceived quality that are accurate, robust, and deployable [1]. Subjective tests remain the gold standard but are costly, slow, and difficult to reproduce at scale. Service operators therefore rely on objective, data-driven models to monitor and optimize quality in the wild; yet, existing models often struggle to generalize when content, device, user context, or network conditions change. This thesis is motivated by the gap between what can be measured automatically and what users actually perceive in diverse, evolving multimedia ecosystems.

A first driver is heterogeneity and domain shift [2]. Modern services span heterogeneous contents (natural scenes, screens, videos, point clouds), devices (mobile/desktop/VR), codecs, and transport conditions. Models trained in one setting frequently degrade in another due to distribution shifts and dataset bias. A practical methodology must therefore emphasize signal design (choosing the right indicators), regularization and fusion across different views, and validation protocols that test cross-context robustness rather than single-dataset accuracy alone.

A second driver is temporal dynamics. Perception is not a static function of independent frames or aggregates; therefore, recency, duration, and order of events (e.g., start-up delay, quality switches, stalls, jitter) shape how users judge overall quality. Conversational Quality of Experience (QoE) depends on interactive latency and smoothness, while streaming QoE depends on event sequences across the session. This motivates sequence-aware models that encode long-range dependencies and attention to salient events, moving beyond snapshot predictions.

A third driver is data fragmentation and privacy. Valuable signals are often distributed across organizations, applications, or devices, and in many cases cannot be pooled due to business sensitivity or privacy regulations (e.g., GDPR). Classical centralized training is unrealistic; instead, methods

must leverage complementary feature views and partial datasets while avoiding raw-data sharing. This motivates multi-view learning that fuses information at the representation level and supports collaborative modeling under real constraints.

A fourth driver is the rise of new media types where reference signals are scarce or unavailable. Point clouds and volumetric content introduce geometric sparsity, view dependence, and rendering pipelines that differ fundamentally from 2D video. Many operational scenarios require a no-reference (NR) assessment that remains aligned with human judgments. This motivates multi-projection feature extraction and adaptive view weighting that capture perceptually relevant cues without needing pristine references.

Finally, a fifth driver is deployment efficiency [3]. Real-time and edge settings impose strict latency, memory, and power budgets. Models must be compact and computationally tractable without sacrificing fidelity, and they should expose clear trade-offs between accuracy and efficiency. This motivates architectures and fusion strategies that are lightweight by design, with explicit reporting of Floating Point Operations (FLOPs), memory, and inference time alongside accuracy metrics.

Accordingly, this dissertation conceptualizes QoE modeling as a two-layer problem. The service-level layer focuses on how network and application behaviors, such as delay, jitter, and adaptation dynamics, influence the user's conversational and streaming experience. The content-level layer, by contrast, targets the perceptual quality of visual and immersive media itself, including 2D video, 3D point clouds, and biometric imagery. Together, these layers form a unified perspective that connects system behavior with perceptual outcomes, enabling the design of scalable and adaptive QoE models applicable across heterogeneous multimedia contexts.

Taken together, these considerations motivate a thesis that formulates a two-layer conceptual, learning-based framework for objective modeling of perceived quality, and then instantiates and validates the framework across complementary service and content-level domains and use cases, each corresponding to a dedicated chapter.

1.2 | Research Challenges

Advancing objective models of human perceived quality requires understanding both service-level factors that shape user experience and content-level factors that influence perceptual quality, which is a lined with bridging a persistent gap between what systems can measure automatically and what users actually experience. Multimedia ecosystems are inherently heterogeneous; contents range from natural scenes and screen recordings to images and 3D point clouds; delivery spans mobile and desktop devices, variable codecs, and fluctuating networks. Models that perform well in one regime often degrade in another because of distribution shift and dataset bias. A central challenge, therefore, is to design modeling pipelines that remain reliable under such cross-domain variation rather than

optimizing narrowly for a single dataset or operating point.

Perception in communication and streaming is also fundamentally temporal. Start-up delays, quality switches, stalls, jitter, and conversational latencies do not affect users independently; their order, duration, and recency shape the overall judgment of quality. Static or aggregate features obscure these dependencies, leading to predictors that miss key perceptual phenomena (e.g., "stall memory" effects or the asymmetry between early and late impairments). Capturing long-range structure and event salience is thus a prerequisite for precise estimation of perceived quality in sequential services.

At the same time, observability is imperfect. Application-layer telemetry, such as Web Real-Time Communication (WebRTC) logs, contains rich but noisy, partially redundant indicators. Without principled screening and ablation, models risk overfitting to imprecise correlations and becoming fragile in deployment. The methodological challenge is to extract compact, stable signal sets that retain explanatory strength while remaining interpretable enough to guide engineering decisions.

Data governance introduces a further constraint. Useful signals are frequently fragmented across organizations, entities, studies, or dataset features, and privacy or business rules prevent raw-data pooling. Classical centralized training is often infeasible. The challenge is to recover the benefits of comprehensive information under partial observability, enabling collaborative learning that respects privacy constraints while still exploiting complementarity between different views.

Emerging content-level modalities, such as point clouds and volumetric video, intensify these issues. Volumetric and point cloud content exhibit geometric sparsity, view dependence, and rendering artifacts that differ from 2D imagery. In many operational scenarios, pristine references are unavailable, so models must infer perceptual quality without reference signals. Designing NR assessors that remain aligned with human judgments and that can adapt multiple projections or viewpoints derived by a 3D object through adaptive fusion and integration poses both algorithmic and computational challenges.

Finally, deployment imposes hard resource constraints. Real-time and edge settings demand predictable latency and a low memory footprint. Highly accurate but heavy models are unusable in practice; the challenge is to make accuracy-efficiency trade-offs explicit and to deliver architectures whose runtime profiles match system budgets.

Taken together, these challenges motivate a conceptual, learning-based QoE framework that (i) selects and stabilizes task-relevant signals under heterogeneous conditions, (ii) encodes temporal structure where perception is sequential, (iii) enables learning under fragmented and privacy-constrained data, (iv) supports NR assessment for emerging modalities, and (v) remains efficient and deployable (FIQA use case). Subsequent chapters instantiate this framework in concrete use cases, each addressing different aspects of these challenges, from service-level QoE in conversational and streaming systems to content-level QoE in immersive and perceptual quality applications, quantifying its effectiveness and generality across diverse multimedia contexts.

1.3 | Research Questions

Based on the challenges outlined overhead, this doctoral research is guided by a set of core questions. The research questions operationalize the thesis aims into testable inquiries that cover predictive validity, generalization across domains and devices, privacy-aware learning with partial information, and deployability under resource constraints.

Each question is grounded in one or more pillars of the proposed conceptual framework, signal design, temporal modeling, multi-view fusion, multi-projection fusion, and computational efficiency, and corresponds to a concrete modeling strategy validated through controlled experiments, ablation analyses, and cross-context evaluations. Performance is reported with rank and linear correlations (Spearman Correlation Coefficient (SCC)/ Pearson Correlation Coefficient (PCC)), error metric (RMSE), and, where applicable, R^2, alongside statistical significance, calibration, and resource profiles (memory, FLOPs). Together, these research questions form a coherent progression from service-level QoE modeling (interactive and streaming systems) to content-level QoE modeling (3D and perceptual imagery), ultimately converging in a methodological paradigm for perceptual quality prediction.

RQ1. (Signals that matter - WebRTC) To what extent can carefully curated application-layer telemetry from WebRTC predict conversational QoE under controlled impairments, and which indicators (e.g., freeze burstiness, RTT tails, encoder/adaptation dynamics, audio concealment) are the dominant drivers of prediction across contents and impairment types? (Chapter 2: WebRTC QoE)

RQ2. (Time matters - Streaming) Does a sequence-aware estimator that encodes per-segment/session Key Performance Indicators (KPIs) and attends to salient events (start-up delay, switches, stalls) outperform standard streaming models across devices, and how robust is it under train/test device shifts and variations in event timing and duration? (Chapter 3: Streaming QoE)

RQ3. (Share without sharing - Multi-view QoE) Can representation-level fusion of complementary feature views approximate full-view accuracy without sharing raw data, and how does performance degrade (or remain stable) under view imbalance, missing views, and cross-dataset settings? (Chapter 4: Multi-View QoE)

RQ4. (Modality-agnostic design - NR-PCQA) Can an NR, multi-projection point cloud model with adaptive view weighting achieve state-of-the-art SCC/PCC and reduced RMSE under practical compute budgets, while remaining robust to content-level diversity, projection choices, and rendering conditions? (Chapter 5: NR-PCQA)

RQ5. (Deployability - Efficiency coverages) What inference time, memory, and FLOPs envelopes are sufficient for online estimation across the studied scenarios, and what accuracy-efficiency trade-offs emerge across backbones and fusion strategies when targeting real-time and edge constraints? (Chapter 6: FIQA case study)

A concluding synthesis chapter revisits these research questions collectively, reflecting on the

empirical findings, methodological trade-offs, and open challenges observed across both service-level (interactive and streaming) and content-level (volumetric and perceptual) QoE modeling, figure 1.1. It further outlines future directions in personalization, uncertainty-aware decision support, and cross-domain adaptation.

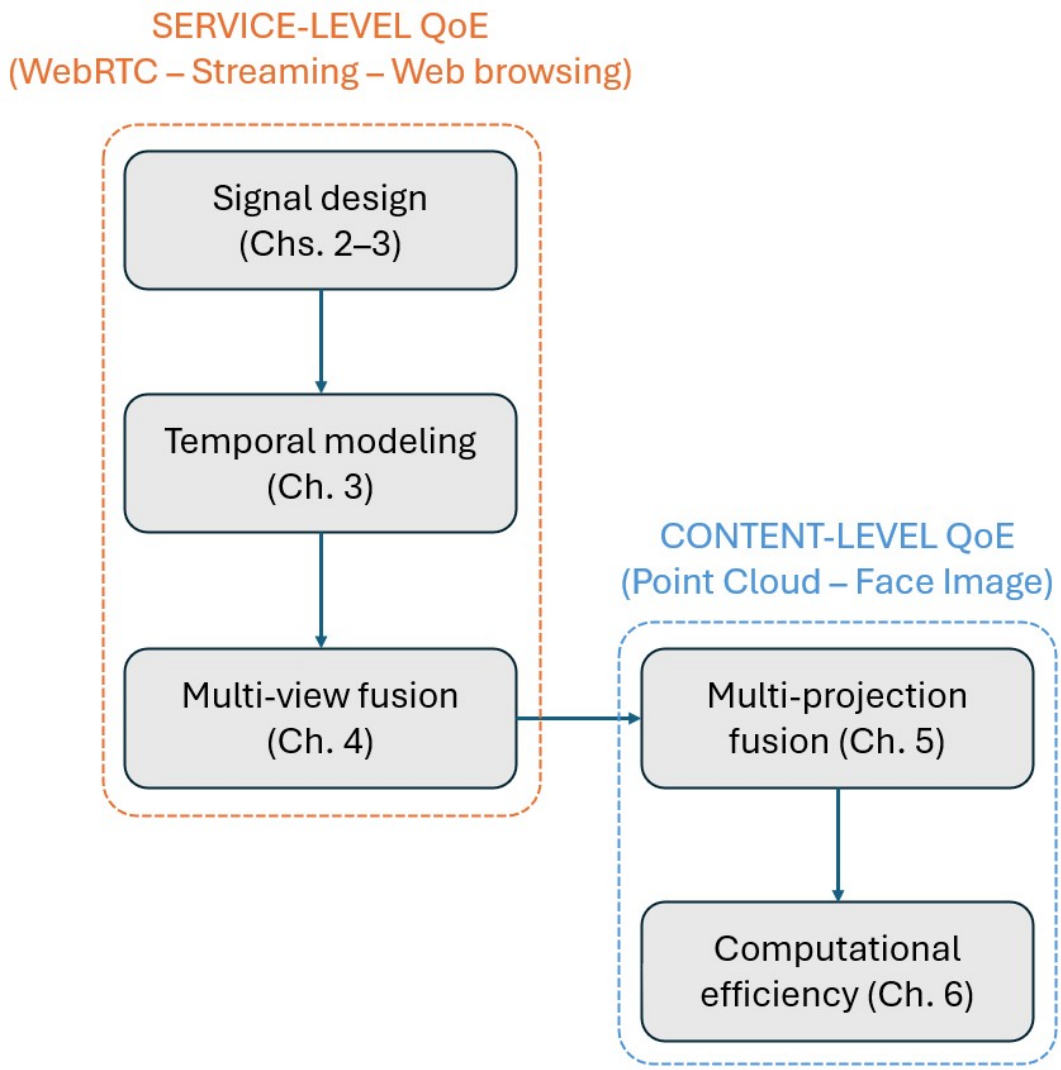

Figure 1.1: Illustration of the conceptual organization of the dissertation, mapping the research questions to the corresponding modeling activities and chapter-level contributions across service-level and content-level QoE.

1.4 | Aims and Objectives

Together, these objectives operationalize the proposed two-layer framework for learning-based QoE modeling, integrating service-level and content-level perspectives under a unified methodological paradigm.

The two-layer conceptual framework proposed in this thesis directly addresses the core questions introduced earlier, with each solution instantiated and validated in peer-reviewed publications. Together, these objectives form a systematic progression of studies, where successive publications build on earlier insights to gradually shape a methodological paradigm for objective QoE prediction. The overarching aim of this dissertation is to develop and validate a learning-based framework for objective modeling of human-perceived quality that remains reliable across heterogeneous contents, devices, and network conditions, and that is efficient enough for online monitoring and control. The doctoral study treats both service-level QoE (e.g., WebRTC conversations, adaptive video streaming) and content-level perceptual quality assessment (e.g., Image Quality Assessment (IQA) and Point Cloud Quality Assessment (PCQA)), showing how common design principles, signal selection, temporal encoding, multi-view fusion, and efficiency can be instantiated across distinct use cases. Corresponding to the research questions outlined above, the main objectives of this dissertation are as follows.

Objective 1: Signal design for conversational QoE - Addressed through the research presented in Chapter 2, which is based on the publication [4]. Application-layer telemetry from webrtc-internals is rich but noisy. This objective establishes a principled pipeline that collects audiovisual call sessions under controlled impairments (packet loss, jitter, rate limiting, delay) with matched MOS, harmonizes the logs, and selects a compact, non-redundant set of task-relevant signals via correlation/Analysis of Variance (ANOVA) statistical analysis and redundancy checks, followed by training lightweight regressors to predict conversational QoE. The result is a reproducible feature-curation recipe and an accurate, compact predictor suitable for online use, with sensitivity and ablation analyses identifying the most influential factors. (Thesis chapter 2: WebRTC QoE)

Objective 2: Temporal modeling for adaptive streaming QoE - Investigated through the study presented in Chapter 3, derived from the work published in [5]. Session quality in adaptive streaming depends on the sequence and timing of events, including start-up delay, quality switches, and stalls. So this objective develops a sequence-aware QoE estimator that encodes per-segment and session KPIs and uses transformer attention to capture long-range dependencies and emphasize salient impairments. The pipeline standardizes sequentialization and normalization of KPIs, handles variable-length sessions, and trains lightweight models that outperform standard baselines while remaining robust under cross-device and train/test shifts. The result is a generalizable temporal QoE modeling framework with systematic comparisons to established baselines and explicit cross-device validation. (Thesis chapter 3: Streaming QoE).

Objective 3: Collaborative multi-view learning under data constraints - This ob-

jective is fulfilled through the methodology and results reported in Chapter 4, which extend the work published in [6]. Informative QoE signals are often fragmented across entities or datasets, where privacy and governance rules prevent raw data pooling. This objective introduces a multi-view learning scheme that fuses intermediate representations from complementary feature views to approximate full-view performance from partial observations, without sharing raw samples. The chapter provides an architectural recipe (per-view encoders with a shared fusion head), training and consistency strategies for uneven/missing views, and empirical evidence of privacy-preserving, collaborative QoE modeling, including analyses of view complementarity, robustness under view ablations, and identified failure modes. (Thesis chapter 4: Multi-View QoE).

Objective 4: NR modeling for PCQA - This objective is realized through the study presented in Chapter 5, which builds upon the study in [7]. Volumetric/point cloud media often lack pristine references, yet operators need NR predictors that remain aligned with human perception. This objective proposes a multi-projection, adaptive weighting NR-PCQA model; a fixed set of 2D projections is rendered from each point cloud, view-specific features are extracted with lightweight backbones, and a content-adaptive fusion module assigns weights to aggregate these cues into a single quality score without relying on references. The result is a modality-agnostic NR model that attains state-of-the-art rank/linear correlations with reduced error under practical compute budgets; ablation studies quantify the contribution of each view, the effect of projection count, and robustness across diverse contents and rendering conditions. (Thesis chapter 5: NR-PCQA).

Objective 5: Efficiency and deployability - Pursued through the study presented in Chapter 6, an extended version of the research in [8]. This objective tests whether the thesis principles transfer to face image quality assessment (FIQA) under tight efficiency constraints. The unified methodology is adapted by (i) curating face-specific signals (e.g., sharpness/blur, exposure, pose, occlusion, alignment confidence) alongside compact deep features; (ii) employing efficient backbones (e.g., MobileNetV3/ShuffleNet) with lightweight heads; and (iii) optimizing correlation-aware losses to maximize SCC/PCC while controlling RMSE. The pipeline is profiled end-to-end under strict memory and hardware budgets, with ablations over input size and model width to trace accuracy and efficiency trade-offs. The result is a deployment-feasible FIQA model and evidence that the same design choices, like signal selection, efficient architectures, and careful objective design, generalize beyond media quality, reinforcing the overall approach. (Thesis chapter 6: FIQA case study).

Objective 6: Research insights and future work - Addressed in Chapter 7. The final chapter summarizes the proposed conceptual two-layer QoE assessment framework, highlighting its substantial benefits, wide applicability, scalability, and strong potential for real-time multimedia service implementation. It concludes the dissertation by synthesizing the key findings from all chapters and outlining future research directions. (Thesis chapter 7: Research insights)

1.5 | List of Publications and Awards

1.5.1 | International Conferences

- Ma, S., Chen, W. T., Gao, Q., Wang, J., Zhou, C. W., Sun, W., ... & Yue, Y. (2025). VQualA 2025 Challenge on Face Image Quality Assessment: Methods and Results. Accepted in ICCVW 2025. arXiv preprint arXiv:2508.18445.
- Hamidi, M., Amirpour, H., Atzori, L., & Timmerer, C. (2025). A lightweight ensemble-based face image quality assessment method with correlation-aware loss. Accepted in ICCVW 2025. arXiv preprint arXiv:2509.10114.
- Hamidi, M., Amirpour, H., Atzori, L., & Timmerer, C. (2025). Perceptual JND Prediction for VMAF Using Content-Adaptive Dual-Path Attention. In 2025 International Conference on Visual Communications and Image Processing (VCIP). Accepted.
- Hamidi, M. (2025). Learning-Based Objective Perceptual Quality Models Across Interactive and Immersive Media. In 2025 International Conference on Visual Communications and Image Processing (VCIP) - Doctoral Symposium Track. Accepted.
- Hamidi, M., Porcu, S., Floris, A., & Atzori, L. (2025, Sep). MVAW-PCQA: A No-reference Point Cloud Quality Assessment via Multi-View Adaptive Weighting. In 2025 17th International Conference on Quality of Multimedia Experience (QoMEX). Accepted.
- Hamidi, M. A., Porcu, S., Floris, A., & Atzori, L. (2025, June). A Transformer-Based Modeling Approach for Robust QoE Estimation in Video Streaming. In 2025 23rd Mediterranean Communication and Computer Networking Conference (MedComNet) (pp. 1-6). IEEE.
- Hamidi, M., Porcu, S., Floris, A., & Atzori, L. (2024, June). Towards the Application of Multiview Learning in Quality of Experience Collaborative MModeling In 2024 16th International Conference on Quality of Multimedia Experience (QoMEX) (pp. 286-292). IEEE.
- Hamidi, M., Bingöl, G., Floris, A., Porcu, S., & Atzori, L. (2023, December). Analysis of Application-layer Data to Estimate the QoE of WebRTC-based Audiovisual Conversations. In 2023 IEEE Globecom Workshops (GC Wkshps) (pp. 365-370). IEEE.

1.5.2 | Awards

- **First Rank** – VCIP 2025 Grand Challenge on Live Broadcasting Video Quality Assessment.
- **Second Rank** – VCIP 2025 Grand Challenge on Image Manipulation Quality Assessment.

- **Fourth Rank (Top 5)** – VQualA (ICCV 2025) Face Image Quality Assessment Challenge.

1.6 | Dissertation Structure

The present thesis addresses the development of learning-based objective models for estimating the QoE across interactive and immersive multimedia services.

It develops a two-layer conceptual framework encompassing:

- a service-level QoE layer, which models user experience as a function of network and application behavior, and
- a content-level QoE layer, which evaluates the perceptual quality of visual and volumetric media.

Across these complementary layers, the proposed framework integrates statistical analysis, feature selection, and machine learning strategies to bridge the gap between measurable system indicators and subjective human perception. Each chapter of the thesis contributes to one or more pillars of this framework, including signal design, temporal modeling, multi-view fusion, multi-projection fusion, and computational efficiency, progressively building toward a methodological paradigm for learning-based QoE prediction.

The thesis is organized into six chapters as follows:

- **Chapter 1 – Introduction.** This chapter presents the background and motivation for the research, highlighting the need for reliable, data-driven QoE models that generalize across heterogeneous contents, devices, and network conditions. It introduces the main research challenges, objectives, and questions, and outlines the methodological pillars of the proposed framework.
- **Chapter 2 – WebRTC Conversational QoE from Application-Layer Telemetry (WebRTC QoE).** This chapter investigates the feasibility of predicting user-perceived quality in real-time audiovisual communications using only application-layer telemetry collected from *webrtc-internals*. A statistical feature selection pipeline based on the PCC and ANOVA is proposed, followed by regression modeling using tree-based and neural approaches. Experimental results demonstrate that the selected features achieve strong correlation with subjective ratings, and the Multi-Layer Perceptron (MLP) model outperforms state-of-the-art QoE predictors.
- **Chapter 3 – Temporal Modeling for Adaptive Video Streaming QoE (Streaming QoE).** This chapter extends the proposed framework to *HTTP Adaptive Streaming (HAS)* services, where user-perceived quality depends on time-varying events such as start-up delays, quality switches, and playback stalls. A sequence-aware transformer-based model is developed

to capture the temporal dependencies between streaming events and user judgments. The chapter highlights the advantages of temporal encoding over snapshot-based models and provides extensive cross-device validation.

- **Chapter 4 – Collaborative Multi-View Learning for QoE Prediction (Multi-View QoE).** This chapter explores QoE estimation in distributed environments where data are fragmented across multiple entities or feature views and cannot be shared directly due to privacy constraints. A multi-view learning approach is proposed to fuse latent representations across views without exchanging raw data. The results demonstrate that collaborative representation fusion can recover performance close to that of centralized models while preserving data privacy.

- **Chapter 5 – No-Reference Point Cloud Quality Assessment (NR-PCQA).** The methodology is extended to volumetric media, where reference content is often unavailable. A PCQA model is proposed, combining multi-projection feature extraction with adaptive view weighting. This approach achieves state-of-the-art performance in terms of correlation with subjective scores and robustness across different datasets and projection settings.

- **Chapter 6 – Efficiency and Practical Deployment (FIQA Case Study).** This chapter validates the generality and deployability of the proposed methodology by applying it to *Face Image Quality Assessment (FIQA)*. The same design principles signal selection, lightweight feature extraction, and regression modeling are adapted under technical constraints, including Memory used, FLOP operations, and inference time. The results demonstrate that the proposed framework remains effective for real-time and edge-based systems, achieving competitive accuracy with compact architectures and correlation-aware loss functions.

- **Chapter 7 – Research Insights and future perspectives.** The final chapter summarizes the main findings and contributions of the thesis. It discusses the implications of the proposed unified methodology, emphasizing its generality, scalability, and potential for deployment in real-time multimedia services. Future directions include extending the framework to federated learning, uncertainty-aware decision support, and cross-modal quality prediction across emerging immersive applications.

Overall, the thesis contributes a consistent and validated framework for objective QoE modeling that integrates statistical interpretability with data-driven efficiency. The proposed methodology demonstrates strong predictive power, robustness, and scalability across multiple modalities, forming a solid foundation for the next generation of perception-aware, data-driven multimedia systems.

2

WebRTC Conversational QoE from Application-Layer Telemetry (WebRTC QoE)

2.1 | Introduction

Following the research framework and objectives presented in Chapter 1, this chapter initiates the service-level layer of the proposed framework by addressing Objective 1: Signal Design for Conversational QoE. The focus is on developing an interpretable and data-driven methodology for predicting QoE in real-time conversational services based on Web Real-Time Communication (WebRTC) telemetry. As emphasized in the introductory chapter, conversational services such as video calls and virtual meetings are key enablers of modern communication ecosystems, where perceived quality depends on both network-level performance and application-level behavior.

Traditional QoE estimation models often rely on network parameters that fail to capture user perception accurately in application-layer adaptive systems. To bridge this gap, this chapter proposes a fully application-layer QoE estimation method that leverages the rich telemetry exposed by WebRTC. By extracting and modeling meaningful signal features, such as jitter, bitrate, and delay, from real session data, the study aims to build interpretable QoE predictors that are accurate, reproducible, and suitable for real-time monitoring. This represents the first practical step toward achieving the dissertation's broader goal of learning-based, deployable QoE models that can operate autonomously across diverse multimedia contexts.

The increasing adoption of real-time communication (RTC) technologies has fundamentally transformed how users interact in audiovisual services such as video conferencing, online education, telemedicine, and remote collaboration. Over the last decade, WebRTC has emerged as the leading open-source framework that enables peer-to-peer audiovisual communication directly in web browsers

[9], without the need for additional software or plug-ins. Its flexibility, low latency, and native integration into browsers have made it the foundation for modern platforms such as Google Meet, Zoom, and Microsoft Teams.

With the growing reliance on such real-time services, the perceived quality of communication, or QoE, has become a central concern. While Quality of Service (QoS) metrics, such as packet loss, jitter, and latency, describe network behavior, they do not directly reflect how users perceive the quality of communication. In contrast, QoE captures the subjective satisfaction of users, influenced not only by network conditions but also by encoding parameters, device capabilities, and even conversational dynamics. As WebRTC operates over heterogeneous networks and devices, understanding and predicting QoE from observable system metrics is both crucial and challenging.

Traditional approaches to QoE estimation often rely on subjective testing following standardized methodologies such as ITU-T P.910 or P.913 [10, 11]. These studies, while accurate, are time-consuming, expensive, and not feasible for large-scale or real-time monitoring. As a result, there has been increasing interest in objective, data-driven QoE estimation models that can automatically infer user-perceived quality from network or application metrics [12, 13]. Most existing methods, however, depend heavily on network-layer statistics or simulation environments, which may not fully capture the application-level adaptation mechanisms used by WebRTC (e.g., congestion control, bitrate adaptation, frame rate regulation).

In this context, this chapter focuses on modeling QoE at the application layer, directly from telemetry data generated by WebRTC itself. Unlike network-level traces, these logs provide a detailed view of the audiovisual session as managed by the browser, including send/receive bitrates, packet loss, jitter buffer size, round-trip time (RTT), frame rate, and resolution changes. Such data offer a valuable opportunity to build lightweight, interpretable models that can predict QoE in real time without intrusive measurement tools.

The core objective of this chapter study is to investigate the relationship between application-layer parameters and user-perceived QoE, and to develop an ML-based framework that can predict perceived quality from WebRTC telemetry. To achieve this, several steps are taken:

1. A dataset of audiovisual conversations is collected under controlled network impairments, including delay, jitter, and packet loss.
2. Subjective scores are obtained using standardized Mean Opinion Score (MOS) ratings from participants.
3. Statistical analyses, such as the Pearson correlation coefficient (PCC) and the Analysis of Variance (ANOVA), are employed to identify the most influential parameters.
4. Different machine learning (ML) models (e.g., Random Forest, XGBoost, and Regression Trees) are trained and compared for their ability to predict QoE based on selected parameters.

The significance of this study lies in its application-oriented perspective. By using the data that WebRTC already provides, this approach eliminates the need for specialized hardware or deep packet inspection, making it practical for real-time deployment in browsers or at the edge. Furthermore, the insights obtained from this study inform the signal design pillar of the broader QoE modeling framework introduced in this thesis, establishing which metrics are most predictive of human perception and how they should be processed.

This chapter is organized as follows:

Section 2.2 reviews related work and research gaps on network and application-level QoE estimation. Section 2.3 outlines the methodology, which includes data collection, feature selection, and model training. Section 2.4 proposes ML models and implementations for QoE estimation modeling, followed by the results and comparison with the state-of-the-art approaches. Section 2.5 reports the model performance results in terms of RMSE and R^2. Section 2.6 presents key findings and contributions of the chapter. Finally, Section 2.7 concludes the chapter with reflections and directions for future work.

2.2 | Background and Related Work

The estimation and prediction of QoE for real-time audiovisual communication have been widely studied over the last two decades. The research field has evolved from early network-level QoS-based models to more recent application-aware and data-driven approaches, many of which apply ML or statistical methods. This section provides an overview of related work organized along three main directions: (i) network-level QoE estimation, (ii) application-layer and hybrid approaches, and (iii) WebRTC-specific modeling efforts.

Early studies on QoE prediction primarily focused on establishing empirical relationships between network performance indicators such as packet loss, delay, and jitter and user-perceived quality. The common assumption was that QoE could be expressed as a deterministic function of a limited number of network metrics, often through non-linear regression or exponential decay models. For example, a study in [14] conducted a subjective campaign of quality assessment on artificially generated VoIP calls, collecting the values of network metrics associated with each test call. Similarly, [15] explored the relationship between one-way delay, jitter, and MOS in conversational scenarios, confirming that even small latency variations could significantly impact interactivity. These models were simple and interpretable, but they often failed to generalize because they neglected the adaptive mechanisms of modern multimedia applications.

More sophisticated approaches began integrating multiple QoS metrics. In the study [16], they provided studies on three network-impaired video sets with the aim to provide a comprehensive evaluation of the effects of networks on video quality. In study [17], a parametric packet-based

model has been created to estimate user-perceived audiovisual quality of Internet Protocol Television (IPTV) services. It is divided into three modules: audio, video, and audiovisual quality. However, these models remained technology-dependent and relied heavily on controlled lab conditions. A key limitation of network-level QoE models is that they require access to packet traces or deep packet inspection (DPI). Such data are often unavailable in real deployments due to encryption, privacy regulations, or performance overhead. Moreover, network parameters alone cannot describe how codecs, buffers, or adaptive controls mitigate impairments. As a result, there has been a gradual transition toward application-layer and hybrid QoE modeling approaches.

Application-layer QoE estimation methods exploit information available within the application's control or adaptation layer. Unlike network metrics, these indicators—such as sender bitrate, frame rate, frame drops, buffer occupancy, or resolution changes reflect the actual experience of the end user more directly. For adaptive video streaming, work in [18] introduced the concept of QoE-aware bitrate adaptation, showing that metrics like rebuffering duration, quality switches, and start-up delay are primary determinants of user satisfaction. Similarly, a study in [19] proposed the Model Predictive Control (MPC) approach for Adaptive Bitrate (ABR) streaming, explicitly modeling how bitrate and rebuffering jointly affect QoE. In real-time applications, a study in [20] demonstrated that application-level parameters (e.g., encoder bitrate, frame rate, and freeze ratio) could be used to predict perceived quality in live streaming.

Hybrid approaches combine network and application-layer data, achieving improved accuracy but at the cost of complexity. The work in [21] argued that hybrid models can better capture user experience under heterogeneous networks, but their deployment remains limited because of the difficulty in synchronizing metrics from multiple layers.

The evolution of application-layer modeling has been strongly influenced by the growing adoption of ML. Instead of fitting pre-defined equations, ML-based models learn non-linear relationships directly from data. The work done by [22] and [13] used ML algorithms to predict QoE based on the collected session data, achieving higher accuracy than classical regression models. These advances set the foundation for WebRTC-specific QoE prediction, where ML methods can directly exploit browser telemetry. WebRTC has become the standard framework for browser-based real-time communication, yet its QoE modeling presents unique challenges. Because it operates under highly variable conditions—different browsers, operating systems, and devices QoE estimation must account for both network variability and application-level adaptation.

Early work in [23] presents a model-free Deep Reinforcement Learning approach aimed at enhancing user experience by managing the data rate of media streams transmitted in the uplink direction by a moving, remote-controlled device. The model incorporates WebRTC-compliant metrics to ensure its seamless integration into real-world applications. Its primary objective is to maximize a reward function specifically designed to align with users' perceptions of video streams.

Another study in [24] presents a comprehensive analysis of *webrtc-internals* tool[1], a powerful tool integrated into the Google Chrome browser for gathering WebRTC statistics. Our primary objective is to demonstrate that these statistics can effectively predict the QoE for WebRTC video calls. Through a series of rigorous experiments and an end-user questionnaire, we successfully collected WebRTC Internals statistics and MOS. By applying Multiple Linear Regression (MLR), we quantified the relationship between selected WebRTC Internals statistics and QoE, ultimately leading to a robust prediction model for QoE in WebRTC video calls.

2.2.1 | Research Gap and Contribution

Despite the growing body of research on WebRTC and real-time audiovisual QoE, existing approaches still face several limitations that hinder their applicability in live, user-centric environments. A comprehensive review of the literature reveals persistent challenges related to data accessibility, feature selection, and model generalization.

The main research gaps can be summarized as follows:

1. Limited integration of multiple impairment factors. Most studies analyze one impairment type (e.g., packet loss) at a time, whereas real WebRTC calls are affected by combinations of delay, jitter, and rate constraints.

2. Lack of systematic feature analysis. Previous models often rely on arbitrary or fixed sets of parameters without rigorous statistical selection.

3. Absence of a reproducible pipeline. Few works document their data preprocessing, feature correlation, or validation process, which hinders comparison and reuse.

4. Over-reliance on network metrics. Many approaches still depend on network-layer data, limiting their deployability in end-user applications.

5. Limited focus on real-time feasibility. Complex models may provide good accuracy but are not optimized for online or browser-based prediction.

Overall, this chapter contributes to the dissertation's Objective 1: Signal Design for Conversational QoE, by establishing the foundational framework for learning-based QoE modeling from application-layer telemetry. This chapter addresses these gaps by introducing a fully application-layer QoE estimation framework for WebRTC audiovisual communication. The proposed method leverages telemetry available from webrtc-internals to construct an interpretable, data-driven model that predicts subjective MOS scores under multiple impairment scenarios. It employs a systematic

[1] chrome://webrtcinternals/

signal design approach, combining correlation analysis, ANOVA-based feature selection, and ML regression to build a reproducible and efficient QoE predictor. Moreover, the analysis provides insight into which features matter most for human-perceived quality, forming a foundation for adaptive QoE management and the broader signal-design pillar of this dissertation.

This work not only validates the feasibility of application-layer QoE prediction but also provides the methodological basis upon which the subsequent chapters expand, first toward temporal modeling for adaptive streaming QoE (Chapter 3) and later toward collaborative, modality-agnostic, and deployable QoE frameworks.

2.3 | Methodology

This section describes the experimental design, data acquisition, and modeling process developed to estimate the QoE of WebRTC-based audiovisual communication using application-layer telemetry. The goal is to establish a reproducible and interpretable pipeline capable of predicting user-perceived quality directly from browser-generated metrics, without requiring deep packet inspection or network-layer data. So, in a nutshell, the main objectives of this work are to: i) investigate the relationship between WebRTC session parameters and the users' QoE through statistical analysis; ii) utilize the identified parameters to build ML-based QoE estimation models; iii) compare and discuss experimental findings and model performance with the state-of-the-art.

2.3.1 | Subjective experiment

The subjective test aimed to assess the QoE of WebRTC-based audiovisual conversations under a variety of network impairment conditions. The goal was to understand how degradations such as delay, jitter, and packet loss rate (PLR) jointly affect users' perception of communication quality.

Table 2.2 summarizes the test conditions (TCs) used in the study, representing different combinations of the three impairment factors. These parameters were chosen to emulate realistic yet controlled network scenarios capable of stressing the built-in adaptation mechanisms of WebRTC. The experiments were carried out using Google Chrome, which natively supports WebRTC-based calls. Two laptop computers were connected in a peer-to-peer setup, allowing real-time audiovisual communication through the browser.

A dedicated local network was created between the two laptops using a managed router, thereby ensuring that the test environment remained isolated from unpredictable Internet traffic and background congestion. This isolation was crucial to guarantee reproducibility and to make sure that the only distortions affecting the communication were those intentionally introduced in each test condition.

The PyNetem tool was used to inject network impairments at the link layer. PyNetem extends the Linux Traffic Control (tc) module by allowing precise manipulation of network delay, jitter, and packet loss. Each combination of parameters in Table 2.2 was applied for the entire duration of a conversation, ensuring that both participants experienced a consistent degradation throughout the session.

The chosen impairment values may appear high compared with those typically used in Voice over IP (VoIP) experiments. However, this choice is justified by the resilience mechanisms implemented within the WebRTC stack and the Google Chrome browser. These include adaptive bitrate control, forward error correction (FEC), and packet retransmission strategies provided by the Google Congestion Control (GCC) algorithm [25]. Due to these compensation techniques, WebRTC-based applications can often maintain acceptable audiovisual quality even under moderate levels of network stress. Therefore, stronger network distortions were necessary in this experiment to produce observable differences in perceived quality and to align with the levels of distortion considered in related QoE research [26, 27].

A total of 20 participants (11 females and 9 males), aged between 23 and 36 years, took part in the experiment. All participants had normal or corrected-to-normal vision and hearing and were familiar with online video calls. They were paired into 10 conversational dyads, each performing a series of 15 conversations (one per test condition). The conversational content followed the "Who am I?" game—a lighthearted, semi-structured task where each participant attempts to guess a celebrity selected by their partner by asking yes/no questions in turn. This task was chosen because it creates a natural two-way interaction that mirrors real-life conversational dynamics while ensuring consistent speech activity across all sessions.

During each conversation, the webrtc-internals tool in Chrome was enabled to capture detailed statistics about the ongoing communication session. These statistics included audio and video stream metrics such as packet loss, jitter, frame rate, encoder/decoder statistics, and round-trip time. All logs were saved locally as JSON files for later analysis and feature extraction. Each conversation lasted approximately 60 seconds, balancing natural interaction time with participant fatigue.

After completing each session, both participants were immediately asked to rate the perceived audiovisual quality of the conversation using the 5-point Absolute Category Rating (ACR) scale, as recommended by ITU-T Rec. P.800 [28]. The ACR scale includes the following levels: 1 - Bad, 2 - Poor, 3 - Fair, 4 - Good, and 5 - Excellent (see Table 2.1).

All scores were collected individually and later averaged across participants to obtain the MOS for each test condition. Before computing the MOS, data consistency was checked, and two participants were identified as outliers due to inconsistent rating patterns (e.g., uniform responses across conditions). Consequently, the final MOS values reported in Table 2.2 represent the average ratings from 18 valid participants.

The resulting MOS trends clearly show that packet loss is the most critical factor affecting au-

diovisual quality in WebRTC conversations. Even at moderate levels (15%), packet loss significantly reduced the perceived smoothness of the conversation, often leading to noticeable freezes and desynchronization between audio and video. The impact was even more severe when packet loss was combined with delay and jitter, causing frequent audio dropouts and visible frame stuttering, which participants rated as "Poor" (average MOS ≈ 2).

Conversely, conversations with only delay or jitter, but no packet loss, were perceived as tolerable. Most participants described them as "slightly annoying but understandable," corresponding to MOS values between 3 ("Fair") and 4 ("Good"). This suggests that WebRTC's congestion control and adaptive encoding strategies effectively mitigated delay and jitter alone but could not fully compensate for packet losses that resulted in data corruption or missing frames.

Overall, the subjective results confirm that the perceived QoE in WebRTC-based audiovisual calls depends on a complex interplay of network factors. Among these, packet loss remains the dominant impairment, while delay and jitter primarily amplify its impact when occurring simultaneously. These findings provide a strong motivation for analyzing application-layer telemetry parameters, as they capture how WebRTC internally reacts to such distortions and how these reactions correlate with user-perceived quality.

Table 2.1: Quality levels based on subjective scores.

Score	Quality Level	Description
5	Excellent	Imperceptible degradation
4	Good	Perceptible but not annoying
3	Fair	Slightly annoying
2	Poor	Annoying
1	Bad	Very annoying

2.3.2 | Statistical analysis

Figure 2.1 illustrates the complete statistical analysis and data preprocessing pipeline adopted to extract informative features from the raw telemetry logs collected through the webrtc-internals tool. The goal of this phase was to transform the heterogeneous, time-varying WebRTC parameters into a consistent numerical representation suitable for ML-based QoE modeling.

For each test condition TC_i (with $i \in \{1, 2, \ldots, 15\}$), the webrtc-internals tool recorded a total of 116 WebRTC session parameters ($p_{i,j,k}$, where $j \in \{1, 2, \ldots, 116\}$) indexes the parameters and $k \in \{1, 2, \ldots, 18\}$ represents the test participants. These parameters include a broad range of indicators describing the audio, video, and data channel performance of each WebRTC session.

Examples of such parameters are the Round Trip Time (RTT), jitter buffer delay, packets lost, bytes sent, frames encoded, and keyframes received. Each parameter was captured as a time series

Table 2.2: Test conditions (TCs) and MOS

TC	Delay (ms)	Jitter (ms)	PLR (%)	MOS
1	0	0	0	3.89
2	500	0	0	3.67
3	1000	0	0	3.72
4	500	500	0	3.44
5	1000	500	0	3.44
6	0	0	15	3.22
7	500	0	15	3.06
8	1000	0	15	2.72
9	500	500	15	2.33
10	1000	500	15	2.44
11	0	0	30	2.56
12	500	0	30	2.11
13	1000	0	30	2.11
14	500	500	30	1.89
15	1000	500	30	1.44

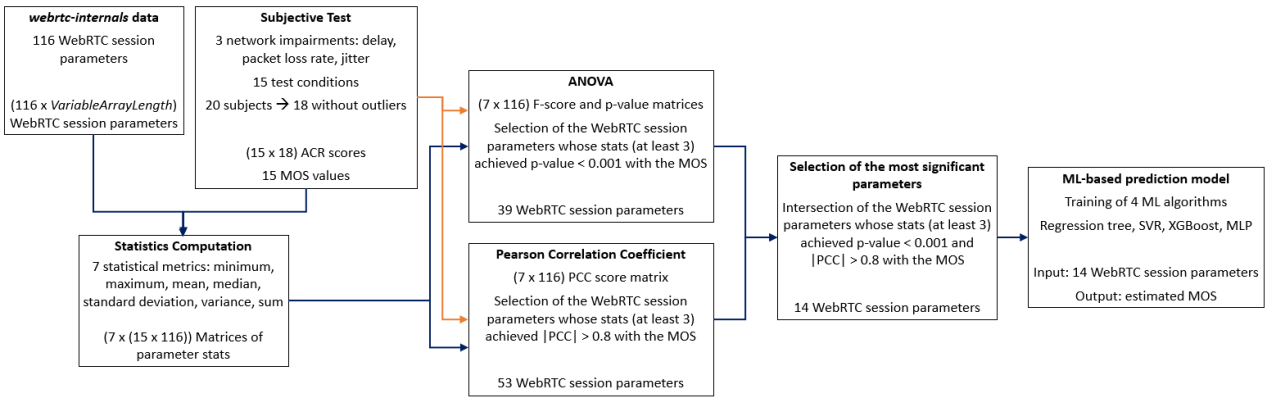

Figure 2.1: Overview of the statistical analysis process.

sampled throughout the duration of the conversation, typically with one data point per second. As a result, each WebRTC parameter was stored as an array of variable length depending on the sampling frequency and session duration.

To make the dataset homogeneous and manageable, the first processing step involved the aggregation of telemetry data across participants for each test condition. Algorithm 1 formalizes this procedure. Specifically, for each parameter $p_{i,j,k}$ collected from all 18 participants under the same TC_i the parameter data were concatenated into a unified vector $p_{i,j}$. This operation ensured that random variations due to individual participant behavior were smoothed out, producing a single consolidated distribution per parameter and test condition. Next, we computed a set of seven descriptive statistics for each joined parameter vector $p_{i,j}$.

1. Maximum (max) – captures the upper bound of the parameter's fluctuation range.
2. Minimum (min) – indicates the lowest observed value.
3. Mean – reflects the average behavior over the session.
4. Median – represents the central tendency, less sensitive to outliers than the mean.
5. Standard deviation (std) – quantifies short-term variability.
6. Variance (var) – measures the spread of the data distribution.
7. Sum – provides a measure of the total accumulated activity (e.g., total bytes or packets).

Algorithm 1 Calculation of statistical metrics for the WebRTC session parameters.

for all TC_i **do** ▷ $i \in \{1, 2, \ldots, 15\}$
  for all $p_{i,j}$ **do** ▷ $j \in \{1, 2, \ldots, 116\}$
    $p_{i,j} = \bigcup_{k=1}^{18} p_{i,j,k}$
    $p_{i,j}^{min} = min(p_{i,j})$
    $p_{i,j}^{max} = max(p_{i,j})$
    $p_{i,j}^{mean} = mean(p_{i,j})$
    $p_{i,j}^{med} = median(p_{i,j})$
    $p_{i,j}^{stdev} = standardDeviation(p_{i,j})$
    $p_{i,j}^{var} = variance(p_{i,j})$
    $p_{i,j}^{sum} = sum(p_{i,j})$
  end for
end for

Each of these seven metrics was computed for all 116 parameters and for all 15 test conditions, resulting in seven distinct 15×116 matrices — one per statistical metric. For instance, the "mean" matrix contains the average values of all 116 parameters across the 15 network conditions, while

the "std" matrix captures their variability. Collectively, these matrices form what we refer to as the parameter stats, a structured dataset summarizing the dynamic behavior of WebRTC sessions across all impairment scenarios. This process provided a compact yet information-rich representation of the original telemetry data. Instead of working with raw time series of varying length, which are difficult to align and compare, each test condition was now described by a fixed-size feature vector derived from the parameter stats. This transformation facilitated the subsequent statistical and machine-learning analyses while maintaining interpretability — each feature retained its physical meaning (e.g., average bitrate, variance of RTT).

The next step involved the identification of the most relevant WebRTC session parameters in relation to the MOS collected during the subjective experiment. The main objective was to discover which session metrics were statistically correlated with perceived quality, thus serving as effective predictors for QoE modeling.

To this end, we applied two complementary statistical techniques: the PCC and the ANOVA. The PCC analysis quantifies the strength and direction of the linear relationship between each parameter stat and the MOS values. For a given parameter $p_{i,j}$, the PCC was computed between its vector of statistical values (across the 15 TCs) and the corresponding MOS vector. Features exhibiting a strong correlation — defined as $|r| > 0.8$ for at least three of the seven statistics — were retained as PCC-selected parameters. This filtering process ensured that only those parameters with consistent linear dependencies on perceived quality were kept, reducing noise and redundancy.

The ANOVA test complements this by assessing whether differences in parameter distributions correspond to statistically significant differences in MOS across test conditions. Unlike PCC, which captures correlation, ANOVA tests the discriminative power of each parameter with respect to QoE. For each parameter, we performed seven ANOVA tests (one per statistic), obtaining the associated p-values. Parameters that showed significance ($p < 0.001$ in at least three metrics) were considered ANOVA-selected parameters.

Finally, the intersection of the PCC and ANOVA sets was computed to obtain a joint parameter set, representing the most robust indicators — those that are both strongly correlated with and significantly discriminative for perceived quality. This systematic feature selection yielded three groups of parameters:

- PCC-selected parameters (24 total).
- ANOVA-selected parameters (34 total), and
- Joint parameters (22 total), which represent their intersection.

The parameter stats and their statistical selection process play a pivotal role in this study, as they form the bridge between raw telemetry data and interpretable QoE indicators. Through this two-step

statistical analysis, we effectively reduced the complexity of the feature space while preserving the most perceptually relevant information. The resulting parameter subsets were subsequently used as input features for machine-learning-based QoE prediction models, described in the next section.

2.3.3 | Pearson Correlation Coefficient

The PCC is one of the most widely used statistical tools for evaluating the linear relationship between two continuous variables. It provides an intuitive, quantitative indication of how one variable changes in relation to another. Mathematically, the PCC is defined as:

$$r_{X,Y} = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y} \tag{2.1}$$

where $\text{cov}(X,Y)$ represents the covariance between the two variables X and Y, σ_X and σ_Y are their standard deviations. The resulting correlation coefficient $r_{X,Y}$ ranges from -1 to +1.

- A PCC value close to +1 indicates a strong positive correlation, meaning that as one variable increases, the other tends to increase as well.
- A PCC value close to -1 indicates a strong negative correlation, where one variable increases while the other decreases.
- Values near 0 suggest little or no linear relationship between the two variables.

In practice, absolute correlation values greater than 0.8 are typically interpreted as strong correlations, while those below 0.3 are weak or negligible. Intermediate values may still indicate some association but are often influenced by noise or non-linear dependencies.

In the context of this study, the two datasets under comparison were the MOS obtained from the subjective experiment and each statistical feature (or "parameter stat") derived from the webrtc-internals logs. The MOS reflects human-perceived audiovisual quality, while the parameter stats capture measurable technical indicators of the communication session (e.g., bitrate, jitter, or frame rate). The objective of this analysis was to determine which technical parameters were most linearly associated with the perceived QoE, thus potentially serving as strong predictors for subsequent model training.

For each of the 116 WebRTC session parameters, we computed seven PCC values — one for each statistical descriptor (maximum, minimum, mean, median, standard deviation, variance, and sum) previously obtained during data preprocessing. This approach allowed us to capture not only the average behavior of a parameter but also how its variability and extremes relate to QoE. For instance, a parameter might have a weak correlation in its mean but a strong correlation in its variability, indicating that fluctuations rather than absolute levels drive user perception.

To systematically identify the most relevant parameters, we established a correlation selection rule: A WebRTC session parameter is considered correlated with the MOS if at least three out of its seven parameter stats exhibit an absolute PCC value greater than 0.8.

This rule was designed to ensure robustness — avoiding the inclusion of parameters that showed a strong correlation only sporadically or due to random fluctuations. By requiring at least three independent statistics to exceed the threshold, the method effectively filtered out unstable or context-dependent relationships.

As an illustrative example, consider the parameter RTCInboundRTPAudioStreamInsertedSamplesForDeceleration, which represents the number of additional audio samples inserted by WebRTC's jitter buffer to compensate for late packet arrivals. High values of this parameter generally indicate buffer expansion due to jitter, leading to perceptible audio delays or distortions. The computed PCC values for this parameter across its seven statistics, including mean: -0.885, min: 0.153, max: -0.613, median: -0.860, stdev: -0.834, var: -0.816, sum: -0.900.

Since five out of seven of these metrics exceed the absolute threshold of 0.8, this parameter was classified as strongly correlated with the MOS. The negative sign of most coefficients indicates an inverse relationship — higher buffer deceleration corresponds to lower perceived quality, as expected from a user-experience standpoint.

Overall, the PCC analysis identified 24 WebRTC session parameters that demonstrated a strong linear correlation with MOS across multiple statistical measures. These parameters collectively form what we refer to as the PCC-selected feature set, denoted as pcc_params. This subset represents the most perceptually relevant technical indicators in our dataset, each reflecting measurable aspects of WebRTC's adaptive behavior that users notice during impaired communication.

The importance of this analysis goes beyond numerical selection. The identified pcc_params offer valuable insights into how WebRTC manages impairments at the application level. For example, parameters related to frame encoding rate, packet loss counters, and jitter buffer occupancy all exhibited high correlations, reinforcing that QoE in real-time audiovisual calls is jointly influenced by both video continuity and audio smoothness. These observations provide an interpretable foundation for the subsequent ML phase, where the pcc_params are used as key inputs to train regression models aimed at predicting QoE directly from WebRTC telemetry.

2.3.4 | Analysis of variance

The ANOVA is a fundamental statistical method used to assess whether there are significant differences between the means of two or more groups. In essence, ANOVA evaluates whether the observed variability in a dataset can be attributed to differences among group means or is instead the result of random noise. It is a hypothesis-driven approach that provides insight into how strongly one variable (the independent factor) influences another (the dependent factor).

In the context of this research, ANOVA was employed to determine which WebRTC session parameters exhibited statistically significant relationships with MOS, the subjective measure of perceived audiovisual quality. The underlying question was whether variations in a given telemetry feature corresponded to meaningful variations in user-perceived QoE across different network impairment conditions.

Mathematically, ANOVA tests the null hypothesis (H_0) that all group means are equal, i.e.,

$$H_0 : \mu_1 = \mu_2 = ... = \mu_k \tag{2.2}$$

where each μ_i represents the mean of MOS values corresponding to a particular test condition or feature grouping. The alternative hypothesis (H_a) states that at least one of these means differs from the others, implying a non-random effect of the parameter on QoE.

When the ANOVA test yields a p-value below a chosen significance threshold, the null hypothesis is rejected, suggesting that the observed differences among means are unlikely to be due to chance alone. In this study, we adopted a strict significance threshold of $p < 0.001$. This conservative choice ensures that only parameters with very strong statistical evidence of association with MOS are retained, minimizing the likelihood of false positives.

To carry out the analysis, ANOVA was computed between the MOS values (dependent variable) and each parameter stat (independent variable). As in the PCC analysis, each WebRTC session parameter was represented by seven statistical descriptors—maximum, minimum, mean, median, standard deviation, variance, and sum—capturing different aspects of its distribution across the 15 test conditions.

For each parameter, we therefore performed seven separate ANOVA tests, one for each statistic, obtaining a corresponding p-value that quantifies how strongly the parameter's variation is linked to changes in perceived quality. This process allowed us to assess not only whether a parameter influenced MOS but also whether specific statistical aspects (e.g., variability vs. central tendency) were more perceptually relevant.

To ensure robustness and reduce sensitivity to random fluctuations, we applied the following selection criterion: A WebRTC session parameter is considered significant for the MOS if at least three out of its seven statistics achieve a p-value < 0.001. This rule mirrors the approach used in the PCC-based selection and emphasizes stability across multiple statistical views of the same parameter. By requiring significance in at least three metrics, we guarantee that the parameter's influence on QoE is consistent and not an artifact of a single distributional property.

As an illustrative example, the parameter RTCInboundRTPAudioStreamInsertedSamplesForDeceleration—which measures the number of additional audio samples inserted by the jitter buffer to handle late packets—showed the following p-values across its statistics: mean: < 0.001, min: < 0.001, max: 0.0149, median: < 0.001, stdev: < 0.001, var: < 0.001, sum: < 0.001. Since six out of seven statistics passed the significance threshold, this parameter was classified as highly significant for the

MOS. This result aligns with the earlier PCC findings, reinforcing the importance of this parameter as a key indicator of perceived quality. In practical terms, frequent buffer decelerations indicate packet arrival instability, which users perceive as degraded smoothness or delayed audio, thus lowering QoE.

Overall, the ANOVA analysis identified 34 WebRTC session parameters that met the significance criterion, confirming that these parameters exhibit systematic and measurable effects on user-perceived quality across the evaluated test conditions. These parameters are hereafter referred to as the ANOVA-selected feature set, or anova_params.

Together with the PCC-based analysis, ANOVA provides complementary insights:

- While PCC highlights linear correlations between technical and perceptual metrics,
- ANOVA focuses on variance-based significance, identifying features that cause perceptible differences in quality ratings.

The combination of these two methods yields a robust feature selection strategy that captures both correlation and discriminative strength. This dual-stage selection ensures that the retained parameters are not only statistically relevant but also perceptually meaningful, forming a solid foundation for training ML models that can accurately estimate QoE based solely on application-layer telemetry.

2.4 | QoE Estimation Models

In this section, we describe the proposed ML-based models used for estimating the QoE from application-layer telemetry, along with the achieved performance results. The overarching goal of this stage was to evaluate how effectively the WebRTC session parameters identified through the PCC and ANOVA analyses could predict user-perceived quality, as expressed by the MOS obtained during the subjective tests.

The dataset used for model training consisted of the values achieved by the selected WebRTC session parameters across the 15 test conditions (TCs), each representing a specific combination of delay, jitter, and packet loss. For each TC, the webrtc-internals telemetry provided 116 parameters whose statistical features (as detailed in the previous section) were aggregated into structured matrices. After feature selection, these were reduced to smaller subsets representing the most perceptually relevant metrics. The target output to be estimated by each regression model was the corresponding MOS value, as reported in Table 2.2. Thus, the input–output pairs consisted of numerical features derived from WebRTC telemetry and the subjective quality ratings given by human participants.

To systematically explore the impact of different feature selection strategies on model performance, we trained and evaluated a total of nine regression models, organized into three groups

according to the feature sets used as input. Each group employed one of the following feature subsets:

- *pcc_params*: the 24 WebRTC session parameters identified as correlated with MOS according to the PCC analysis in Section 2.3.3. These parameters capture strong linear relationships between network/application-level statistics and perceived quality.
- *anova_params*: the 34 WebRTC session parameters determined to be significant for MOS via the ANOVA test presented in Section 2.3.4. These features exhibit statistically meaningful differences in distribution corresponding to changes in user-perceived quality.
- *joint_params*: the 22 parameters common to both *pcc_params* and *anova_params*, representing the intersection of the two sets. This subset contains features that are simultaneously strongly correlated with and statistically significant for QoE. It includes all parameters from *pcc_params* except for two specific metrics [2] [3] which did not meet the ANOVA significance threshold.

The rationale for training separate models with each feature subset was to compare the predictive capacity and robustness of correlation-based, significance-based, and combined feature selection strategies. By analyzing their relative performance, we could determine which statistical approach most effectively isolates the parameters that drive perceived quality in WebRTC-based audiovisual conversations.

To ensure methodological consistency and fair comparison, each of the three feature sets was used to train three different regression algorithms, covering a spectrum from interpretable to highly expressive models:

1. Decision Tree Regressor (DT) — a transparent model that recursively partitions the feature space based on thresholds, providing interpretability and insight into the hierarchical importance of individual parameters.
2. Extreme Gradient Boosting (XGBoost) — a powerful ensemble of decision trees optimized for predictive performance, capable of capturing non-linear dependencies between features and MOS.
3. Multilayer Perceptron (MLP) — a feedforward neural network with one or more hidden layers, used to explore the potential of deep learning for mapping complex feature–QoE relationships.

[2] $RTCMediaStreamTrack\_sender\_hugeFramesSent$
[3] $RTCRemoteInboundRtpAudioStream\_packetsLost$

Each model was trained and validated using a 5-fold cross-validation procedure to mitigate overfitting and to ensure statistical robustness despite the limited number of test conditions. Within each fold, 80% of the TCs were used for training and 20% for testing, and the process was repeated across all folds to obtain mean and standard deviation performance metrics.

Model performance was evaluated using two standard regression metrics:

- The coefficient of determination (R^2), indicating how well the model explains the variance in the MOS; and
- The Root Mean Square Error (RMSE), representing the average prediction error in MOS units.

These metrics were chosen for their interpretability in QoE studies: a high R^2 value close to 1 denotes strong predictive alignment with user scores, while a low RMSE indicates precise estimation with minimal deviation from subjective ratings.

The combination of these regression techniques and feature subsets allowed us to perform a comprehensive comparison across three dimensions:

1. The statistical nature of the input features (correlation-based, significance-based, or combined),
2. The model complexity and interpretability (tree-based vs. neural models), and
3. The stability of results across multiple cross-validation folds.

Through this structured design, the experimental pipeline not only assessed the predictive strength of selected parameters but also provided insights into the most suitable modeling approach for real-time QoE estimation from WebRTC telemetry. The following subsection presents and discusses the achieved results, including a detailed performance comparison among the trained models and an interpretation of which feature selection strategy produced the most reliable and generalizable QoE predictor.

2.4.1 | ML algorithms

The ML algorithms adopted in this study were designed to predict the MOS directly from the selected WebRTC session parameters. Three regression models were trained and compared in order to capture different types of relationships between the input features and the subjective quality ratings. The selected models were chosen to cover a range of complexity, interpretability, and representational power, from simple tree-based learners to non-linear neural networks.

- **Regression Tree:** This model is a form of decision tree regressor where the target variable is continuous rather than categorical. A regression tree partitions the feature space into regions

by recursively splitting the data according to feature thresholds that minimize prediction error. Each internal node in the tree represents a decision based on one feature, while each leaf node corresponds to a predicted continuous output value. The primary advantages of regression trees are their interpretability and their ability to model non-linear relationships without requiring explicit feature transformations. In this work, the regression tree served as a baseline model, providing insights into the hierarchical structure of the most influential WebRTC parameters that drive variations in perceived QoE.

- **Extreme Gradient Boosting (XGBoost):** This algorithm is an efficient and scalable implementation of the gradient boosting framework for supervised learning. XGBoost constructs an ensemble of decision trees, where each new tree is trained to correct the residual errors made by the ensemble of previous trees. The boosting process continues iteratively, gradually improving predictive performance while preventing overfitting through regularization techniques such as shrinkage and subsampling. The term *gradient* refers to the fact that each tree is optimized with respect to the gradient of a differentiable loss function, typically using gradient descent. XGBoost is particularly effective for small to medium-sized datasets like those in this study, as it balances accuracy, generalization, and computational efficiency. In the context of QoE prediction, XGBoost captures subtle non-linear dependencies between network parameters and user perception, allowing the model to identify complex interactions between delay, jitter, and packet loss.

- **Multi-Layer Perceptron (MLP):** The MLP is a feedforward artificial neural network capable of learning complex non-linear mappings between input and output spaces. It consists of an input layer, one or more hidden layers with non-linear activation functions, and an output layer. In this study, the MLP regressor was trained using the backpropagation algorithm with the *Adam* optimizer, minimizing the mean squared error loss. The activation functions used in the hidden layers introduce non-linearity, enabling the network to model intricate relationships that simpler models may fail to capture. Since the target output (MOS) is continuous, the output layer used a linear (identity) activation function. This configuration allows the MLP to approximate continuous quality scores directly rather than discrete classes. The MLP model was particularly suitable for exploring whether neural architectures could generalize across heterogeneous feature sets, as WebRTC telemetry data often contains interdependent metrics such as bitrate adaptation, buffer delay, and packet retransmission counts.

The combination of these three algorithms enables a comprehensive evaluation of different modeling paradigms for QoE prediction. The regression tree provides a simple and interpretable reference, XGBoost introduces an ensemble approach with strong generalization capability, and the MLP explores deep, non-linear interactions between features. Together, these models form a balanced

methodological suite that supports both analytical understanding and high predictive accuracy, ensuring the robustness of the proposed framework for real-time QoE estimation.

2.4.2 | Implementation and optimization

The selected ML algorithms were implemented using the Python programming language, leveraging its open-source ecosystem and rich set of libraries for data analysis and model development. In particular, the *scikit-learn* library was used as the main framework for building, training, and evaluating the regression models. This library provides reliable implementations of most classical ML algorithms and offers efficient tools for model selection, cross-validation, and performance evaluation.

Before model training, all input features were normalized using the *MinMaxScaler* function from *scikit-learn*. This preprocessing step scaled each feature to a fixed range, typically between 0 and 1. Feature scaling was a necessary step because WebRTC session parameters have widely different numerical magnitudes (for example, packet counts may range in the thousands, while ratios or percentages remain below 1). Without normalization, features with larger numeric ranges could dominate the optimization process, causing biased model convergence. By scaling all features to a uniform range, the training process became more stable, and each parameter contributed equally to the prediction task.

To maximize model performance, an extensive process of *hyperparameter optimization* was performed as a first step. Hyperparameters control the behavior and flexibility of each ML algorithm (for example, tree depth in decision trees or learning rate in neural networks), and tuning them is critical to achieve high predictive accuracy. Poorly chosen hyperparameters can lead to underfitting or overfitting, whereas optimal values improve the generalization ability of the model and reduce prediction error.

The objective of the tuning phase was to find the hyperparameter configurations that minimized the RMSE and maximized the coefficient of determination (R^2). A lower RMSE indicates that the model's predictions are close to the subjective MOS ratings, while an R^2 value close to 1 demonstrates that the model explains most of the observed variance in the data.

To identify the optimal hyperparameters efficiently, two complementary search methods from the *scikit-learn* library were used: *GridSearchCV* and *RandomizedSearchCV*. These functions automate the process of evaluating different combinations of hyperparameter values using cross-validation to prevent overfitting.

- **GridSearchCV:** This method performs an exhaustive search over a predefined grid of possible hyperparameter values. Each combination is tested systematically, and the configuration yielding the best cross-validation performance is selected. Although this approach guarantees

finding the optimal combination within the tested grid, it can be computationally expensive when the search space is large.

- **RandomizedSearchCV:** This method randomly samples combinations of hyperparameters from specified distributions. While not exhaustive, it allows a broader exploration of the search space with significantly reduced computational cost. It is particularly useful when there are many hyperparameters or when prior knowledge about optimal ranges is limited.

Both search strategies were combined with k-fold cross-validation, with $k = 5$, to ensure robust model evaluation. In this procedure, the dataset was split into five equally sized folds. For each iteration, four folds (80% of the data) were used for training and one fold (20%) for validation. The process was repeated five times so that every sample was used for validation exactly once, and the average performance across folds was recorded. This technique improves the reliability of model evaluation by reducing variance associated with random data splits and by making better use of the limited number of test conditions available.

For each of the three ML algorithms described in Section 2.4.1, a comprehensive set of hyperparameters was explored. The investigated hyperparameters and the best-performing values (highlighted in bold) are reported in Table 2.3. These include, for example, the maximum depth of trees and minimum samples per leaf for the Regression Tree, the number of estimators and learning rate for XGBoost, and the number of layers, neurons, and learning rate for the MLP. The optimization process ensured that each model configuration achieved its best possible trade-off between bias and variance.

The combination of scaling, hyperparameter optimization, and cross-validation provided a strong foundation for the subsequent performance evaluation. By following a consistent and systematic procedure for all models, the comparison across algorithms and feature sets became fair and meaningful, allowing the analysis to focus on which modeling approach most effectively predicts the perceived QoE from WebRTC telemetry.

Table 2.3: The hyperparameters for the three ML methods; optimal values are highlighted in bold. MLP used the ADAM optimizer.

Regression tree		**XGBoost**			**MLP**		
Max dep.	**Num. leafs**	**Num. est.**	**Max dep.**	**Lrn. rate**	**Num. lay.**	**Neur.**	**Lrn. rate**
11	500	700	7	10^{-1}	1	8	10^{-1}
12	600	800	8	$\mathbf{10^{-2}}$	2	16	10^{-2}
13	**700**	**900**	**9**	10^{-3}	**3**	32	$\mathbf{10^{-3}}$
14	800	1000	10	10^{-4}	4	**64**	10^{-4}
15	900	1100	11	10^{-5}	5	128	10^{-5}

2.5 | Results

Figures 2.2 and 2.3 show the performance of the three regression models, each trained with the three feature datasets and optimized using the hyperparameters highlighted in bold in Table 2.3. Model performance is reported in terms of the coefficient of determination (R^2) and the RMSE. For both metrics, the results represent the mean values and standard deviations obtained from the 5-fold cross-validation procedure, where each fold provided an independent evaluation of model generalization. The error bars in the figures thus reflect the variability of the models' predictive stability across different training–validation splits.

Figure 2.2: Coefficient of determination R^2 results achieved by the 3 regression models when trained on the 3 features datasets.

A general observation from the results is that all three regression models achieved their best performance when trained with the feature set based on *pcc_params*, which includes the 24 WebRTC session parameters identified through the PCC analysis. Specifically, these models consistently achieved the highest R^2 values and the lowest RMSE scores compared to those trained on the other two feature sets. This finding demonstrates that the PCC-based statistical analysis effectively identified the most informative WebRTC session parameters for estimating end-user QoE. The parameters in this set capture the strongest linear relationships between network-level indicators and subjective quality ratings, confirming that features exhibiting consistent linear correlations with MOS contribute most to accurate QoE prediction. The complete list of these selected parameters is reported in Table 2.4.

By contrast, models trained using the *anova_params* feature set, which includes 34 parame-

Figure 2.3: RMSE results achieved by the 3 regression models when trained on the 3 features datasets.

ters deemed statistically significant according to the ANOVA analysis, achieved comparatively lower performance. This result suggests that the additional parameters identified by ANOVA, though statistically significant, may introduce redundancy or noise into the feature space. Instead of improving estimation accuracy, these extra features appear to slightly degrade performance, likely due to over-parameterization and increased model variance in such a small dataset. This behavior is consistent with the general principle that adding weakly correlated or redundant features can confuse the model and reduce its generalization capability.

When considering the *joint_params* feature set, which includes the 22 parameters common to both the PCC and ANOVA analyses, a further slight performance decrease is observed. This reduction can be attributed to the loss of two key parameters from the PCC-only subset, namely [4] and [5], which are among the features most directly related to frame transmission and packet reliability. Their exclusion reduces the model's ability to capture some of the key degradations that users perceive as quality drops. Nevertheless, the *joint_params* set remains valuable as it represents a more conservative feature subset, containing only those parameters that are both linearly correlated and statistically significant for QoE estimation.

While the overall trend holds across all regression models, the specific behaviors differ slightly among them. The Regression Tree model exhibited moderate performance across all feature sets, showing better results when trained with *anova_params* compared to *joint_params*. This can be explained by the model's limited ability to capture complex non-linear interactions between param-

[4] $RTCMediaStreamTrack_{s}ender_{h}ugeFramesSent$
[5] $RTCRemoteInboundRtpAudioStream_{p}acketsLost$

Table 2.4: The 24 *pcc_params*. Note: Dec. is the abbreviation for deceleration.

WebRTC session feature
RTCIceCandidatePair_bytesReceived
RTCInboundRTPAudioStream_concealedSamples
RTCInboundRTPAudioStream_concealmentEvents
RTCInboundRTPAudioStream_headerBytesReceived
RTCInboundRTPAudioStream_insertedSamplesForDec.
RTCInboundRTPAudioStream_jitterBufferEmittedCount
RTCInboundRTPAudioStream_packetsLost
RTCInboundRTPVideo_framesDecoded
RTCInboundRTPVideo_framesDecoded_s
RTCInboundRTPVideo_framesPerSecond
RTCInboundRTPVideo_framesReceived
RTCInboundRTPVideo_headerBytesReceived
RTCInboundRTPVideo_keyFramesDecoded
RTCInboundRTPVideo_pliCount
RTCInboundRTPVideo_totalSquaredInterFrameDelay
RTCMediaStreamTrack_receiver_concealedSamples
RTCMediaStreamTrack_receiver_concealmentEvents
RTCMediaStreamTrack_receiver_insertedSamplesForDec.
RTCMediaStreamTrack_sender_hugeFramesSent
RTCOutboundRTPVideoStream_hugeFramesSent
RTCOutboundRTPVideoStream_keyFramesEncoded
RTCOutboundRTPVideoStream_pliCount
RTCOutboundRTPVideoStream_qpSum_framesEncoded
RTCRemoteInboundRtpAudioStream_packetsLost

eters; the addition of a few extra ANOVA-selected features may have compensated for some missing relationships at the cost of slight overfitting.

The Extreme Gradient Boosting (XGBoost) model performed strongly in general and demonstrated relatively stable performance across the three feature sets. Its best results were achieved when trained with *anova_params*, where it reached the highest R^2 and lowest RMSE among the tree-based models. This improvement over the single Regression Tree is expected, as XGBoost leverages ensemble learning by combining multiple decision trees, each correcting the residual errors of its predecessors. This iterative refinement allows XGBoost to model more complex dependencies between features, providing a more accurate representation of the underlying QoE dynamics. The ensemble nature of XGBoost typically enhances robustness to noise and reduces the variance associated with single-tree predictors, which aligns well with the characteristics of the WebRTC telemetry data.

The Multi-Layer Perceptron (MLP), representing the neural network approach, achieved the best

overall performance among the three algorithms. When trained with both the *pcc_params* and *joint_params* feature sets, it produced the highest R^2 values and the lowest RMSE, outperforming the Regression Tree and XGBoost models. This indicates that the MLP's ability to capture complex, non-linear relationships between input features and QoE leads to superior predictive accuracy. Unlike tree-based models, which rely on threshold-based splits, the MLP learns smooth functional mappings by optimizing weights through gradient-based backpropagation. Such flexibility enables it to adapt to intricate relationships in the WebRTC feature space, where quality degradation may depend on subtle combinations of network metrics (for example, interactions between jitter variability and video frame loss).

It is worth noting, however, that neural networks like MLP do not guarantee convergence to a global minimum, and their performance depends strongly on initialization and training stability. Despite these challenges, the MLP in this study was able to converge effectively, achieving consistent results across folds. This robustness can be attributed to proper feature normalization, careful hyper-parameter optimization, and the relatively compact size of the dataset, which prevented overfitting.

In summary, the experimental results confirm that the feature selection method based on PCC produced the most compact and effective parameter set for predicting QoE in WebRTC-based audiovisual conversations. Among the tested regression algorithms, the MLP achieved the best balance between predictive accuracy and model stability, followed by XGBoost and Regression Tree. These findings demonstrate that even with a relatively small set of well-chosen features, it is possible to build a reliable, data-driven model capable of estimating user-perceived quality with strong alignment to subjective evaluations.

2.5.1 | Comparison with the state-of-the-art

To evaluate the effectiveness of the proposed QoE estimation models, we compared their performance with that of state-of-the-art approaches reported in the literature. To the best of the author's knowledge, the study in [12] is the only existing work that specifically proposed a regression-based QoE prediction model trained on features derived from the *webrtc-internals* telemetry data. In that study, a *Multiple Linear Regression (MLR)* model was trained using only four selected WebRTC session parameters, achieving a coefficient of determination $R^2 = 0.732$. This result provides a valuable baseline for evaluating subsequent developments in application-layer QoE estimation.

In contrast, all the models proposed in this work achieved substantially higher performance in terms of both R^2 and RMSE. When trained with the parameters identified through the PCC-based feature selection, the proposed MLP model achieved $R^2 = 0.852$ and RMSE $= 0.282$, clearly outperforming the state-of-the-art MLR model by a significant margin. Even the simpler Regression Tree and XGBoost models surpassed the previously reported results, highlighting the advantage of using a more rigorous statistical preprocessing pipeline combined with non-linear regression techniques.

These findings confirm that the combination of statistical feature selection (PCC and ANOVA) and modern regression methods yields a significant improvement in predictive accuracy.

Furthermore, to provide an additional benchmark, we applied the *DQX model* proposed in [27], which estimates audiovisual communication quality using a set of multiple network-level variables. We computed the estimated MOS values for the 15 test conditions listed in Table 2.2 and compared them with the subjective MOS obtained from our experiment. The resulting RMSE between the predicted and actual MOS values was 0.672, which is more than twice the RMSE achieved by any of the proposed models. This comparison reinforces the effectiveness of the presented approach in predicting user-perceived QoE, showing that models trained directly on application-layer telemetry can capture perceptually meaningful effects more accurately than traditional network-level QoE models.

Overall, these results demonstrate that the proposed methodology not only advances the state-of-the-art in WebRTC QoE prediction but also establishes a strong foundation for future models targeting real-time, data-driven QoE estimation in communication systems.

2.6 | Key findings and contributions

The proposed analysis and modeling framework for WebRTC-based audiovisual conversations provides several important insights into how application-layer telemetry can be used to predict user-perceived QoE. The key findings and contributions of this study are summarized as follows:

- A complete dataset was created by collecting *webrtc-internals* telemetry from controlled audiovisual conversations under fifteen test conditions combining delay, jitter, and packet loss. This dataset links objective WebRTC statistics with subjective MOS, enabling reproducible QoE modeling and analysis.
- A statistical feature analysis pipeline was designed to transform raw telemetry into structured parameter statistics and select the most informative features using the PCC and ANOVA. This approach ensured interpretability and reduced feature redundancy while preserving perceptually relevant information.
- The PCC-based feature selection method proved to be the most effective for QoE estimation. The identified 24 parameters (*pcc_params*) captured the strongest correlations with MOS and yielded the highest prediction accuracy across all tested models.
- Three regression algorithms—Regression Tree, XGBoost, and Multi-Layer Perceptron (MLP)—were implemented, optimized, and evaluated using a 5-fold cross-validation procedure. Among them, the MLP achieved the best performance with $R^2 = 0.852$ and RMSE $= 0.282$, demonstrating its ability to model non-linear relationships between application-level features and perceived quality.

- A comprehensive hyperparameter optimization was performed using *GridSearchCV* and *RandomizedSearchCV*, ensuring that each algorithm achieved its best accuracy–efficiency trade-off. The evaluation confirmed the reliability and reproducibility of the models across different feature subsets.

- The proposed models significantly outperformed existing state-of-the-art approaches, including the Multiple Linear Regression (MLR) model in [12] and the DQX model in [27]. The achieved improvements highlight the advantage of using feature sets extracted directly from application-level telemetry rather than relying on traditional network-layer metrics.

- The study demonstrated that careful statistical preprocessing and feature selection are as critical as the choice of the regression model itself. Reducing noise and focusing on the most discriminative parameters resulted in robust models that generalize well across various network impairment scenarios.

- Finally, this work provides a transferable methodology for QoE estimation in real-time communication systems. The proposed statistical machine learning framework can serve as a blueprint for future studies on QoE prediction in other multimedia domains, such as adaptive video streaming, VoIP, and immersive communication.

These findings collectively validate the hypothesis that objective QoE estimation can be achieved directly from application-layer data without requiring access to low-level network information.

Together, these findings reinforce one of the key objectives of this dissertation: establishing a data-driven and interpretable foundation for service-level QoE estimation using accessible, real-world telemetry data. This foundation enables the development of increasingly adaptive, generalizable, and deployment-ready QoE models in the subsequent chapters. The chapter establishes a solid foundation for extending the methodology to other multimedia services in the subsequent chapters of this dissertation. These results provide the first empirical validation of the proposed thesis framework, proving that objective QoE modeling is feasible using application-level data alone. The following chapter extends this methodology to *HTTP adaptive video streaming*, where temporal dynamics such as start-up delay, quality switches, and playback stalls become the primary determinants of perceived quality.

Building upon the signal-design methodology and feature analysis framework introduced in this chapter, the next study extends QoE modeling to adaptive video streaming, where perceived quality varies dynamically throughout the viewing session. While the WebRTC model focused on static session-level estimation, Chapter 3 introduces temporal deep learning methods capable of capturing sequential variations in quality over time, marking the next step toward context-aware and temporally adaptive QoE prediction.

2.7 | Conclusion

This study presented a comprehensive workflow for identifying and modeling the most relevant WebRTC session parameters that influence user-perceived QoE in audiovisual conversations affected by network distortions. The proposed methodology combined descriptive statistical analysis, correlation and variance-based feature selection, and data-driven regression modeling to create a reliable and interpretable QoE prediction pipeline. By systematically integrating these steps, we demonstrated how subjective perception data and objective telemetry logs can be jointly leveraged to build accurate and generalizable QoE estimators.

The feature selection process, based on the PCC and ANOVA, proved essential in isolating parameters that have a consistent and statistically significant relationship with perceived quality. The experiments demonstrated that the parameters identified through PCC yielded the most compact and effective feature set for QoE estimation, resulting in superior prediction performance across all regression algorithms. Among the three trained models (Regression Tree, XGBoost, and MLP), the Multi-Layer Perceptron achieved the best results, with $R^2 = 0.852$ and RMSE $= 0.282$. This performance represents a substantial improvement over existing state-of-the-art models, including the MLR method from [12] and the DQX model from [27].

The results also confirmed that conducting a rigorous statistical analysis before training ML-based estimators significantly enhances model robustness and generalization. By reducing noise and eliminating redundant features, the proposed feature selection pipeline enables ML algorithms to focus on the most informative signals that directly correspond to user perception.

In summary, this study provides empirical evidence that objective QoE estimation for WebRTC applications can be effectively achieved through data-driven approaches grounded in statistical analysis and ML modeling. The combination of PCC and ANOVA-based feature selection with efficient regression models forms a generalizable methodology that can be extended to other multimedia services. In future work, we plan to apply the same framework to additional application domains such as adaptive video streaming and VoIP communication. This extension will enable a more comprehensive validation of the proposed methodology and contribute to the development of unified, real-time QoE monitoring tools for interactive multimedia systems.

3

Temporal Modeling for Adaptive Video Streaming QoE (Streaming QoE)

3.1 | Introduction

Following the research directions outlined in Chapter 1, this chapter deepens the service-level layer by focusing on the temporal dynamics of adaptive video streaming to achieve the dissertation's Objective 2: Temporal modeling for adaptive streaming QoE. While the previous chapter addressed QoE prediction from application-layer telemetry in conversational scenarios (WebRTC), adaptive video streaming introduces a more complex and dynamic context, where the perceived quality evolves continuously over time due to bitrate adaptation, stalling, and resolution switching events. This aligns directly with the dissertation's overarching aim of developing data-driven, context-aware QoE estimators that can generalize across different service types and time-varying conditions.

The central hypothesis explored in this chapter is that the temporal evolution of streaming impairments carries valuable information about user-perceived quality, operationalizing the second pillar of the dissertation: temporal modeling, which seeks to represent how time-dependent distortions influence human-perceived quality. By explicitly modeling these temporal patterns through deep learning architectures capable of sequence understanding, such as recurrent and attention-based networks, the study seeks to capture long-term dependencies between network events and perceived QoE, thereby addressing one of the key challenges highlighted in Chapter 1: the dynamic and time-varying nature of multimedia experience.

In today's rapidly evolving digital landscape, the seamless delivery of multimedia content has become an essential component of our daily lives. From video-on-demand platforms and live streaming services to online gaming and virtual communication tools, the demand for high-quality multimedia experiences has grown exponentially. However, guaranteeing a consistent and satisfying user experience across such dynamic and heterogeneous environments remains a significant challenge for

network operators and service providers. QoE encompasses the user's holistic perception of the performance and usability of a service, integrating technical quality aspects with cognitive and emotional responses. It reflects how end users subjectively assess the quality of multimedia services based on audiovisual fidelity, responsiveness, interactivity, and overall satisfaction [29].

In recent years, increasing research attention has been devoted to quantifying and predicting user-perceived QoE both through subjective and objective methodologies. Subjective approaches rely on user studies, where participants rate their experiences following standardized procedures, such as the ACR method. Although these studies are invaluable for understanding human perception, they are costly, time-consuming, and difficult to scale. Conversely, objective QoE models aim to computationally estimate perceived quality using measurable system or media parameters without the need for direct human feedback. Such models are often trained or calibrated using ground-truth subjective data collected under controlled laboratory conditions to ensure their perceptual relevance. Nevertheless, a persistent limitation lies in the variability of users' perceptual judgments, which are influenced by numerous factors, including personal expectations, content familiarity, cultural background, and prior exposure to impairments [30].

Modeling these subjective variations remains a fundamental challenge in QoE research. Therefore, an effective QoE estimation model must not only map the technical performance of the system to perceived quality but also generalize across diverse network conditions, device types, and usage contexts. Traditional parametric or data-driven methods, such as autoregressive models [31], deep neural networks [32], and ensemble learning approaches [33], have achieved promising results in predicting QoE for specific services. However, they often lack robustness when applied to unseen contexts or new streaming conditions. These models are typically sensitive to the range of parameters used during training, which limits their adaptability and generalization capabilities. Consequently, QoE estimation models may suffer from overfitting and fail to accurately represent user experience in real-world, dynamic environments.

To overcome these challenges, recent research trends have turned toward architectures capable of modeling complex temporal dependencies and heterogeneous data sources. Among these, the transformer deep learning architecture [34] has emerged as a powerful solution. Initially proposed for natural language processing (NLP), transformers have achieved state-of-the-art performance in a wide range of sequential learning tasks, including sentiment analysis [35], machine translation, and large-scale language modeling [36]. Unlike conventional recurrent neural networks (RNNs) or Long Short-Term Memory (LSTM) models, transformers rely entirely on the *multi-head attention mechanism*, which enables them to capture both local and global dependencies within a sequence. By dynamically attending to the most relevant parts of the input data, the transformer can model intricate interactions and contextual relationships between variables that occur at different temporal distances.

These properties make transformers particularly well-suited for the QoE estimation problem in

video streaming scenarios, where users' experiences are influenced by a sequence of time-varying factors such as video bitrate fluctuations, rebuffering events, resolution changes, and playback delays. The ability to learn long-range dependencies allows the transformer to effectively interpret the temporal dynamics of streaming sessions, capturing how impairments at different time instants cumulatively impact the overall user perception. Furthermore, the attention mechanism provides interpretability by revealing which parts of the sequence contribute most to the final QoE prediction, thereby offering insights into the perceptual relevance of different streaming events. In addition, since QoE perception may differ across viewing devices and usage contexts, the transformer's attention mechanism can implicitly learn these contextual variations, enhancing the robustness of the model across diverse operating conditions.

In this chapter, we investigate the application of transformer learning to QoE modeling for adaptive video streaming services. Specifically, we design a novel transformer-based QoE estimation model capable of processing sequential input data that encapsulates both video-related and session-level key performance indicators (KPIs). To achieve this, we propose a systematic workflow that includes data collection, encoding, and sequentialization, ensuring that the information is represented in a form suitable for the transformer's learning process.

In a nutshell, we design a complete workflow to support sequence learning:

1. Encode and sequentialize session KPIs into per-segment tokens that reflect start-up delay, bitrate, and resolution levels, quality switches, rebuffering onsets, and durations.

2. Normalize and pad variable-length sessions with appropriate masking.

3. Train a compact transformer with a regression head to predict MOS. The workflow is instantiated on two open datasets from the ITU-T P.1203 standardization process [37], covering 82 impaired sequences and subjective ratings on both PC and mobile devices.

4. Evaluation follows the metrics used throughout the thesis, such as Root Mean Square Error (RMSE), Pearson correlation coefficient (PCC), and Spearman correlation coefficient (SCC) — and includes cross-device tests to probe robustness when training and testing devices differ.

Each HRC includes diverse combinations of quality switches between different resolutions and bitrates, initial loading delay, and stalling events.

Experimental results demonstrate that the proposed approach consistently outperforms the ITU-T P.1203 model in QoE prediction accuracy for both device types. Moreover, it exhibits strong cross-device generalization capability, meaning that a model trained on data from one device (e.g., mobile) can effectively estimate the QoE of videos viewed on another device (e.g., PC). These findings indicate that transformer learning represents a promising direction for developing robust and context-adaptive QoE estimation models.

The remainder of this chapter is structured as follows. Section 3.2 reviews the related works and research gaps on QoE modeling and the use of deep learning architectures in this field. Section 3.3 presents the methodology and workflow process behind the proposed transformer-based QoE model. Section 3.4 details the implementation steps and training configuration, while Section 3.5 reports and discusses the experimental results. Section 3.6 presents the key findings and contributions of the chapter, and finally, Section 3.7 concludes the chapter with perspectives for future work.

3.2 | Background and Related Work

Objective QoE models aim to predict the perceived quality of multimedia applications and content using measurable technical parameters rather than direct subjective feedback from users [38]. Over the last decade, several models have been developed for video streaming services to bridge the gap between system-level performance metrics and end-user perception [39]. Among these, the ITU-T Recommendation P.1203 represents the most comprehensive and widely adopted framework for estimating the QoE of adaptive audiovisual streaming services [37].

The P.1203 model follows a scalable design that accounts for spatial and temporal degradations across four operational modes, which correspond to different levels of available input data and encryption constraints. *Mode 0* operates under the most restricted conditions, requiring minimal access to information such as codec type, target bitrate, resolution, frame rate, and segment durations, thereby maintaining low computational cost. In contrast, *Mode 3* operates with full access to the video bitstream, enabling a more detailed quality estimation at the expense of higher complexity. Each mode combines the outputs of several submodules—responsible for audiovisual quality, stalling, and integration—into a final prediction of session-level quality expressed on the MOS scale ranging from 1 (Bad) to 5 (Excellent). Despite its strong standardization value and scalability, P.1203's fixed analytical formulation limits its flexibility in dynamically changing environments and in cases where the distribution of input parameters differs from those used during model calibration.

In recent years, a growing body of research has focused on leveraging ML methods to enhance the accuracy and generalization of QoE prediction models [40]. ML approaches can learn complex, non-linear relationships between KPIs and user-perceived quality, offering a data-driven alternative to fixed parametric functions. For instance, the nonlinear Auto-Regressive model with exogenous variables (NARX) proposed in [31] was designed to capture temporal dependencies in QoE by processing previous subjective quality measurements along with contextual features such as playback quality, rebuffering traces, and memory effects from prior events. The NARX model demonstrated that accounting for temporal correlations between impairments significantly improves the prediction of time-varying QoE.

Similarly, the authors of [41] trained a large set of ML models, including regression and classi-

fication algorithms, to estimate QoE and related service parameters such as startup delay, bitrate, and resolution under diverse network conditions. Their findings revealed that model performance strongly depends on dataset diversity and highlighted the importance of cross-dataset generalization, where models trained on one dataset are evaluated on others with different characteristics. This line of research emphasized the need for robust learning strategies capable of handling unseen streaming conditions, user devices, and service types.

Deep learning approaches have further advanced the field by automatically learning abstract representations from raw or high-dimensional features. The DeepQoE framework proposed in [42] integrates multi-modal data, such as video features and textual metadata, using 3D Convolutional Neural Networks (CNNs) and word embeddings. These representations are fused and fed into a neural network to produce a learned embedding space suitable for both classification and regression of QoE. DeepQoE exemplified the potential of end-to-end learning pipelines in capturing perceptual factors beyond simple bitrate or stall duration features.

In [32], the authors presented DeSVQ, a deep learning model combining CNN and Long Short-Term Memory (LSTM) networks to model both spatial and temporal dependencies affecting QoE. The CNN layers capture frame-level visual degradations, while the LSTM layers model sequential dependencies that arise from temporal quality fluctuations. The integration of these two feature processing stages significantly improved correlation metrics compared to either architecture alone, highlighting the importance of jointly modeling time and content effects.

Despite these advances, most deep learning models exhibit limitations in adapting to highly diverse and dynamic contexts, such as heterogeneous network conditions, device capabilities, and user expectations. CNN and LSTM-based models, while powerful, often struggle with scalability and long-range dependencies because recurrent connections inherently constrain temporal learning to short or medium-length sequences. This has motivated the exploration of transformer architectures as a more flexible and efficient alternative.

Transformers have become a state-of-the-art solution for sequence modeling and transduction tasks, including natural language processing, speech recognition, and machine translation, where they have surpassed recurrent architectures such as LSTMs and GRUs in both accuracy and computational efficiency [36]. Their self-attention mechanism allows direct modeling of relationships between any two points in a sequence, enabling the capture of global temporal dependencies. This architecture's ability to weigh the relevance of each input element makes it particularly suitable for QoE estimation, where user perception depends on the cumulative and contextual influence of temporally distributed events (e.g., stalling, switching, bitrate changes).

In recent literature, transformer-based models have been increasingly applied beyond text processing to address multimedia and network management problems. The FlowFormers algorithm introduced in [43] represents one of the first attempts to apply transformer encoders to network traffic analysis. By exploiting self-attention to capture dependencies between packet sequences, FlowForm-

ers achieved superior performance in real-time network flow classification tasks, demonstrating the capability of transformers to handle sequential network data more effectively than CNN and LSTM counterparts.

Another relevant study [44] applied transformers to user satisfaction prediction in proactive dialogue systems. In this case, the model processed both structured numerical inputs and textual dialogue turns, learning dependencies between current and past user interactions. The results indicated a 19% improvement in satisfaction prediction accuracy and a 2.3% increase in overall user experience, showcasing the model's strength in multimodal temporal reasoning.

In the domain of visual quality assessment, hierarchical transformer architectures have recently been proposed to handle multi-scale temporal dependencies in video perception. The Hierarchical Transformer model introduced in [45] combines two transformer modules: one operating at the clip level to extract short-term embeddings and another at the frame level to derive a holistic video-level representation. The final quality score is predicted using a linear regressor on the aggregated embeddings. This approach outperformed several conventional deep learning methods for video quality assessment, reinforcing the transformer's capacity to capture fine-grained and long-term temporal structures. However, it focuses primarily on visual fidelity and does not explicitly incorporate streaming impairments such as delay or rebuffering, which are crucial determinants of QoE in adaptive video streaming.

These prior studies collectively demonstrate the transformative potential of self-attention mechanisms in modeling sequential data where temporal dependencies play a central role. However, their application to end-to-end QoE estimation for streaming services remains underexplored. Most existing transformer-based methods either concentrate on visual quality or operate in non-streaming contexts where temporal degradations are limited. To the best of the authors' knowledge, the present work constitutes the first study that employs transformer learning to jointly model video quality and session-related factors for QoE estimation. By encoding and sequentializing key performance indicators such as resolution, bitrate, stalling, and delay, the proposed model leverages the transformer's attention mechanism to learn the complex, time-dependent interactions that shape user-perceived quality. The output is expressed in the MOS scale, enabling direct comparison with standardized models such as ITU-T P.1203 and facilitating integration into broader QoE evaluation frameworks.

3.2.1 | Research Gap and Contribution

A systematic analysis of state-of-the-art approaches reveals persistent challenges in modeling the temporal and contextual aspects of user-perceived QoE. From the reviewed literature in Section 3.2, several research gaps can be identified:

1. Limited modeling of temporal dependencies. Most existing QoE estimation models rely on

aggregated session statistics or frame-based features, which disregard the temporal order and duration of playback impairments. This limits their ability to reflect how sequential events, such as the position and frequency of stalls or quality switches, affect perceived experience.

2. Lack of integration between video quality and session dynamics. Prior studies often focus on visual quality assessment alone, neglecting session-related parameters such as rebuffering, delay, and adaptation behavior that strongly influence user satisfaction in adaptive streaming environments.

3. Restricted adaptability to heterogeneous contexts. Many deep learning models are trained and tested under homogeneous conditions (single dataset or device type) and fail to generalize when network environments, device characteristics, or content types vary.

4. Underexplored use of attention-based architectures for QoE estimation. Although transformers have achieved remarkable success in sequence modeling tasks across other domains, their application to QoE prediction, particularly for adaptive video streaming, remains largely unexplored.

5. Limited interpretability of deep learning QoE models. Conventional CNN or LSTM-based models operate as black boxes, offering little insight into which temporal events most influence the predicted QoE, making them less suitable for diagnostic or optimization purposes.

Overall, this chapter addresses these gaps by introducing a sequence-aware transformer-based modeling framework for video streaming QoE estimation and contributes to the dissertation's goal of developing temporal, data-driven QoE models that capture both instantaneous and evolving aspects of user experience. The proposed approach represents each streaming session as a per-second sequence of KPIs capturing bitrate level, delay, and stalling events, thus preserving the temporal evolution of playback quality. By leveraging the multi-head attention mechanism, the model learns long-range dependencies between impairments and identifies which temporal segments are most influential for user perception.

Extensive experiments on two open ITU-T P.1203 datasets demonstrate that this architecture achieves higher accuracy and stronger cross-device generalization than the standardized ITU-T P.1203 model. Furthermore, the attention weights provide an interpretable representation of QoE dynamics, offering valuable insights for streaming optimization and adaptive service management. This contribution extends the thesis's broader goal of developing data-driven, generalizable, and explainable QoE estimation models applicable across diverse multimedia services.

3.3 | Methodology

The variability and dynamic nature of the QoE are deeply rooted in human subjectivity and are shaped by multiple interacting factors, including system-level performance, human perception, and the surrounding context [29]. Each of these dimensions contributes differently to how users perceive and evaluate the quality of a multimedia service. As a result, modeling QoE represents a multi-faceted challenge that requires capturing both objective indicators of service delivery and their subjective interpretation by users.

Traditional objective QoE models generally define a deterministic or statistical relationship between a selected subset of *QoE influence factors* and the perceived quality for a specific application domain. In the case of adaptive video streaming, these influence factors are typically derived from session-level parameters, such as the resolution or bitrate of delivered segments, the frequency and duration of rebuffering events, and the magnitude of quality switches between segments. Although such parameters effectively capture observable impairments, the predictive accuracy of classical models is often constrained by the specific configuration and value ranges represented in the training data. Consequently, models that rely on static or aggregated metrics struggle to generalize to dynamic or previously unseen streaming conditions.

The ITU-T Recommendation P.1203 has established an important step forward by providing a scalable, modular, and standardized framework for QoE estimation. The P.1203 model can operate at multiple levels of input granularity, ranging from low-level bitstream features to high-level playback data, and can be fed with per-second measurements of video and session-related variables [46]. This temporal input capability allows P.1203 to capture fine-grained variations in playback behavior; however, its analytical formulation still limits flexibility when learning complex, non-linear dependencies between time-varying impairments and perceived quality.

Inspired by the sequential modeling concept underlying P.1203, this work proposes a **transformer-based QoE estimation model** that leverages per-second sequential encoding of video and session-level information. The transformer architecture is inherently designed to process ordered sequences and to capture dependencies across distant temporal points through its *multi-head self-attention mechanism*. This mechanism enables the model to assign varying levels of importance to each time step in the sequence, allowing it to focus on the most perceptually relevant segments when estimating QoE. By representing session data as sequences of KPIs, the model learns to interpret not only the individual impairments but also their cumulative and contextual impact on user experience.

The core rationale for encoding KPIs as sequential data lies in the hypothesis that user-perceived quality is not determined by isolated events but by the temporal structure of impairments. For example, an early stall followed by smooth playback may be perceived differently from a late stall of the same duration, even if both sessions share identical average statistics. By capturing data at a per-second resolution, the model can identify patterns, transitions, and dependencies that would

be otherwise lost in coarser temporal aggregations. Furthermore, by directly feeding this sequential data into the transformer, the model naturally exploits its capacity to learn high-order temporal relationships and non-linear interactions among multiple KPIs.

To guide the investigation, three main research questions are formulated:

- **Q1:** *What performance can the transformer-based model achieve in predicting the QoE of videos with different lengths and diverse impairment patterns (e.g., quality switching, delay, and stalling)?*
- **Q2:** *How effectively can the transformer-based model predict the QoE of videos viewed on different devices, such as mobile and PC platforms?*
- **Q3:** *How does the performance of the proposed transformer-based model compare to the standardized ITU-T P.1203 model in terms of accuracy and generalization?*

Figure 3.1 illustrates the proposed workflow for data processing and model training. The process begins with the collection of KPIs at one-second intervals during each video streaming session. These KPIs are extracted from session logs and include information on video bitrate, resolution, quality level, segment duration, occurrence of initial delay, and stalling events. The data is then **encoded** into a numerical representation that captures both the categorical and continuous attributes of each second, forming a sequence suitable for input to the transformer encoder.

After encoding, the data undergoes a **sequentialization** step, where the per-second tuples are ordered chronologically to form a temporal stream of KPIs. This sequence effectively represents the evolution of playback quality throughout the session, enabling the model to learn dependencies between earlier and later events. Additional preprocessing steps, such as normalization, padding, and masking, are applied to standardize sequence lengths and ensure compatibility with the model's input requirements.

In parallel with the KPI collection, the overall video quality perceived by users is obtained through controlled subjective experiments. Each test sequence is rated by a pool of participants, and their individual scores are averaged to compute the MOS for both PC and mobile viewing conditions. The MOS serves as the ground-truth target for model training and evaluation, establishing a direct mapping between the temporal sequence of KPIs and the perceived session quality.

During training, the transformer learns to minimize the prediction error between the estimated QoE values and the ground-truth MOS. The attention mechanism adaptively identifies which time steps contribute most to the predicted quality, providing interpretability and highlighting the perceptual relevance of specific events such as stalls or quality drops. The model is designed to handle sessions of varying duration and to generalize across devices and datasets by relying on shared temporal patterns rather than device-specific distributions.

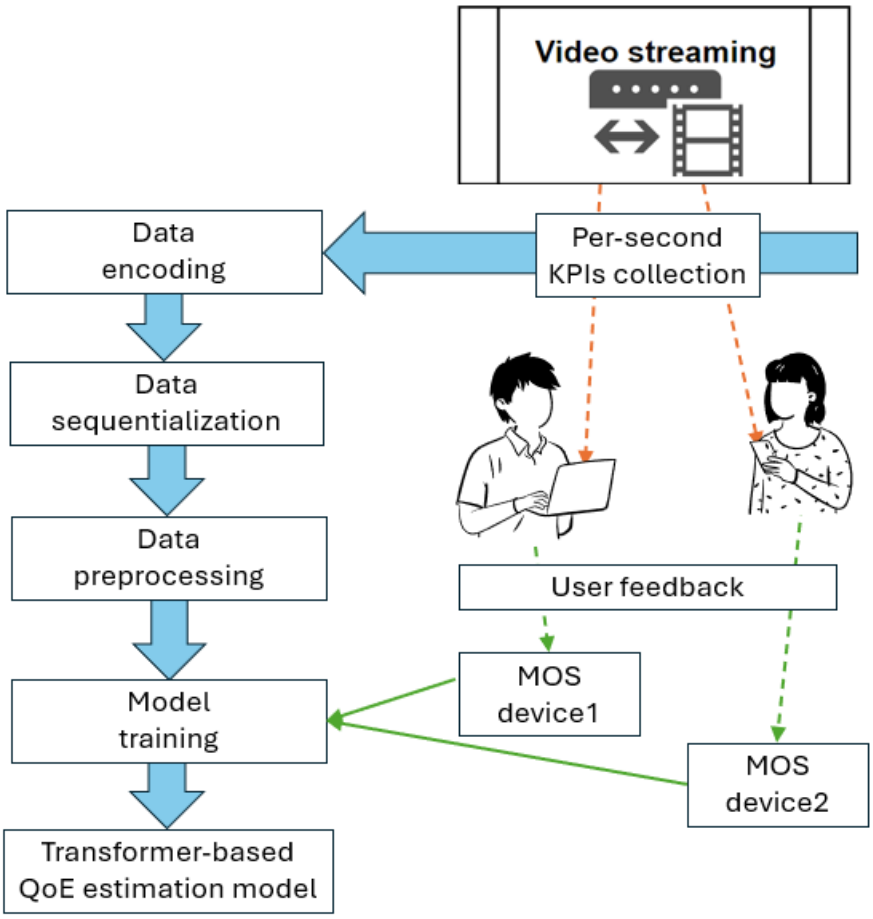

Figure 3.1: The proposed workflow process.

In summary, the proposed methodology combines sequential encoding of video KPIs with transformer learning to build a robust QoE estimation framework. The final output is a trained transformer-based model capable of predicting the perceived QoE of video sequences streamed over the Internet, expressed in the MOS scale. The subsequent sections describe the detailed implementation of the model architecture, the dataset characteristics, and the comparative evaluation with the ITU-T P.1203 benchmark.

3.4 | Model Implementation

3.4.1 | Datasets

We consider two open datasets[1] released within the ITU-T P.1203 standardization procedure (P.NATS) [47]. Following the original nomenclature [47], they are denoted as *TR04* and *TR06*. Both datasets were designed to probe the impact of HAS-typical network and playback conditions referred to as Hypothetical Reference Circuits (HRCs) on user-perceived quality. Each HRC defines a specific combination of quality switching behavior between resolutions and bitrates, the presence or absence of an initial loading delay, and one or more stalling events. TR04 comprises 20 HRCs applied to 3 distinct contents, each 60 seconds in duration, resulting in 60 test sequences. TR06 comprises 11 HRCs applied to 2 contents, each 180 seconds in duration, resulting in 22 test sequences. In total, the corpus contains 82 impaired sequences that represent a diverse set of temporal impairment patterns and content types. In the original subjective study [47], each sequence was viewed and rated on both PC and mobile devices under controlled conditions. Thus, for every sequence there are two target labels, MOS_{PC} and MOS_{M}, both reported on the ACR scale from 1 (Bad) to 5 (Excellent). Averaging across subjects yields continuous MOS targets within $[1, 5]$. This setup is well aligned with our goals for sequence-aware QoE estimation. The two datasets differ in content duration and impairment configurations, which allows us to examine model behavior across short and long sessions and across heterogeneous device contexts. The presence of MOS labels on both devices enables same-device and cross-device evaluations using a common ground-truth scale.

3.4.2 | Data encoding and sequentialization

As discussed in Section 3.2, transformer architectures are specifically designed to operate on ordered sequences, making them well-suited for representing the temporal structure of streaming sessions. Consequently, the raw data contained in the ITU-T P.1203 datasets was transformed into sequential inputs that encode how the viewing conditions evolve over time. This procedure consists of two main stages, including data encoding and data sequentialization.

In the data encoding phase, the goal is to create a compact numerical representation of the KPIs describing each second of playback. For every test video sequence, a tuple was generated for each second, defined as:

$$t^{d}_{v,n} = (Q, D, S), \tag{3.1}$$

where d identifies the dataset (TR04 or TR06), v denotes the index of the test video sequence, and n represents the current second of the video, ranging from $\{1, ..., N^{d}_{v}\}$. Each tuple, therefore, encapsulates the information observed at that particular instant. The variable Q corresponds to the video quality level at that second, represented by the bitrate–resolution pair reported in Table 3.1.

[1] `https://github.com/itu-p1203/open-dataset`

The variable D is a binary flag indicating whether an initial delay is present ($D = 1$ for a delayed frame, 0 otherwise), while S denotes the occurrence of a stalling event ($S = 1$ for a stall, 0 otherwise).

Table 3.1: Quality levels of the test video sequences.

Q	***Bitrate (kbps)***	***Resolution - height (px)***
7	10000	1080
6	2500	1080
4	500	480
2	150	240

This per-second encoding captures three major QoE-affecting factors, including quality, delay, and stall in a unified format. Compared with aggregated statistics such as total stall duration or average bitrate, this second-level representation retains fine-grained temporal information that allows the model to learn how the position, frequency, and duration of impairments influence perception. For example, a stall of two seconds occurring in the first few seconds of playback may be perceived differently from the same stall occurring near the end of the session. Such subtle effects can only be identified when the temporal order of events is explicitly modeled.

The sequentialization stage then concatenates the encoded tuples in chronological order to form the complete input sequence for each video session:

$$SQ_v^d = [t_{v,1}^d \oplus t_{v,2}^d \oplus \ldots \oplus t_{v,N_v^d}^d]. \tag{3.2}$$

where $\oplus$ denotes concatenation. Each element $t_{v,n}^d$ becomes a token in the sequence that will be processed by the transformer encoder. This operation produces a structured timeline of playback behavior, preserving not only the occurrence of events but also their relative position in the session. Through this sequential structure, the model can distinguish, for instance, between a single long stall and multiple shorter stalls or between gradual and abrupt quality transitions—differences that often lead to distinct subjective impressions.

For every sequence SQ_v^d, two corresponding target values are available, including the MOS obtained from subjective evaluations on PC devices (MOS_{PC}) and on mobile devices (MOS_M). These labels define the ground truth used during model training and validation. Maintaining both device-specific targets allows us to perform independent device-wise experiments as well as cross-device analyses, where a model trained on one device type is tested on the other. This dual labeling also enables investigation into whether the model can capture perceptual differences attributable to display size, viewing distance, or user interaction patterns.

Overall, this encoding–sequentialization pipeline converts raw playback traces into temporally ordered numerical sequences that serve as the foundation for the proposed transformer-based QoE estimation framework. By treating each second of a session as a token containing (Q, D, S), the

model can exploit the self-attention mechanism to learn which moments in the timeline contribute most to perceived quality, achieving a more context-aware and interpretable estimation process.

3.4.3 | Data preprocessing

Transformers require a fixed input length per batch. Since the actual duration of a session may exceed the nominal content duration due to initial delay or buffering, we standardize the sequence length by padding shorter sequences. Concretely, we apply zero-padding to SQ_v^d up to the maximum length observed in the dataset, adding tuples with $Q = 0$, $D = 0$, $S = 0$. During training and inference, an attention mask is applied so that padded positions do not contribute to the self-attention computation or the subsequent pooling. A second challenge concerns label imbalance. In many QoE datasets, MOS distributions are not uniform across the [1, 5] range, which can bias regression models toward frequent regions and degrade performance at the tails. To mitigate this, we adopt the SMOGN technique for imbalanced regression [48], which leverages the SmoteR interpolation strategy [49]. SMOGN adaptively oversamples underrepresented regions by interpolating nearby samples in feature-label space while reducing the risk of generating unrealistic examples. Using a Random Search procedure over augmentation factors, we found that tripling the size of the training set (we explored factors from 2 to 10) provided the best trade-off between bias reduction and overfitting, leading to the strongest QoE estimation performance in our experiments. Finally, numerical fields are normalized to stabilize optimization. Categorical quality levels Q are handled consistently with Table 3.1 (e.g., encoded as discrete levels mapped to their bitrate-resolution semantics), while binary indicators D and S are kept in $\{0, 1\}$. The same preprocessing pipeline is applied consistently across training and validation splits to avoid leakage.

3.4.4 | Proposed transformer-based model

As anticipated in Section 3.3, the novelty of transformer-based deep learning architectures lies primarily in the multi-head attention mechanism within the encoder block. This component allows the network to attend to different segments of the input sequence simultaneously, learning correlations and dependencies among temporally distant events. Unlike traditional recurrent architectures such as LSTM or GRU, which process data sequentially and may suffer from vanishing gradients, transformers process the entire sequence in parallel. This enables them to capture both short and long-range temporal dependencies with higher computational efficiency and improved representational power.

Each encoder block applies a sequence of self-attention and feed-forward operations, repeated N times, to progressively refine the representation of the input data. Through these stacked layers, the model builds hierarchical abstractions that describe how variations in the input parameters, such as bitrate changes, stalling, or delay, interact over time to influence user perception. The redundancy

resulting from repeated encoding steps allows the model to reinforce stable correlations and smooth out noise, leading to more consistent and accurate predictions of QoE. In this work, these architectural properties are leveraged to perform regression on MOS with high robustness across both content and device types.

The proposed transformer-based network, schematically depicted in Fig. 3.2, is composed of four main building blocks: the Encoder block, the GlobalAveragePooling1D layer, the Multi-Layer Perceptron (MLP), and the Regression head. The encoder block (highlighted by a dashed outline) is the core of the model and contains two key modules: the Multi-Head Attention (MHA) module and the Feed-Forward module. The MHA module is responsible for modeling dependencies among tokens in the sequence. It includes a normalization layer to stabilize learning, followed by an attention layer consisting of four attention heads, each of size 256. Each attention head learns an independent mapping of the input sequence, focusing on different aspects of the data; for instance, one head may learn to emphasize stalling-related segments, while another captures bitrate fluctuation patterns. The parallel combination of multiple heads enhances the model's ability to capture diverse relationships within the same sequence. A dropout layer with a dropping rate of 0.15 is applied after attention to mitigate overfitting. The module concludes with a residual or skip connection that adds the input tensor to the output, allowing gradients to propagate unimpeded through the network and facilitating stable optimization across layers. Following the attention mechanism, the Feed-Forward module performs non-linear transformations to refine the contextualized token representations. It includes a normalization layer, two one-dimensional convolutional layers (Conv1D), each with eight filters and a kernel size of one, and a dropout layer with a rate of 0.15. The Conv1D layers act as lightweight fully connected transformations applied to each token, helping the model to project the attention outputs into a higher-dimensional feature space.

This design captures non-linear relationships that are not directly modeled by attention alone, thereby enhancing the expressive capacity of the encoder. The feed-forward module also employs a skip connection to preserve essential features and ensure that low-level information remains accessible throughout the depth of the network [50]. The encoder block described above is stacked N times, where $N = 4$ was determined to be the optimal configuration after extensive experimentation. Increasing N beyond this value offered negligible performance improvement while leading to higher computational cost and risk of overfitting, whereas using fewer blocks reduced the model's ability to capture long-term dependencies. Hence, the selected configuration represents an effective trade-off between accuracy, generalization, and training efficiency. The output of the encoder is subsequently aggregated by a GlobalAveragePooling1D layer, which computes the average representation of all tokens in the sequence.

This operation produces a fixed-length feature vector independent of sequence duration, allowing the model to process sessions of varying length. The aggregated representation is then passed to the MLP module, which performs high-level feature transformation and regression preparation. The

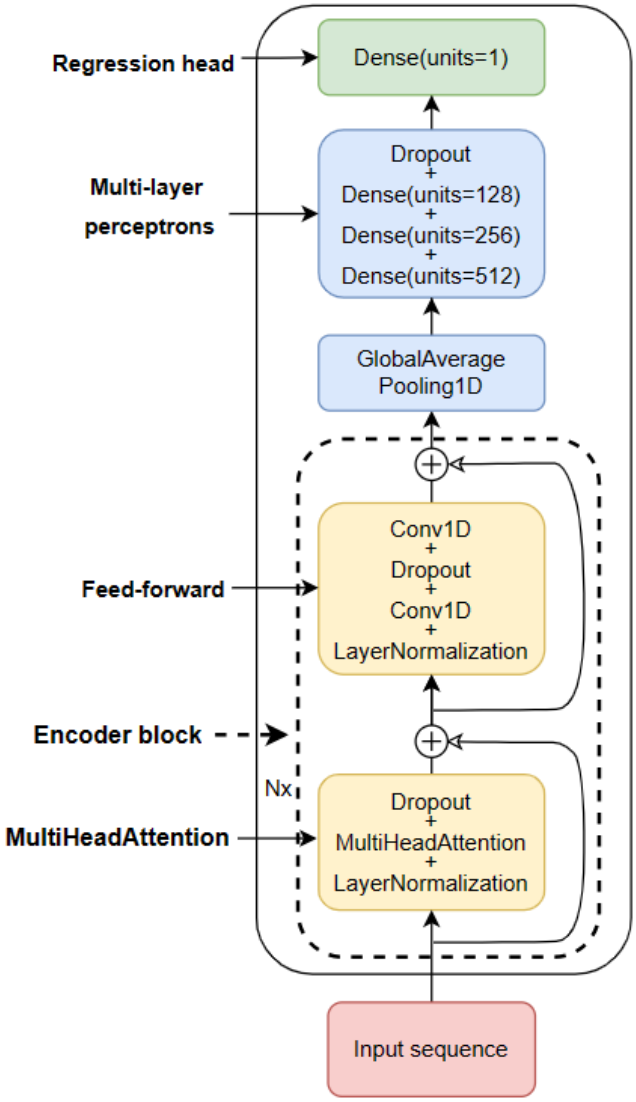

Figure 3.2: The architecture of the proposed Transformer-based model.

MLP consists of three fully connected layers with 512, 256, and 128 neurons, respectively, followed by a dropout layer with a rate of 0.25. Each dense layer uses the Rectified Linear Unit (ReLU) activation function to introduce non-linearity, and L2 kernel regularization to discourage overfitting by penalizing large weight magnitudes. Finally, the Regression head outputs the predicted QoE value in the form of a continuous MOS score. This final layer acts as a linear projection that maps the processed feature representation to a scalar quality estimate. The entire model, therefore, defines a non-linear function:

$$\hat{y} = f_\theta(SQ_v^d), \tag{3.3}$$

where $\hat{y}$ is the estimated MOS and SQ_v^d represents the sequential input for a video v from dataset d. The parameters θ are optimized through backpropagation to minimize the prediction error. To identify the best architectural configuration, the number of layers, neurons, attention heads, and regularization parameters were tuned using a Random Search approach [51]. This procedure was chosen over exhaustive grid search because of its better efficiency in exploring large hyperparameter spaces and its empirical success in identifying near-optimal solutions for deep neural networks. The model was trained to minimize the Mean Squared Error (MSE) loss function, using the Adam optimization algorithm [52], which combines adaptive learning rates with momentum for faster convergence.

Training followed a 5-fold cross-validation procedure to ensure the robustness of the results and reduce the effect of random initialization. In each fold, 70% of the available data was used for training and 30% for validation. An early stopping criterion was applied based on the validation loss, which typically led the model to converge within 40 epochs. This strategy prevented overfitting and reduced unnecessary training time. Overall, this transformer-based design achieves a strong balance between architectural simplicity and predictive capacity. By employing multi-head attention and sequential encoding of KPIs, the model is able to automatically learn which moments in a streaming session are most relevant to user perception and how they interact over time. This approach leads to more accurate, interpretable, and generalizable QoE estimations compared with conventional deep learning architectures.

3.5 | Results

Table 3.2 summarizes the overall QoE estimation performance of the proposed transformer-based model in terms of three widely adopted evaluation metrics, including RMSE, PCC, and SCC. RMSE quantifies the absolute prediction error between the estimated and ground-truth MOS values, while PCC and SCC measure, respectively, the strength of linear and monotonic relationships between the predicted and subjective scores. Together, these indicators provide a comprehensive assessment of both accuracy and ranking consistency, which are equally important for QoE prediction tasks.

Table 3.2: QoE estimation performance of the proposed transformer-based model compared to the state-of-the-art ITU-T P.1203 model in terms of RMSE, PCC, and SCC.

Device	Model	Dataset	RMSE	PCC	SCC
Mobile	P.1203	TR04	0.3850	0.9118	0.8858
		TR06	0.3964	0.9195	0.8994
		TR04+TR06	0.3881	0.9092	0.8869
	Prop.	TR04	0.3552	0.9230	0.8920
		TR06	0.2855	0.9540	0.9525
		TR04+TR06	0.3549	0.9267	0.9152
PC	P.1203	TR04	0.5257	0.8783	0.8235
		TR06	0.3595	0.9548	0.9206
		TR04+TR06	0.4867	0.9014	0.8654
	Prop.	TR04	0.3934	0.9148	0.8714
		TR06	0.3031	0.9589	0.9377
		TR04+TR06	0.3698	0.9289	0.9043
PC + Mobile	P.1203	TR04	0.4608	0.8854	0.8472
		TR06	0.3784	0.9286	0.8998
		TR04+TR06	0.4402	0.8959	0.8692
	Prop.	TR04	0.3987	0.9093	0.8733
		TR06	0.3344	0.9456	0.9442
		TR04+TR06	0.3873	0.9192	0.8972
Cross-Device	Prop.	Trn: Mob. TR06 Val: PC TR06	0.3426	0.9536	0.9541
		Trn: Mob. TR04 Val: PC TR04	0.4006	0.9151	0.8786
		Trn: PC TR06 Val: Mob. TR06	0.3608	0.9566	0.9634
		Trn: PC TR04 Val: Mob. TR04	0.4055	0.9051	0.8538

For comparison, the table also reports the performance achieved by the reference ITU-T P.1203 model in its Mode 0 configuration. This baseline operates under the highest encryption constraint and the lowest computational complexity, using only a limited set of playback and encoding parameters (e.g., bitrate, resolution, frame rate, and segment duration). Although P.1203 Mode 0 is not specifically optimized for the open datasets used in this study, since it was trained on the entire set of 30 datasets generated by the P.NATS group, it provides an important benchmark for gauging the improvements achieved by our transformer-based approach. The remaining P.NATS datasets are not publicly available, which restricts direct model retraining but does not affect the validity of the comparative analysis.

The "Device" column in Table 3.2 specifies the platform on which the subjective assessments were conducted, distinguishing between Mobile and PC viewing conditions. Both the TR04 and TR06

datasets contain subjective ratings from participants who watched the same impaired videos on both devices, which allows device-specific and cross-device evaluations to be performed consistently.

3.5.1 | Performance on same-device datasets

To address Research Questions Q1 and Q3, we first compare the performance of the proposed transformer-based model against the ITU-T P.1203 baseline across both datasets and devices. The results clearly indicate that the transformer-based approach consistently outperforms the standard model in all tested configurations. This improvement is observed not only on individual datasets (TR04 and TR06) but also when combining them into a unified training–validation pool (TR04+TR06).

The most significant gains are obtained for Mobile device evaluations on the TR06 dataset, where the proposed model achieves an RMSE of 0.2855 compared with 0.3964 for P.1203, and a PCC of 0.954 versus 0.9195. This improvement reflects the transformer's ability to better capture long-range dependencies introduced by extended video durations (180 seconds in TR06) and diverse impairment patterns. For PC devices, the proposed model also provides substantial improvements on the TR04 dataset (RMSE 0.3934 vs. 0.5257, PCC 0.9148 vs. 0.8783) and on the combined TR04+TR06 dataset (RMSE 0.3698 vs. 0.4867, PCC 0.9289 vs. 0.9014).

These results demonstrate that attention-based modeling of per-second KPIs leads to better generalization and more accurate MOS prediction across heterogeneous content and conditions. By leveraging the self-attention mechanism, the transformer can identify perceptually relevant moments in the playback sequence, such as the timing of stalls, the duration of buffering events, or the amplitude of bitrate changes, and assign them appropriate importance weights. In contrast, the analytical formulation of P.1203 treats such events using fixed rules that may not fully capture their contextual effect on user perception.

When results from both devices are aggregated (Mobile + PC), the proposed model continues to outperform the ITU-T reference across all metrics. The overall RMSE reduction averages around 0.08, while both correlation coefficients (PCC and SCC) exhibit consistent increases. The joint-device performance highlights the robustness of the proposed approach, suggesting that it successfully learns general patterns of perceptual degradation rather than overfitting to device-specific data.

3.5.2 | Cross-device evaluation

To address Research Question Q2, we further evaluate the robustness of the model under cross-device conditions, where training and validation are conducted on different device types. This setting assesses the model's capacity to generalize across viewing contexts, which is critical for practical QoE monitoring systems deployed in heterogeneous user environments. Specifically, the model was trained on all Mobile data and tested on PC data, and vice versa.

The results reveal that the transformer model maintains high prediction accuracy even when applied to a device different from that used during training. Cross-device performance is particularly strong on the TR06 dataset, where training on Mobile and validating on PC yields an RMSE of 0.3426 and a PCC of 0.9536. When the direction is reversed-training on PC and validating on Mobile, the model achieves an RMSE of 0.3608 and a PCC of 0.9566, indicating minimal degradation and excellent generalization capability. These values remain close to, or even better than, the same device performance, confirming that the attention-based model learns intrinsic QoE dynamics that are largely independent of display characteristics.

On the TR04 dataset, the cross-device correlations remain high, with PCC values around 0.91 and RMSE near 0.40 in both training–validation directions. Although slightly lower than the corresponding TR06 results, these values are still comparable to or better than those obtained by the ITU-T P.1203 model on the complete datasets (Mobile + PC). The difference between TR04 and TR06 can be attributed to the shorter sequence length and less diverse impairment configurations in TR04, which offer fewer temporal cues for the model to exploit. Nonetheless, the consistency across all scenarios indicates that the transformer architecture captures generalizable temporal relationships rather than overfitting to device or content-specific patterns.

Overall, the experimental findings confirm that the proposed transformer-based model achieves superior predictive performance and strong cross-context robustness compared to the ITU-T standard. The results provide empirical evidence for three main conclusions. First, incorporating per-second sequential data into the learning process significantly enhances prediction accuracy relative to models trained on aggregated features. Second, the multi-head attention mechanism enables the model to capture subtle contextual dependencies, such as the perceptual impact of impairment order and recovery dynamics, which are ignored by traditional approaches. Third, the model's ability to generalize across devices demonstrates its potential for real-world deployment, where QoE monitoring systems must operate reliably under diverse playback configurations and user conditions.

In summary, the transformer-based model not only achieves lower estimation error and higher correlation with subjective data but also exhibits stable and interpretable behavior across datasets. These results validate the research hypotheses formulated in Section 3.3 and position the proposed approach as a promising foundation for future QoE prediction frameworks that integrate multi-modal or cross-platform data.

3.6 | Key findings and contributions

This chapter has presented the design, implementation, and evaluation of a transformer-based model for QoE estimation in adaptive video streaming services. The key outcomes and contributions can be summarized as follows:

- Development of a sequence-aware QoE modeling framework: A complete data processing and modeling workflow was proposed, covering the collection, encoding, and sequentialization of streaming session parameters. This framework enables the representation of playback sessions as per-second sequences of KPIs, allowing the model to capture both the order and duration of impairments such as stalling, delay, and quality switching.
- Design of a transformer-based deep learning architecture for MOS regression: A tailored transformer network was implemented to estimate user-perceived quality from sequential data. The architecture integrates multi-head attention, residual connections, and lightweight convolutional transformations, effectively modeling long-range temporal dependencies and complex contextual relationships between playback events.
- Performance improvement over the ITU-T P.1203 standard: Extensive experiments on two open datasets (TR04 and TR06) demonstrated that the proposed model consistently outperforms the ITU-T P.1203 baseline in all configurations. Average reductions in RMSE ranged from 0.08 to 0.13, with corresponding increases in both PCC and SCC correlation coefficients, confirming higher prediction accuracy and perceptual alignment.
- Demonstration of strong cross-device generalization: The model exhibited high robustness when trained and tested on different devices. Cross-device evaluations showed RMSE values of 0.34 - 0.40 and PCC values exceeding 0.95, proving that the learned attention-based representation captures intrinsic QoE patterns independent of device characteristics.
- Improved interpretability and insight into temporal QoE dynamics: The attention mechanism provided interpretability by highlighting which temporal segments most influenced the predicted QoE. This feature enables a deeper understanding of the perceptual impact of event ordering, stall duration, and recovery phases, offering valuable feedback for adaptive bitrate control and streaming optimization.
- Contribution toward a scalable and generalizable QoE modeling paradigm: The proposed approach bridges the gap between analytical parametric models and conventional deep neural networks by combining high predictive accuracy, interpretability, and cross-context adaptability. It establishes a foundation for future extensions toward multimodal and real-time QoE estimation within intelligent service management systems.

Together, these findings strengthen one of the critical objectives of this dissertation: advancing service-level QoE estimation toward temporal and context-aware modeling, enabling dynamic understanding of user experience over time.

Building on the temporal and deep learning foundations established here, the next chapter extends the scope of QoE modeling from temporal sequence prediction to collaborative and privacy-preserving

learning across multiple datasets and entities. While this chapter focused on learning time-dependent patterns within single streaming scenarios, the forthcoming study explores how knowledge can be shared and integrated across diverse feature spaces using an MV learning framework, further generalizing the proposed approach toward scalable and interoperable QoE prediction.

3.7 | Conclusion

This chapter has investigated the potential of transformer learning architectures for modeling and estimating the QoE in adaptive video streaming services. To this end, a complete workflow was proposed, encompassing the key stages of data collection, encoding, sequentialization, and model training. The methodology was designed to exploit the transformer's ability to process ordered sequences and to learn long-range dependencies among temporally distributed impairments. Through this approach, each streaming session was represented as a per-second sequence of KPIs, enabling the model to capture the dynamic nature of user experience over time.

The proposed transformer-based model was rigorously evaluated using two publicly available datasets from the ITU-T P.1203 standardization campaign, TR04 and TR06, which together include 82 video sequences affected by common streaming impairments such as bitrate switching, playback delay, and rebuffering events. Subjective MOS ratings obtained for both mobile and PC devices were used as the ground-truth labels for performance evaluation. The model's predictions were compared against those of the ITU-T P.1203 reference model in Mode 0, which serves as the standard baseline for parametric QoE estimation under limited input information.

Experimental results have shown that the proposed model consistently outperforms the ITU-T P.1203 model across all tested conditions. In particular, substantial improvements were observed for the MOS prediction of mobile devices on the TR06 dataset, where the RMSE was reduced by approximately 0.11 (from 0.3964 to 0.2855), and for PC devices on the TR04 dataset, where the RMSE decreased by 0.13 (from 0.5257 to 0.3934). A similar performance gain was achieved for the combined TR04+TR06 dataset, with the RMSE lowered by about 0.117. When both mobile and PC evaluations were aggregated, the average RMSE reduction remained around 0.08, accompanied by notable increases in correlation metrics (PCC and SCC). These improvements confirm that modeling temporal dependencies through attention mechanisms yields more accurate and perceptually aligned QoE predictions than rule-based parametric approaches.

Another important finding concerns the cross-device generalization capability of the proposed model. The transformer demonstrated strong robustness when evaluated on a device different from that used for training. For the TR06 dataset, training on mobile data and validating on PC data yielded an RMSE of 0.3426, while the inverse configuration (training on PC and validating on mobile) achieved an RMSE of 0.3608. Both results correspond to PCC values above 0.95, indicating an almost

linear agreement with subjective scores. On the TR04 dataset, cross-device experiments produced RMSE values around 0.40, comparable to those obtained on the combined dataset and superior to the standard ITU-T P.1203 model. These findings demonstrate that the transformer architecture captures intrinsic QoE patterns that generalize across devices, rather than learning device-specific artifacts.

In summary, the results obtained in this study validate all three research hypotheses formulated in Section 3.3. First, sequential modeling of per-second KPIs significantly improves prediction accuracy compared to models trained on aggregated statistics. Second, the multi-head attention mechanism enables the model to discover temporal and contextual dependencies that are difficult to represent with conventional deep learning architectures. Third, the cross-device experiments confirm that the proposed approach generalizes well to heterogeneous viewing contexts, a key requirement for practical QoE monitoring systems.

Beyond numerical performance, the interpretability of the attention mechanism provides an additional advantage: it allows visualization of which temporal segments most strongly influence the predicted QoE. This feature not only enhances the model's transparency but also provides valuable insights for adaptive streaming optimization, such as identifying which types of impairments most strongly affect user satisfaction.

Overall, the proposed transformer-based framework represents an effective and generalizable solution for QoE estimation in adaptive video streaming. It bridges the gap between parametric analytical models and purely data-driven approaches by combining explainability with high predictive accuracy. The encouraging results obtained here lay the foundation for future work aimed at integrating multimodal features (e.g., audio quality, network statistics, and user feedback) and extending transformer-based architectures toward real-time QoE prediction and adaptive service management within next-generation multimedia networks.

4

Collaborative Multi-View Learning for QoE Prediction (Multi-View QoE)

4.1 | Introduction

Following the methodological roadmap introduced in Chapter 1, this chapter extends the investigation of objective QoE modeling beyond single-dataset or single-entity approaches toward a collaborative learning paradigm. While the previous chapters focused on modeling temporal and streaming-specific QoE using individual datasets, this study explores how knowledge can be shared and integrated across multiple data sources without compromising privacy or requiring data exchange. This transition directly addresses one of the dissertation's central challenges, outlined in Chapter 1, the fragmentation and non-interoperability of QoE models trained under isolated experimental conditions. To this end, this chapter introduces a Collaborative Multi-View Learning (MVL) framework that leverages multiple partial datasets, or "views," to jointly predict user-perceived QoE. Each view represents a distinct subset of features or influence factors collected by different entities (e.g., Internet Service Providers (ISPs), Over-The-Top (OTTs), or research groups). By employing MV learning techniques, the proposed approach enables knowledge sharing through model-level fusion rather than raw data exchange, ensuring both privacy preservation and improved generalization across diverse operational contexts.

This study operationalizes the third pillar of the proposed framework: MV fusion, which focuses on integrating heterogeneous or partially shared data sources to improve prediction accuracy, robustness, and generalization. Therefore, it establishes a crucial link between the early-stage signal-design and temporal-modeling studies and the later content-level investigations on perceptual and volumetric quality. In doing so, it demonstrates how collaborative learning can overcome data silos and build unified, extensible QoE predictors.

In today's highly digitalized society, user-perceived quality has become a decisive factor in the

success and sustainability of Web-based services. The ability to deliver multimedia content that meets users' expectations is increasingly recognized as a core competitive advantage for both ISPs and OTT application providers. The user perception of quality is commonly evaluated through the concept of QoE, which is defined as "the degree of delight or annoyance of the user of an application or service" [29]. QoE goes beyond purely technical indicators by integrating the human dimension, capturing how users perceive, react to, and evaluate the quality of multimedia applications as a result of network and service conditions.

Objective QoE models are employed to automatically estimate user-perceived quality without direct user involvement. These models rely on measurable QoE influence factors (IFs), which can be network-related (e.g., delay, packet loss, jitter, throughput) or application-related (e.g., playout buffering, bitrate adaptation, audiovisual fidelity) [38]. They are widely used by ISPs and OTT providers not only for real-time quality monitoring but also for root cause analysis, helping to detect inefficiencies, diagnose performance bottlenecks, and guide network and service optimization strategies aimed at improving user satisfaction.

The development of accurate objective QoE models, however, fundamentally depends on the availability of subjective test data. Such data are typically obtained through controlled experiments in which human participants rate the quality of multimedia sessions under various network and service configurations. Despite extensive research efforts and numerous subjective studies, the coverage of existing datasets remains limited. Subjective tests are inherently expensive, time-consuming, and context-dependent, making it infeasible to capture all possible scenarios of interest, especially as usage conditions, devices, and service architectures evolve over time. Furthermore, most studies consider only a restricted set of influence factors, leading to models that generalize poorly when new parameters or configurations are introduced. When experiments are conducted independently by different research groups, the lack of synchronization and consistency in parameter definitions often prevents the merging of datasets, as the recorded variables are neither standardized nor directly compatible. This heterogeneity complicates the construction of unified datasets suitable for training large-scale data-driven models, such as neural networks (NNs).

Recent advances in Artificial Intelligence (AI) have provided promising pathways to overcome these limitations. Among them, MV learning has emerged as a particularly effective paradigm for integrating heterogeneous datasets that describe the same target phenomenon from different perspectives [53]. In MV learning, each dataset or "view" represents a distinct but complementary subset of features, and the learning algorithm jointly exploits multiple views to improve predictive accuracy and generalization. By leveraging shared information across views, MV models can capture a broader understanding of the underlying relationships while preserving the specificity of each data source. This paradigm has proven successful in several domains, such as image recognition, speech processing, and social network analysis, where data from multiple sources contribute unique but related insights.

Building on these foundations, this work explores the potential of MV learning for QoE modeling. We hypothesize that integrating information from multiple subjective datasets, even when collected independently and containing different IFs, can yield more robust and generalizable QoE predictors. Beyond accuracy improvements, the MV paradigm also offers important privacy and scalability advantages: since each view is processed separately, raw data sharing between entities is unnecessary. Instead, the integration occurs at the feature or representation level (e.g., through intermediate neural layers), allowing multiple organizations to collaboratively improve prediction models without disclosing proprietary or sensitive data.

To empirically investigate this hypothesis, we consider three modeling strategies: Full View (FV), Partial View (PV), and MV. To ensure full experimental control, we start from a single comprehensive dataset of Web browsing sessions [54], which is then artificially partitioned into two distinct subsets (views) simulating independent datasets collected by separate entities. In the FV configuration, an NN is trained on the entire dataset, serving as an upper-bound reference. The PV configuration trains the same NN architecture on only one view, representing a scenario with limited feature availability. The MV configuration implements a data fusion mechanism that combines intermediate representations (learned features) extracted from the two NNs trained on separate views, forming a unified feature space for final QoE prediction. To ensure completeness, all possible combinations of influence factors in the two views are tested, allowing systematic evaluation of the impact of feature partitioning and integration.

The experimental results demonstrate that the proposed MV approach achieves QoE estimation performance comparable to that of the FV model, even though it is trained using only the two separated views. In contrast, models trained on a single view (PV) exhibit a noticeable decline in prediction accuracy, particularly when fewer IFs are available. This outcome confirms that integrating complementary information from multiple datasets can substantially improve model robustness and generalization. Moreover, the benefits of MV learning become especially evident when data scarcity or feature incompleteness restricts the performance of conventional single-view approaches.

In summary, this chapter contributes to the broader objectives of this dissertation by validating that collaborative and privacy-preserving learning strategies can achieve accuracy comparable to centralized, FV models. By confirming the feasibility of knowledge fusion across heterogeneous datasets, the MV QoE framework lays the methodological foundation for the subsequent chapters, where the focus expands toward modality scalability (3D point clouds) and deployment efficiency (FIQA).

The remainder of this chapter is structured as follows. Section 4.2 reviews related work and research gaps on QoE modeling and multi-view learning. Section 4.3 presents the proposed methodology and model design behind the suggested multi-view QoE learning method, followed by Section 4.4, which describes the implementation details and neural network configurations. Section 4.5 reports and discusses the experimental results, Section 4.6 summarizes the findings and contributions of the

chapter, and finally, Section 4.7 concludes the chapter and outlines the potential research directions.

4.2 | Background and Related Works

The MV learning paradigm has gained considerable attention in recent years as a powerful strategy to improve the performance and robustness of machine learning models, particularly in fields where information can be represented from multiple complementary perspectives. In MV learning, each "view" corresponds to a distinct feature subset or data modality that describes the same target phenomenon. By integrating these multiple representations, MV approaches can capture richer and more discriminative information, thereby achieving better generalization and interpretability compared to single-view learning. This section reviews relevant applications of MV learning in multimedia and related domains, followed by a discussion of its potential for QoE modeling.

In the domain of image captioning, the work presented in [55] proposed an MV multimodal transformer architecture that integrates visual and textual information to generate descriptive captions for images. The model includes two separate neural branches, one processing the image and the other the textual description which are subsequently fused using a transformer network. This fusion mechanism enables the model to align semantic representations across modalities, effectively capturing the relationship between image content and language structure. The proposed MV architecture significantly outperformed state-of-the-art single-view and unimodal approaches, demonstrating the benefits of learning from complementary representations.

A similar concept was applied in the biomedical field, where [56] developed a MV, multi-class disease classification system based on voice features. This framework leverages several acoustic descriptors, such as Mel-Frequency Cepstral Coefficients (MFCC), Log Mel-filter bank coefficients (logFBANK), and Spectral Subband Centroids. These heterogeneous feature sets are treated as different views of the same vocal signal, which are fused within a two-phase classification module. The integration of distinct acoustic cues significantly enhanced the model's predictive performance in identifying disease presence from speech recordings, illustrating how MV learning can effectively handle feature heterogeneity while improving discriminative power.

In the field of facial expression recognition (FER), the OCA-MTL (Orthogonal Channel Attention-based Multi-Task Learning) approach proposed in [57] adopts a Siamese CNN architecture to learn view-independent representations of human emotions. The system processes two input streams simultaneously, one from frontal and one from non-frontal viewpoints of the same subject's face, covering head rotations from $-90°$ to $+90°$. By jointly training these parallel networks with shared and task-specific objectives, the model learns pose-invariant facial features, achieving a mean accuracy of 88.4% across six emotion categories. This outperformed existing single-view FER methods and confirmed that MV learning can effectively generalize across spatial variations.

Another notable contribution is the Noise-aware Incomplete MV Learning Network (NIM-Nets) proposed in [58]. This framework was specifically designed to address incomplete or noisy data scenarios, which are common in real-world MV applications. NIM-Nets generate a shared latent representation across multiple views, ensuring consistency and information completeness even when some views contain missing or corrupted features. By explicitly modeling noise distributions, this method enhances the robustness of MV fusion and maintains predictive accuracy under imperfect data conditions. The ability to handle partial and unreliable inputs makes this approach particularly relevant for scenarios where data from multiple entities or sensors are difficult to synchronize or standardize.

The above studies collectively demonstrate that leveraging multiple views or modalities leads to more accurate, stable, and interpretable predictions compared to single-view architectures. MV learning has proven effective in diverse applications, ranging from computer vision and speech processing to medical diagnostics, where integrating information from heterogeneous sources provides complementary insights into the target variable. These works also highlight the versatility of MV frameworks in managing incomplete data, improving model robustness, and enabling fusion across different representation levels, from raw features to deep embeddings.

However, despite these successes, the use of MV learning in QoE modeling remains largely unexplored. Existing QoE studies primarily rely on single-dataset or single-modality approaches, where subjective ratings are linked to a predefined set of technical parameters, such as bitrate, rebuffering events, or network delay. In most cases, these datasets are collected independently, using distinct configurations and variable naming conventions, which limits their interoperability. Consequently, training a unified model across different studies is often infeasible without extensive data normalization or re-annotation. Moreover, classical QoE models, whether parametric or data-driven, typically ignore the opportunity to fuse information from complementary datasets that may capture different dimensions of user perception.

Several recent studies have explored privacy-preserving and collaborative learning paradigms for QoE modeling, particularly through federated and round-robin learning schemes, where multiple entities jointly train a model without sharing raw data. Notably, in [**?**], introduced collaborative learning mechanisms for QoE estimation by sequentially or federatively exchanging neural network weights across isolated data silos, demonstrating that near-centralized performance can be achieved while preserving privacy.

However, these approaches primarily assume a homogeneous feature space shared across all participating entities and rely on weight aggregation or sequential fine-tuning of a single global model. As a result, they do not explicitly address scenarios where different entities observe complementary or partially overlapping feature subsets, nor do they investigate how heterogeneous feature representations can be fused at the model level.

In contrast, many practical QoE scenarios involve MV data, where different stakeholders collect

different types of information (e.g., network KPIs, application-level statistics, or content descriptors). This distinction motivates the need for MV learning frameworks that operate at the representation level rather than solely at the model-weight level. The proposed framework evaluates three configurations, including Full View (FV), Partial View (PV), and MV to systematically assess the feasibility and benefits of learning from multiple perspectives of QoE data. By doing so, it aims to establish whether MV-based integration can lead to better QoE predictors, improved feature utilization, and increased generalization across heterogeneous data sources.

4.2.1 | Research Gap and Contribution

From the reviewed literature, several research gaps can be identified:

1. Limited distinction between collaborative learning and MV learning in QoE modeling. Existing collaborative QoE approaches focus on federated or sequential weight sharing over homogeneous feature spaces, without explicitly modeling multiple complementary feature views within a unified architecture.

2. Lack of representation-level fusion for heterogeneous QoE features. Prior works aggregate model parameters but do not investigate how latent representations extracted from different feature subsets can be fused to improve perceptual prediction.

3. Insufficient handling of partial observability and view imbalance. Existing collaborative frameworks assume equal feature availability across participants, whereas real-world QoE data often exhibit missing, uneven, or asymmetric feature views.

4. Absence of systematic analysis of view contribution and redundancy. Few studies quantify how individual views contribute to QoE prediction accuracy or how performance degrades under reduced feature availability.

5. Limited evaluation across independent subjective experiments. Most collaborative QoE studies rely on partitioned versions of a single dataset, rather than integrating genuinely independent subjective datasets with differing feature definitions.

Overall, this chapter contributes to the dissertation's overarching objective of developing generalizable, collaborative, and privacy-aware QoE modeling frameworks. Unlike prior collaborative learning approaches that focus on weight sharing over homogeneous feature spaces, this study introduces a MV learning framework for QoE prediction that explicitly models heterogeneous and partially overlapping feature views at the representation level. By demonstrating that MV fusion can approximate full information performance without raw data exchange and by systematically analyzing robustness under missing and imbalanced views, this work extends collaborative QoE modeling beyond federated

optimization and establishes a methodological foundation for cross-modality scalability (Chapter 5) and efficiency-driven deployment (Chapter 6).

4.3 | Proposed approach

The research presented in this chapter investigates the application of MV learning to QoE modeling. QoE predictors, or objective QoE models, describe the functional relationship between a set of measurable QoE IFs and the perceived user quality. In practice, these models take monitored QoE IFs as inputs and produce an estimated QoE score as output. Despite the large number of models proposed in the literature, existing QoE predictors are often incompatible across datasets or services, meaning that a model trained under certain conditions performs poorly when applied to data collected under different configurations. This lack of generalization can be attributed mainly to two reasons: (i) QoE depends on multiple IFs of different nature, including both network-based factors (e.g., delay, packet loss, throughput) typically measured by ISPs, and application-based factors (e.g., playout buffering, multimedia fidelity) collected by OTT service providers [29]; (ii) different studies employ dissimilar sets of IFs, or the same IFs measured within different value ranges, leading to inconsistent mappings between features and perceived quality.

Consequently, each model tends to achieve its best estimation accuracy only when the input variables match those used during its training phase. Because the impact of each IF, and the magnitude of its variation on user perception differs across contexts, transferring a model to another dataset usually introduces bias. Furthermore, the ground-truth data used to train objective QoE models are obtained through controlled subjective experiments, which are expensive and time-consuming to repeat. For example, consider two entities, an ISP and an OTT platform, both monitoring the QoE of the same video streaming service. The ISP develops a model that estimates QoE primarily as a function of network delay, whereas the OTT's model focuses on application-level metrics such as playout buffering. Although both delay and buffering jointly influence the perceived QoE, each model remains limited by its own data perspective. The two entities, moreover, are typically unwilling or unable to share raw data due to privacy concerns, regulatory constraints, or commercial sensitivity. This raises a key research question:

Is it possible for these entities to share abstract model information rather than data, so that each model can enhance its predictive accuracy by leveraging the learned knowledge of the other, without exposing proprietary datasets or conducting new joint subjective tests?

To explore this question, we build upon the extensive body of existing QoE models and propose an MV-based framework that can integrate the knowledge extracted from multiple independent datasets. Here, an entity refers to any organization or research group (e.g., ISP, OTT provider, or academic laboratory) conducting subjective assessments and developing QoE models for Web-based multimedia

applications. The MV learning paradigm enables such entities to combine knowledge from several views—each corresponding to a distinct subset of IFs to predict a common target (user-perceived quality). By taking advantage of multiple views, MV-based predictors can capture complementary aspects of the phenomenon, improving both accuracy and robustness compared to single-view learning. Crucially, MV learning allows this integration to occur without sharing raw data: instead, the entities exchange intermediate model representations or latent features, typically extracted from the hidden layers of neural networks trained on their own datasets. This ensures privacy preservation while enabling collaborative enhancement of model performance.

Formally, let K_i denote the vector encoding all IFs considered by entity i and used to predict user quality through model $Q_i(K_i; D_i)$, defined as

$$Q_i^{est} = Q_i(K_i; D_i), \quad 1 \leq Q_i^{est} \leq 5, \tag{4.1}$$

where the predicted QoE Q_i^{est} lies on the five-level Absolute Category Rating (ACR) scale [28]. Each model Q_i is trained on its corresponding dataset D_i, collected independently during a subjective experiment. The goal is to build an enhanced quality model $\hat{Q}_i(K_i; D_i, O_f)$ that improves estimation accuracy by integrating shared information O_f derived from the fusion layer outputs of other collaborating models Q_j, trained on their respective datasets D_j for $j \neq i$ and $j \in 1, \ldots, J$, where J is the number of entities participating in the collaboration.

To illustrate the concept, consider two models $Q_1(K_1; D_1)$ and $Q_2(K_2; D_2)$, represented in Fig. 4.1, which estimate QoE for the same service based on distinct feature sets K_1 and K_2. Their basic forms are:

$$Q_1^{est} = Q_1(K_1; D_1), \tag{4.2}$$

$$Q_2^{est} = Q_2(K_2; D_2). \tag{4.3}$$

The proposed enhancement consists of exchanging model-level information through a fusion layer output O_f, yielding improved models:

$$\hat{Q}_1^{est} = \hat{Q}_1(K_1; D_1, O_f), \quad 1 \leq \hat{Q}_1^{est} \leq 5, \tag{4.4}$$

$$\hat{Q}_2^{est} = \hat{Q}_2(K_2; D_2, O_f), \quad 1 \leq \hat{Q}_2^{est} \leq 5, \tag{4.5}$$

where $\hat{Q}_x(K_x; D_x, O_f)$ represents the refined model trained on its local dataset D_x and augmented with shared knowledge O_f (with $x \in 1, 2$). Through this mechanism, each entity benefits from the complementary insights captured by the others while keeping its raw data private.

To comprehensively evaluate the feasibility and performance of this framework, three learning strategies are implemented and compared:

1. **Partial View (PV):** Each model $Q_i(K_i; D_i)$ is trained solely on its local dataset D_i, which contains the IFs K_i and corresponding QoE scores. This configuration represents the conventional case in which each entity independently develops its predictor, producing an output Q_i^{est}.

2. **Multi-View (MV):** Each enhanced model $\hat{Q}_x(K_x; D_x, O_f)$ is trained using its own dataset D_x and the fused information O_f derived from the other collaborating models. This is the proposed collaborative configuration, producing outputs $\hat{Q}_x^{est}$ that reflect both local and shared knowledge.

3. **Full View (FV):** A single unified QoE model $Q(K; D)$ is trained on a combined dataset D containing all features $K = \bigcup_i K_i$ and corresponding QoE scores. This represents the idealized case in which all entities share their raw data to train one centralized model. It serves as an upper-bound reference, even though such data sharing is rarely feasible in practice due to privacy and operational constraints.

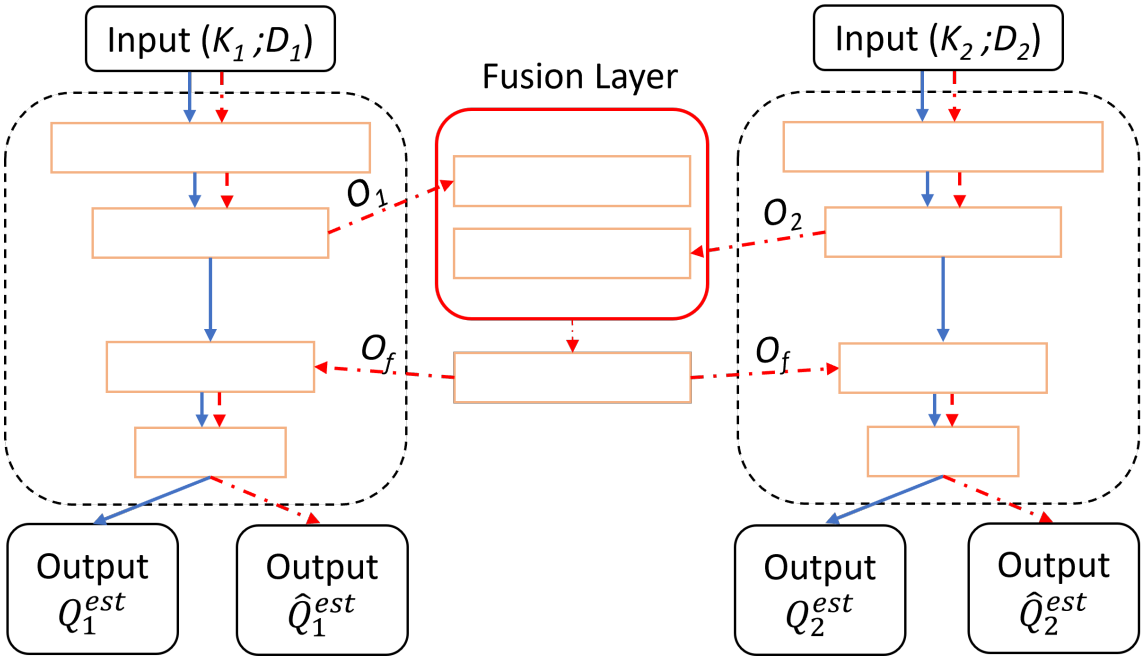

Figure 4.1: PV approach: the blue solid lines indicate the $Q_x(K_x; D_x)$ models, whose outputs are Q_x^{est} $(x = 1, 2)$. MV approach: the red dashed lines indicate the $\hat{Q}_x(K_x; D_x, O_f)$ models, which are trained with the support of the fusion layer output O_f, and whose outputs are $\hat{Q}_x^{est}$ $(x = 1, 2)$; O_x is the output of one of the hidden layers of $\hat{Q}_x$ models $(x = 1, 2)$.

By comparing these three strategies, including PV, MV, and FV, the study systematically quantifies the benefits of model-level fusion relative to isolated and fully shared configurations. The

analysis demonstrates how MV learning can serve as a bridge between isolated data silos and centralized learning, enabling collaborative QoE modeling that balances predictive accuracy, data privacy, and real-world deployability.

4.4 | Implementation details

This section describes the implementation framework adopted for evaluating the proposed MV learning approach for QoE modeling. In Section 4.4.1, we first introduce the Web QoE dataset [54], which contains subjective quality ratings of web browsing sessions and the corresponding IFs. Then, in Section 4.4.2, we present the structure and configuration of the Fully Connected Deep Neural Networks (FC-DNNs) used to model QoE in the three investigated approaches (Partial View, MV, and Full View). Finally, Section 4.4.3 details the implementation process for each approach, including the data fusion scheme, training configuration, and evaluation setup.

4.4.1 | Dataset

The experimental analysis relies on the publicly available Web browsing QoE dataset presented in [54], which provides a comprehensive mapping between network and application-level indicators and user-perceived quality scores. This dataset includes 3,400 individual web browsing sessions, each rated by human participants using the ACR scale, where 1 corresponds to Bad and 5 corresponds to Excellent. A total of 135 users participated in the subjective tests, each interacting with a variety of web pages characterized by different page sizes, numbers of embedded objects, and loading times, thus reflecting diverse real-world browsing conditions. These ratings serve as the ground truth for training and validating the proposed QoE models.

To identify the most influential factors for QoE prediction, the authors of [54] performed a Pearson correlation coefficient (PCC) analysis between each measured feature and the corresponding subjective scores. Among all the available parameters, the nine features that exhibited the strongest linear correlation with QoE (PCC > 0.7) were selected for modeling. These include:

- (1) the time required to load the Document Object Model (DOM),
- (2) the time to load the last visible multimedia element, known as the Approximate Above-The-Fold (AATF) time,
- (3) the Page Load Time (PLT) corresponding to the triggering of the onLoad event,
- (4–5) two ByteIndex (BI) metrics that quantify the progressive download behavior in terms of bytes,

- (6–7) two ObjectIndex (OI) metrics that capture the completion timing of web objects, and
- (8–9) two ImageIndex (II) metrics reflecting the incremental rendering of images during page loading.

These nine IFs jointly capture the key temporal and structural aspects of the page-loading process that are perceptually relevant to users. By focusing on the subset of features exhibiting the strongest correlation with subjective ratings, the analysis ensures that the modeling task remains both interpretable and computationally tractable. Hence, the original dataset D is composed of feature vectors each containing $d = 9$ IFs and their corresponding MOS values.

To simulate a multi-entity scenario, in which different organizations collect different subsets of features, the dataset D was artificially partitioned into two non-overlapping subsets (views), denoted by D_1 and D_2. Each subset represents the data available to an independent entity, such as an ISP or OTT provider, measuring QoE-related parameters in isolation. Formally, $D_1, D_2 \subset D$ and $D_1 \cap D_2 = \emptyset$, with $|D_1| = d_1$ and $|D_2| = d_2$, such that $d_1 + d_2 = d = 9$. This artificial split emulates a realistic use case where different stakeholders monitor distinct sets of IFs, for example, one entity focusing on network-related metrics and another on application-layer metrics without direct data sharing between them.

The motivation behind this division is twofold. First, it allows systematic evaluation of the PV scenario, in which a QoE model is trained using only a subset of features, thus mimicking real-world conditions of limited observability. Second, it enables the MV configuration, where feature representations extracted from each subset are later fused to improve prediction accuracy without merging the underlying data. The resulting framework provides a controlled environment to compare the three modeling paradigms (PV, MV, and FV) on the same baseline dataset, ensuring that performance differences are attributable solely to the modeling strategy rather than variations in data distribution.

4.4.2 | Neural Network Architectures

To implement the QoE prediction models for the three considered approaches (Partial View, MV, and Full View), we employed Fully Connected Deep Neural Networks (FC-DNNs). This choice was motivated by their capacity to learn complex, non-linear relationships between input features and target variables through hierarchical feature transformations. The fully connected structure allows each neuron in a layer to interact with all neurons in the subsequent layer, ensuring dense information flow across the network and facilitating the learning of high-order feature representations. Such representations can be easily mapped to distinct quality classes, making FC-DNNs particularly suitable for QoE estimation tasks [59]. Through a combination of matrix multiplications, bias additions, and non-linear activation functions, FC-DNNs progressively transform raw input data into meaningful latent embeddings that capture both direct and indirect dependencies among the QoE IFs.

We designed and implemented two neural network architectures, including FC-DNN1 and FC-DNN2 to address the different experimental configurations. The first network, FC-DNN1, was used for the FV configuration, whereas the second network, FC-DNN2, was used for both the PV and MV configurations. Although both architectures share similar design principles, FC-DNN2 includes a dual-branch structure and a fusion mechanism to handle the feature sets from multiple views.

4.4.2.1 | FC-DNN1: Full-View model

The first model, FC-DNN1 (depicted in Fig. 4.2), takes as input the complete dataset X consisting of x features (QoE IFs) and outputs the predicted QoE score. It comprises eight hidden dense layers with progressively decreasing neuron counts, such as 800, 600, 500, 400, 256, 128, 64, and 32 units, followed by an output layer of 5 neurons activated by a SoftMax function to produce probabilities corresponding to the 5 ACR quality levels. The hierarchical reduction of neurons across layers encourages feature abstraction and compresses high-dimensional input information into a compact latent representation.

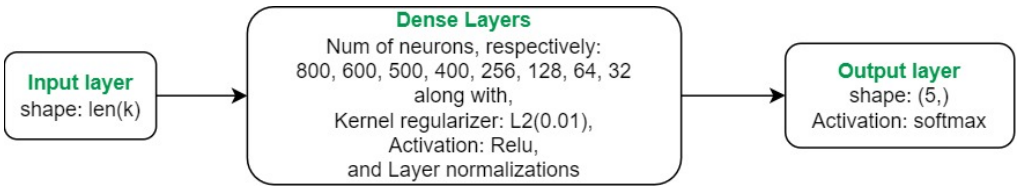

Figure 4.2: The architecture of FC-DNN1. The input dataset X is D_1 for PV_1, D_2 for PV_2, and D for FV.

Each hidden dense layer uses the Rectified Linear Unit (ReLU) activation function combined with an L2 kernel regularization term (regularization factor set to 0.01) to prevent overfitting by penalizing large weight magnitudes. To further stabilize the training process, a normalization layer is applied after each dense layer, ensuring that neuron activations maintain a stable distribution across batches and improving convergence speed.

The operations of FC-DNN1 can be formally expressed as:

$$d_{nu}(.) \coloneqq nl(ReLU(.)), \tag{4.6}$$

$$\begin{aligned} Res = SoftMax(d_{nu_8}(...(d_{nu_2}(d_{nu_1}(Input))))), \\ \text{with } nu = [800, 600, 500, 400, 256, 128, 64, 32], \end{aligned} \tag{4.7}$$

where $Input$ is the input feature vector, Res is the predicted quality distribution, and the vector of neuron units is defined as $nu = [800, 600, 500, 400, 256, 128, 64, 32]$. The normalization function $nl(.)$ normalizes the activation values within each batch according to:

$$nl = \frac{ReLU(activations) - \text{mean}(ReLU(activations))}{\sqrt{\text{var}(ReLU(activations))}}. \tag{4.8}$$

This normalization improves gradient stability and reduces the risk of internal covariate shift.

Overall, FC-DNN1 acts as the reference model capable of leveraging all available IFs simultaneously, thus serving as an upper bound for evaluating the benefits of MV integration.

4.4.2.2 | FC-DNN2: MV and Partial-View model

The second architecture, FC-DNN2 (shown in Fig. 4.3), is specifically designed for the MV and PV learning configurations. It adopts a two-branch structure, where each branch processes a different dataset (D_1 or D_2), corresponding to the separate views defined in Section 5.4.1. Each branch extracts view-specific feature representations independently before they are fused into a shared latent space through a fusion layer. This design enables the network to learn complementary information from distinct feature subsets, even when data sharing between entities is not possible.

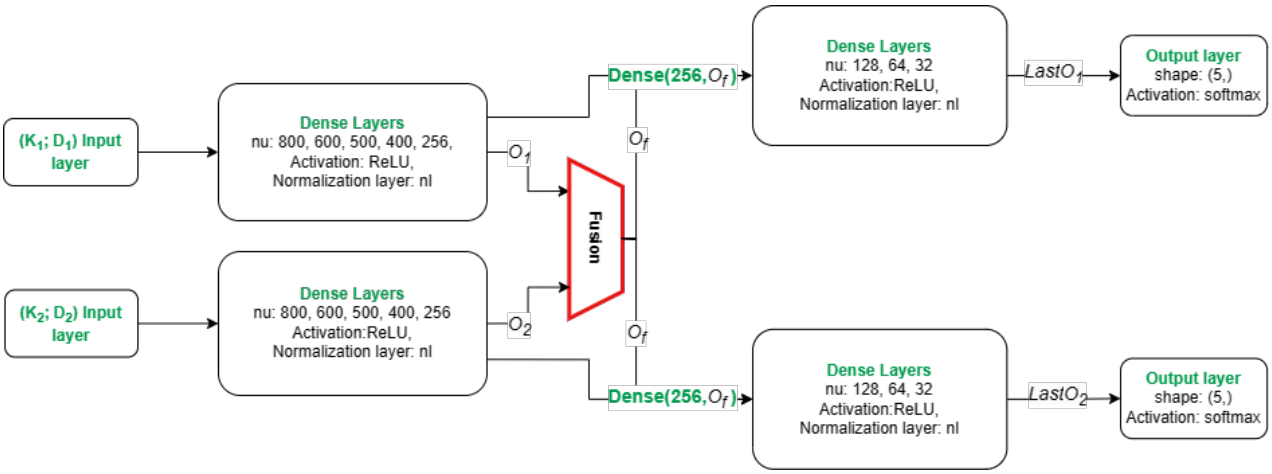

Figure 4.3: The architecture of FC-DNN2.

Each branch consists of five hidden dense layers with neuron counts $nu = [800, 600, 500, 400, 256]$. These layers are activated by ReLU functions and regularized with the same L2 penalty (0.01) and normalization as in FC-DNN1. Formally, the output of each branch before fusion is:

$$O_{bn} = d_{nu_5}(...(d_{nu_2}(d_{nu_1}(Input)))), \tag{4.9}$$

where $Input$ represents the feature vector of one view, O_{bn} denotes the encoded representation of branch bn, and $bn \in 1, 2$ is the branch index.

The Fusion layer combines the intermediate outputs O_1 and O_2 from the two branches using a concatenation operator $\oplus$ along the feature axis, followed by a dense transformation layer with 256 neurons. The fusion operation is defined as:

$$O_f = du_{256}(O_1 \oplus O_2), \tag{4.10}$$

where O_f is the fused feature vector. This shared latent representation integrates complementary patterns learned by each branch and is subsequently fed back to both branches to refine their high-level representations.

Each branch then continues with three additional dense layers of sizes $[128, 64, 32]$ applied sequentially to the fused vector:

$$LastO_{bn} = d_{nu_3}(d_{nu_2}(d_{nu_1}(O_f))), \tag{4.11}$$

producing $LastO_{bn}$ for each branch ($bn \in 1, 2$). Finally, both branches feed into an Output layer consisting of a dense layer with 5 neurons and SoftMax activation to generate class probabilities corresponding to the 5 QoE levels:

$$\hat{Q}_{bn}^{est} = SoftMax(LastO_{bn}), \tag{4.12}$$

where $\hat{Q}_{bn}^{est}$ is the predicted QoE for branch bn.

The fusion mechanism allows both branches to benefit from shared information without direct access to the other's raw data, thus ensuring privacy preservation. The 256-dimensional fusion layer acts as a knowledge-sharing interface that captures the latent relationships between different feature subsets and facilitates their alignment in the prediction space.

4.4.2.3 | Training and optimization setup

The number of hidden layers and neurons in both networks was determined empirically through iterative experimentation to achieve the best balance between model complexity and generalization. Different architectures were initially tested, and the chosen configurations (eight layers for FC-DNN1 and five layers per branch for FC-DNN2) were found to provide the most stable convergence and highest estimation accuracy. Both FC-DNNs were trained using the categorical cross-entropy loss function, optimized via the Adam optimizer. Training was monitored with an early stopping mechanism, set with a patience of 20 epochs, and the maximum number of epochs was limited to 3000. The dataset was split into training and validation subsets following a 70%/30% ratio, and 5-fold cross-validation ($k = 5$) was employed to ensure robustness and reduce the impact of data partitioning on model performance.

This configuration provides a fair comparison between the FV, PV, and MV approaches, isolating the effect of the MV fusion mechanism on QoE prediction accuracy while ensuring methodological consistency across all experiments.

4.4.3 | Approaches

To evaluate the effectiveness of the proposed MV learning approach, we used the dataset described in Section 4.4.1 as the data source D, which contains $d = 9$ IFs identified as the most correlated with the subjective QoE scores. From this dataset, we created two distinct and non-overlapping subsets, D_1 and D_2, each representing a separate "view" of the same phenomenon. Each subset includes a different combination of the nine IFs, simulating the case in which two independent entities (e.g., an ISP and an OTT provider) have access to disjoint sets of QoE-related parameters.

The number of possible combinations of IFs between the two subsets can be expressed as:

$$N_{d_1,d_2} = \frac{d!}{d_1!d_2!}. \tag{4.13}$$

This equation gives the number of ways the original feature set D can be divided into two subsets D_1 and D_2, containing d_1 and d_2 elements, respectively, such that $d_1 + d_2 = d$. In other words, it represents the total number of distinct feature partitions that can be formed to evaluate the model's robustness across different feature configurations.

Based on this formulation, the following feature combination cases were examined:

- $d_1 = 8$ and $d_2 = 1$: 9 combinations.
- $d_1 = 7$ and $d_2 = 2$: 36 combinations.
- $d_1 = 6$ and $d_2 = 3$: 84 combinations.
- $d_1 = 5$ and $d_2 = 4$: 126 combinations.

The total number of possible feature subset combinations is therefore 255, ensuring a comprehensive exploration of all feasible feature splits. This exhaustive testing allows us to assess how the availability and distribution of features between entities influence QoE prediction performance.

4.4.4 | Implementation of the three learning approaches

To ensure a fair and systematic comparison, all three approaches, including PV, MV, and FV, were implemented using the same dataset, network architectures, and training settings described in Section 4.4.2. The main differences between them lie in the number of input features and the presence or absence of the fusion mechanism.

1. Partial View (PV): In this configuration, each entity independently trains a QoE model using its local dataset only. Two separate FC-DNN1 networks were trained: one on D_1 (denoted as PV_1) and another on D_2 (denoted as PV_2). The first model receives as input d_1 features, while the second receives d_2 features. This setup reflects a realistic scenario in which entities build QoE predictors solely based on their own available IFs, without access to additional information from other sources.

2. Multi-View (MV): The proposed MV configuration aims to overcome the limitations of isolated training by integrating knowledge from both views. Here, the FC-DNN2 network is used, featuring two input branches that process D_1 and D_2 in parallel. Each branch independently learns a latent representation of its own features, and these representations are then merged in the fusion layer, as described in Section 4.4.2. This setup allows both entities to indirectly benefit from each other's learned information via model-level fusion, without sharing raw data. The MV configuration thus simulates a collaborative yet privacy-preserving training scenario.

3. Full View (FV): This configuration represents the upper-bound reference and serves as a benchmark for evaluating the MV and PV approaches. In this case, a single FC-DNN1 network is trained using the complete dataset D, which includes all nine IFs. The FV model assumes that all data are centrally available, allowing it to learn from the complete feature space. Although this setup is rarely feasible in practice due to privacy and operational constraints, it provides a valuable comparison point for quantifying the performance trade-offs introduced by MV learning.

To achieve a comprehensive evaluation, the experiment systematically covered the entire feature space. All possible combinations of IFs (d_1, d_2) were generated, and corresponding models were trained for both PV and MV configurations. For each combination, training and validation followed the same data splits, normalization procedure, and optimization settings to ensure consistency. This exhaustive approach makes it possible to analyze how feature completeness, partitioning, and fusion contribute to the final QoE estimation accuracy.

Overall, this methodology allows a detailed comparison of the three learning paradigms, including isolated (PV), collaborative (MV), and centralized (FV), in terms of both their predictive performance and practical applicability to real-world QoE modeling scenarios involving multiple independent data sources.

4.5 | Results

This section presents the experimental results obtained for the validation phase of the three considered approaches, including FV, PV, and MV, in terms of mean accuracy, precision, recall, and F1-score

for the QoE prediction task. The results are based on the dataset described in Section 4.4.1, using the experimental setup and architectures detailed in Section 4.4.2 and Section 4.4.3. Performance metrics were computed for each configuration of feature subsets and then averaged across all 5 folds of the cross-validation process.

Table 4.1 summarizes the mean QoE estimation accuracy for all tested combinations and sizes of D_1 and D_2. Each reported value represents the average performance across all possible feature combinations generated for a given (d_1, d_2) pair.

Since each configuration was evaluated under a 5-fold training/validation scheme, the total number of accuracy values contributing to each mean is given by $5 \times N_{d_1,d_2}$. Specifically:

- $5{\times}9 = 45$ values for $d_1 = 8$ and $d_2 = 1$.
- $5{\times}36 = 180$ values for $d_1 = 7$ and $d_2 = 2$.
- $5{\times}84 = 420$ values for $d_1 = 6$ and $d_2 = 3$.
- $5{\times}126 = 630$ values for $d_1 = 5$ and $d_2 = 4$.
- $5{\times}1 = 5$ values for $d = 9$.

Table 4.1: Mean QoE estimation accuracy of the FV, MV, and PV approaches for different combinations and sizes of D_1 and D_2.

Input features	$\boldsymbol{MV_1}$	$\boldsymbol{MV_2}$	$\boldsymbol{PV_1}$	$\boldsymbol{PV_2}$	**FV**
$d_1 = 8$, $d_2 = 1$	0.68	0.67	0.69	0.20	-
$d_1 = 7$, $d_2 = 2$	0.68	0.68	0.68	0.32	-
$d_1 = 6$, $d_2 = 3$	0.68	0.68	0.65	0.44	-
$d_1 = 5$, $d_2 = 4$	**0.69**	**0.69**	0.68	0.57	-
$d = 9$	-	-	-	-	**0.72**

Thus, the reported results represent robust averages computed over a total of 1,280 independent training–validation experiments. This extensive evaluation ensures statistical reliability and allows a meaningful comparison among the three modeling strategies.

4.5.1 | Overall performance comparison

As expected, the FV approach achieved the highest overall accuracy (0.72), confirming that training a single neural network (FC-DNN1) on the entire dataset, where all nine features are simultaneously available, provides the most accurate QoE estimation performance. However, the proposed MV approach achieved accuracy values that are only marginally lower but still comparable to FV, reaching 0.69 for both MV_1 and MV_2 when $d_1 = 5$ and $d_2 = 4$. This finding is significant, as the MV

configuration operates under a data-privacy constraint, where no raw data is exchanged between the two branches. Instead, only high-level feature representations are shared through the fusion layer, which integrates information from the hidden dense layers of both branches. Therefore, the MV model achieves nearly the same predictive performance as the ideal FV model while preserving data isolation between the two entities, an essential advantage for real-world applications.

Table 4.1 also reveals that the MV approach consistently outperforms the PV models, particularly when the number of available features per entity is small. The improvement is most pronounced when one of the datasets contains only a few IFs. For example, the mean accuracy achieved by PV_2 increases by 0.47, 0.36, 0.24, and 0.12 when moving from PV_2 to MV_2 for $d_2 = 1$, 2, 3, and 4 features, respectively. This demonstrates that knowledge integration through the fusion layer enables models trained on limited information to benefit from complementary patterns learned by other models, effectively compensating for missing input variables. Hence, the MV framework enhances predictive performance under partial observability, confirming its capability to support collaborative and privacy-preserving QoE estimation.

4.5.2 | Best-performing feature combinations

Table 4.2 reports the mean QoE estimation performance of the MV and PV approaches for the best-performing feature combination, corresponding to $d_1 = 5$ and $d_2 = 4$, with $D_1 = IF_2, IF_6, IF_7, IF_8, IF_9$ and $D_2 = IF_1, IF_3, IF_4, IF_5$. For comparison, the FV results (trained on all nine IFs) are also included in the same table. The findings indicate that not only the number of features available to each dataset but also their specific combination can influence the model's performance. Although the number of IFs per subset remains the primary determinant of accuracy, the selection and distribution of features across D_1 and D_2 still affect the final results to a smaller extent.

With the optimal feature split, the MV approach achieves performance nearly equivalent to FV in all evaluation metrics (accuracy, precision, recall, and F1-score). This confirms that the fusion mechanism effectively combines the most informative aspects of the two feature subsets to approximate the performance of a model trained on the full dataset. Furthermore, the table provides insight into per-class prediction performance for each ACR level. The results show that all models tend to predict lower ACR scores (1 and 2) more accurately than mid-to-high scores (3–5). This behavior is likely due to dataset imbalance, as lower quality sessions are more distinct and easier to classify, whereas high-quality sessions being more frequent and similar to each other, are harder to differentiate. Despite this challenge, the MV configuration maintains stable performance across all classes, whereas PV models show greater variability between classes.

Interestingly, the results also indicate that the MV approach harmonizes the performance between the two branches. While in the PV setup PV_1 outperforms PV_2 (due to one additional feature in its training set), in the MV configuration MV_1 and MV_2 achieve comparable performance, both

Table 4.2: Mean QoE estimation performance of the FV, MV, and PV approaches for the best combination of features when $d_1 = 5$ and $d_2 = 4$, i.e., $D_1 = \{IF_2, IF_6, IF_7, IF_8, IF_9\}$ and $D_2 = \{IF_1, IF_3, IF_4, IF_5\}$. M-AVG is the Macro Average among the 5 ACR scores.

Appr.	Metric	ACR scores					M-AVG
		1	2	3	4	5	
FV	**Mean Acc.**	0.72					
	Precision	0.91	0.76	0.65	0.57	0.67	0.71
	Recall	0.94	0.85	0.66	0.55	0.58	0.72
	F1-Score	0.92	0.80	0.65	0.56	0.62	0.71
***PV*$_1$**	**Mean Acc.**	0.68					
	Precision	0.89	0.73	0.61	0.55	0.59	0.67
	Recall	0.89	0.78	0.57	0.49	0.65	0.68
	F1-Score	0.89	0.75	0.59	0.52	0.62	0.67
***PV*$_2$**	**Mean Acc.**	0.57					
	Precision	0.74	0.61	0.53	0.44	0.51	0.57
	Recall	0.80	0.62	0.48	0.42	0.53	0.57
	F1-Score	0.77	0.62	0.50	0.43	0.52	0.57
***MV*$_1$**	**Mean Acc.**	0.71					
	Precision	0.91	0.76	0.65	0.57	0.64	0.71
	Recall	0.93	0.81	0.65	0.54	0.64	0.71
	F1-Score	0.92	0.78	0.65	0.56	0.64	0.71
***MV*$_2$**	**Mean Acc.**	0.70					
	Precision	0.91	0.78	0.64	0.57	0.61	0.70
	Recall	0.91	0.83	0.63	0.50	0.65	0.71
	F1-Score	0.91	0.80	0.64	0.53	0.63	0.70

approaching that of the FV model. This confirms that the fusion layer's shared representation successfully compensates for information asymmetry between entities, yielding balanced and improved predictions across both branches.

4.5.3 | Evaluation of minimal-feature scenarios

To further evaluate the robustness of the MV approach under extreme conditions, Table 4.3 presents the results for $d_1 = 8$ and $d_2 = 1$, where $D_1 = IF_1, IF_2, IF_3, IF_5, IF_6, IF_7, IF_8, IF_9$ and $D_2 = IF_4$. This configuration simulates a highly unbalanced data-sharing scenario in which one entity has access to nearly all features, while the other possesses only minimal information (a single IF). Such a case is representative of real-world industry collaborations, where smaller organizations or ISPs may monitor only a limited set of metrics compared to large OTT providers.

The results clearly highlight the benefit of the MV mechanism in such asymmetric configurations. Even when D_2 includes only one feature, the integration enabled by the fusion layer allows MV_2 to

reach an accuracy comparable to that of PV_1, which was trained with eight features. This illustrates that even minimal feature contributions from other entities—once transformed into latent knowledge through shared neural representations—can significantly enhance prediction accuracy.

Table 4.3: Mean QoE estimation performance of the MV and PV approaches when $d_1 = 8$ and $d_2 = 1$, with $D_1 = \{IF_1, IF_2, IF_3, IF_5, IF_6, IF_7, IF_8, IF_9\}$ and $D_2 = \{IF_4\}$. M-AVG is the Macro Average among the 5 ACR scores.

Appr.	Metric	ACR scores					M-AVG
		1	2	3	4	5	
PV_1	**Mean Acc.**	0.70					
	Precision	0.90	0.74	0.65	0.57	0.61	0.69
	Recall	0.93	0.82	0.61	0.52	0.61	0.70
	F1-Score	0.91	0.78	0.63	0.54	0.61	0.70
PV_2	**Mean Acc.**	0.20					
	Precision	0.04	0.02	0.10	0.13	0.17	0.09
	Recall	0.20	0.09	0.49	0.11	0.11	0.20
	F1-Score	0.07	0.03	0.16	0.04	0.04	0.07
MV_1	**Mean Acc.**	0.71					
	Precision	0.89	0.78	0.61	0.60	0.62	0.70
	Recall	0.94	0.83	0.72	0.42	0.62	0.71
	F1-Score	0.92	0.81	0.66	0.50	0.62	0.70
MV_2	**Mean Acc.**	0.70					
	Precision	0.87	0.78	0.65	0.58	0.59	0.69
	Recall	0.92	0.81	0.63	0.47	0.66	0.70
	F1-Score	0.89	0.79	0.64	0.51	0.63	0.69

In contrast, the single-view model PV_2 trained on only one feature performs substantially worse, underscoring the importance of cross-model information exchange. Therefore, the proposed FC-DNN2 architecture demonstrates scalability and adaptability: additional branches can be easily added to incorporate more entities, further improving model robustness and performance in collaborative multi-party settings.

The experimental results confirm three major findings. First, the MV learning strategy effectively bridges the performance gap between isolated (PV) and centralized (FV) training, achieving near-optimal QoE prediction accuracy without requiring data sharing. Second, the fusion mechanism within FC-DNN2 enables efficient knowledge transfer between independent models, allowing each to benefit from information learned by the others while maintaining data privacy. Third, the approach proves scalable and generalizable, which is able to maintain strong accuracy even under conditions of limited feature availability or imbalanced data access.

These results validate the hypothesis that integrating multiple independent QoE datasets through model-level fusion can yield accurate, privacy-preserving, and collaborative prediction frameworks.

This finding opens new research directions toward federated QoE modeling, where distributed entities cooperate through model updates rather than data exchange, enabling large-scale, ethical, and effective QoE optimization in real-world multimedia service environments.

4.6 | Key findings and contributions

The main findings and scientific contributions of this chapter can be summarized as follows:

1. Introduction of MV learning for QoE estimation. This work represents one of the first attempts to apply the MV learning paradigm to QoE prediction, enabling the integration of information from multiple independent datasets through feature-space and model-level fusion.

2. Development of a privacy-preserving collaborative framework. The proposed architecture allows separate entities to exchange learned representations rather than raw data, ensuring confidentiality while jointly improving QoE estimation accuracy.

3. Design and implementation of dual-branch FC-DNN architecture (FC-DNN2). A dedicated two-branch neural network with a fusion layer was introduced to merge intermediate latent representations from different data views, enabling efficient cross-model information sharing.

4. Comprehensive evaluation of feature distribution scenarios. A total of 255 feature partition combinations were tested to emulate diverse data-ownership and observability conditions, providing a complete characterization of model behavior across varying feature availability.

5. Demonstration of performance parity with full-view learning. The MV approach achieved accuracy and F1-scores nearly identical to those of the FV model (0.69 vs. 0.72) while significantly outperforming PV models trained on single feature subsets.

6. Scalability to asymmetric and minimal-feature configurations. Experiments with extreme cases (e.g., $d_1 = 8$, $d_2 = 1$) revealed that even when one entity has access to minimal data, MV learning can recover comparable performance through latent fusion, suggesting potential for multi-branch and federated extensions.

7. Foundation for distributed QoE modeling. The proposed framework paves the way for large-scale, real-world QoE prediction systems based on federated or collaborative intelligence, supporting distributed learning across stakeholders in multimedia service ecosystems.

Together, these findings reinforce one of the critical objectives of this dissertation: promoting collaborative and privacy-preserving QoE modeling, where distributed data sources contribute to unified learning without data exchange.

Building upon the collaborative and generalizable learning paradigm established in this chapter, the following study extends the proposed methodologies to the domain of immersive three-dimensional (3D) media. While the MV framework validated the feasibility of knowledge sharing across heterogeneous 2D streaming datasets, the next chapter investigates whether similar data-driven principles can be applied to more complex spatial content-levels, such as point clouds, where visual and geometric fidelity jointly determine user-perceived QoE. This transition marks the evolution of the dissertation's research scope, from 2D and network-centric QoE modeling toward scalable, no-reference quality assessment for 3D immersive environments, reinforcing the adaptability of the proposed learning-based approach across modalities.

4.7 | Conclusion

This chapter presented an in-depth investigation into the application of MV learning for QoE modeling, aiming to address the limitations of existing data-driven models that are constrained by the availability and heterogeneity of QoE-related datasets. Building upon the concept of collaborative learning across independent entities, we proposed an MV-based framework capable of integrating information from distinct feature subsets without requiring raw data sharing. Three complementary approaches, such as FV, PV, and MV, were designed and compared using the publicly available web QoE dataset introduced in [54].

To simulate realistic cross-entity data collection scenarios, the original dataset was divided into multiple combinations of non-overlapping feature subsets (D_1 and D_2), representing independent data views. Two deep learning architectures, FC-DNN1 and FC-DNN2, were implemented to support these experiments: FC-DNN1 was used for the FV and PV cases, while FC-DNN2, equipped with a fusion mechanism, was employed for the MV configuration. The proposed MV model leverages shared latent representations between network branches, enabling privacy-preserving information exchange that enhances prediction accuracy even under feature-limited conditions.

The obtained results demonstrated that the MV approach achieves QoE prediction performance comparable to the FV model, which operates on the complete feature set, and clearly outperforms the PV configurations that rely on partial information only. This performance gain was particularly evident when one of the datasets contained a small number of features. In such cases, the fusion layer of FC-DNN2 effectively compensated for missing information by integrating complementary patterns learned from other views, proving the potential of MV learning in collaborative QoE estimation. Moreover, the MV model maintained stable performance across all evaluated metrics—accuracy, precision, recall, and F1-score, confirming the robustness and generalizability of the proposed approach.

These findings highlight that MV integration offers a viable alternative to centralized data fusion, combining strong predictive accuracy with privacy protection and scalability. By enabling inde-

pendent entities, such as ISPs and OTT providers, to collaboratively improve QoE models without exposing their proprietary data, the proposed MV learning framework provides an effective foundation for future federated QoE modeling and distributed intelligence across networked multimedia systems.

5

No-Reference Point Cloud Quality Assessment (NR-PCQA)

5.1 | Introduction

Following the research directions and objectives outlined in Chapter 1, this chapter advances the investigation of objective QoE modeling toward immersive 3D media, marking a significant step in assessing the scalability of the proposed methodologies across content modalities. While the previous chapters addressed temporal and multi-view QoE estimation for 2D audiovisual and streaming scenarios, here the focus shifts to the spatial dimension of user experience, where depth perception, geometry accuracy, and visual fidelity jointly determine perceived quality. This transition responds directly to the overarching challenge defined in Chapter 1, developing generalizable and data-driven QoE models capable of adapting to emerging media formats such as point clouds, which serve as the backbone of Extended Reality (XR) and Human Digital Twin (HDT) environments. By introducing a no-reference point cloud quality assessment (NR-PCQA) framework, this chapter demonstrates how the thesis's learning-based approach can be effectively extended to immersive communication scenarios, preserving both accuracy and computational efficiency without relying on reference data. This approach aligns directly with the fourth and fifth pillars of the proposed framework, multi-projection fusion and computational efficiency, as it achieves state-of-the-art performance while remaining lightweight and suitable for real-time or edge deployment.

In doing so, it bridges the gap between service-level and content-level QoE modeling, confirming the versatility of the proposed learning-based paradigm across both interactive and immersive multimedia environments.

Immersive technologies are rapidly reshaping the landscape of digital communication and entertainment, enabling users to interact with three-dimensional (3D) content in more engaging and lifelike ways. Applications such as remote collaboration, live entertainment, education, and virtual events

increasingly rely on volumetric and spatial representations to provide realistic and interactive user experiences. The growing accessibility of XR devices, such as augmented, mixed, and virtual reality headsets, has accelerated this trend, allowing a wider audience to explore immersive environments with unprecedented realism. A fundamental technology supporting these experiences is the point cloud, a 3D data format that represents a scene or object through a set of spatially distributed points containing geometry and color information. Point clouds enable precise and dynamic reconstructions of real-world scenes, making them ideal for rendering human subjects or objects in volumetric video and telepresence applications. For instance, performers or speakers can be captured and transmitted as 3D entities, allowing viewers to experience events from freely chosen viewpoints, enhancing spatial immersion and user engagement.

Efficient transmission of such high-dimensional content requires adaptive streaming mechanisms that can dynamically adjust quality according to network conditions. To this end, Dynamic Adaptive Streaming over HTTP (DASH) has been extended to handle point cloud data through the introduction of DASH-PC protocols [60]. These extensions support the end-to-end pipeline of encoding, packaging, and adaptive delivery of point cloud sequences to client devices, where they are decoded and rendered in real time. Within this pipeline, an accurate PCQA mechanism is essential for several reasons: it guides the rate adaptation process, supports optimal representation selection for transmission, and provides actionable information to maintain a high QoE for end users. Consequently, objective PCQA models play a crucial role not only in perceptual optimization but also in the efficient management of network and computation resources.

Objective PCQA methods are commonly categorized into Full-Reference (FR), Reduced-Reference (RR), and No-Reference (NR) models [61]. FR models require access to both the reference and the distorted point cloud, enabling precise comparison but at the cost of high computational load and impractical requirements for real-world streaming, where the reference content is rarely available. RR models reduce this dependency by relying on a subset of extracted features from the reference and distorted versions, trading some accuracy for efficiency. NR models, instead, estimate quality solely from the distorted point cloud itself, requiring no access to the original data. This independence from reference information makes NR models inherently more scalable and deployable in real-time scenarios, as they minimize computational overhead and avoid the need for extra transmission bandwidth. For adaptive streaming systems, where quality assessment must occur continuously and at low latency, NR approaches represent the most practical and lightweight solution.

However, designing accurate and efficient NR-PCQA models remains challenging due to the unique characteristics of point clouds. Unlike conventional 2D images or videos, point clouds are composed of irregularly distributed points in 3D space, often reaching millions of points per frame. This irregular and sparse structure makes it difficult to directly apply traditional convolutional operations or spatial filters. Moreover, point clouds are susceptible to a variety of distortions, introduced during acquisition, compression, transmission, or rendering, that can degrade both geometry and color fi-

delity. These distortions manifest differently depending on the capture setup, compression algorithm, and viewing configuration, making the perceptual modeling of point cloud quality a highly complex task.

Streaming-oriented scenarios introduce further challenges. Geometry-based encoders such as G-PCC (Geometry-based Point Cloud Compression) [62] and hybrid codecs often require significant bitrates and computational resources for encoding and decoding. When quality models depend on full 3D data processing, they inherit these computational burdens, which undermine their usability in latency-sensitive environments. Existing deep learning-based PCQA models, while achieving strong performance, often suffer from high complexity due to operations like 3D convolutions [63], graph-based feature extraction [64], or multimodal fusion techniques [65]. Additionally, many of these approaches involve time-consuming preprocessing steps, such as point alignment, normalization, or voxelization, further hindering their real-time applicability. Hence, there is a growing need for compact and efficient NR-PCQA solutions capable of operating on lower-dimensional data while preserving high prediction accuracy.

To address these limitations, this work proposes a Multi-View Adaptive Weighting Point Cloud Quality Assessment (MVAW-PCQA) model—a novel NR-PCQA framework that leverages 2D projection views instead of direct 3D point cloud data. The key idea is to project the 3D point cloud onto multiple orthogonal planes, transforming the irregular spatial data into structured 2D images that are easier to process using CNNs. This transformation drastically reduces computational complexity while maintaining sensitivity to geometric and color distortions observable across different viewpoints. In particular, our model employs a pre-trained CNN backbone to extract deep features from six orthogonal projections of each point cloud. Each view is processed independently to maintain consistent feature distributions and capture view-specific information, ensuring that the network learns spatial patterns associated with local distortions without prematurely mixing them across views.

To enhance the expressiveness and flexibility of the model, an adaptive weighting fusion mechanism is introduced to combine features extracted from the six projections. Rather than treating all views equally, the fusion module learns to assign dynamic importance weights, emphasizing the most informative projections that contribute more strongly to perceived quality. This adaptive weighting strategy enables the model to better capture inter-view correlations while remaining lightweight and interpretable. Through this design, the proposed MVAW-PCQA framework achieves a strong balance between accuracy, computational efficiency, and model compactness, making it well-suited for real-time and resource-constrained applications.

Experimental evaluations demonstrate that the proposed method outperforms state-of-the-art PCQA models on the SJTU-PCQA dataset, achieving superior correlation with subjective quality scores while maintaining a compact architecture with a moderate number of parameters. These results confirm the potential of 2D projection-based NR-PCQA approaches as practical and efficient

solutions for immersive communication systems.

The remainder of this chapter is organized as follows. Section 5.2 reviews the related work in PCQA, highlighting existing methods and their limitations. Section 5.3 presents the proposed MVAW-PCQA model and details its multi-view and adaptive fusion mechanisms. In Section 5.4, experimental evaluations and comparisons with state-of-the-art methods are discussed. Section 5.5 highlights the findings and contribution of the chapter, and finally, Section 5.6 concludes the chapter and outlines potential directions for future research.

5.2 | Background and Related Works

Objective PCQA methods aim to estimate the perceptual quality of distorted 3D content through mathematical or learning-based models that replicate human perception. Depending on the availability of reference data, they are commonly divided into three main categories: Full-Reference (FR), Reduced-Reference (RR), and No-Reference (NR) approaches. FR models have access to both the original and distorted point clouds and compute quality scores by directly comparing them. RR models require only a limited set of extracted features from the reference and distorted versions. NR models, in contrast, operate solely on the distorted content, without any knowledge of the reference, making them the most suitable choice for real-time and bandwidth-limited applications such as point cloud streaming and immersive communication [61].

5.2.1 | Full-Reference and Reduced-Reference PCQA

FR PCQA methods have historically achieved the highest correlation with subjective quality, as they rely on complete geometric and color information from the reference data. Notable examples include GraphSIM, PointSSIM, and PCQM. GraphSIM [66] computes quality by modeling point clouds as graphs and evaluating color and geometry distortions using graph signal gradients, effectively capturing local structural inconsistencies. PointSSIM [67] adapts the widely used 2D Structural Similarity (SSIM) index to 3D space, assessing perceptual quality in terms of luminance, contrast, and structural fidelity across corresponding point neighborhoods. Similarly, PCQM [68] introduces a perceptually motivated weighted linear combination of geometry- and color-based features, demonstrating strong consistency with human visual perception for compressed point clouds.

Although accurate, FR approaches are impractical for streaming applications, as the original point cloud is rarely available on the client side and the computational cost of point-wise correspondence matching is high. To mitigate this dependency, a few RR approaches have been explored [69]. These extract compact statistical or perceptual descriptors from both the reference and distorted point clouds, transmitting only a reduced subset of features. However, RR methods remain limited in

scope and adoption, since even partial reference data introduces bandwidth overhead and complicates deployment in distributed, real-time environments.

5.2.2 | No-Reference PCQA

NR models eliminate the dependency on reference data and estimate perceptual quality directly from the received point cloud. They have gained increasing attention due to their potential for low-complexity, scalable, and adaptive QoE monitoring in streaming pipelines. Existing NR approaches can be broadly categorized according to their input domain into three groups: (1) projection-based models, which operate on 2D projections of the point cloud; (2) model-based methods, which directly analyze the 3D point set or graph structure; and (3) hybrid architectures, which combine information from multiple modalities.

5.2.2.1 | Projection-Based Approaches

Projection-based NR-PCQA methods are particularly attractive for real-time use because they align well with the MPEG V-PCC standard [70], which encodes point clouds as sequences of 2D texture and geometry patches. These methods process rendered 2D images derived from different viewpoints, leveraging mature 2D deep learning techniques while avoiding the high computational cost of operating on unstructured 3D data.

Among early deep learning approaches, IT-PCQA [71] explores the perceptual relationship between natural images and 3D projections through a hierarchical CNN architecture combined with adversarial domain adaptation, improving generalization across datasets. PQA-Net [72] introduces a multi-view projection framework, where 2D projections of a point cloud are processed by a CNN with dual-task learning for distortion classification and quality regression. Zhang et al. [73] proposed a dynamic view-capture strategy, generating rotating projection paths analyzed using both 2D- and 3D-CNNs to capture view-dependent distortions. More recently, MS-PCQE [74] extended this idea by incorporating multi-scale projections and a dual-branch Vision Transformer (ViT), integrating focal-length-aware feature interactions and mask-aware attention mechanisms to enhance prediction accuracy.

In parallel, other studies have explored regression-based projection models, emphasizing simplicity and interpretability. Van Damme et al. [75] proposed a linear regression model based on handcrafted 2D projection metrics, using content-aware sigmoid mappings to better align predictions with subjective scores. Weil et al. [76] developed a Gradient Boost regressor that estimates quality using compression parameters, frame rate, and viewing distance as explanatory variables. Nguyen et al. [77] adapted the ITU-T P.1203 model, originally designed for video QoE, to dynamic point cloud streaming by re-parameterizing its coefficients using subjective datasets. Finally, the bitstreamPCQ

model [78] bypasses decoded projections entirely, estimating quality analytically from encoding parameters such as quantization levels and texture bitrates, offering a computationally efficient yet less perceptually driven alternative.

Projection-based approaches thus benefit from their alignment with practical streaming pipelines and the availability of mature 2D feature extractors. However, they may suffer from information loss due to the projection process, which collapses 3D geometry into 2D planes, potentially omitting depth cues and fine structural variations relevant to human perception.

5.2.2.2 | Model-Based Approaches

Model-based NR-PCQA methods analyze the 3D point cloud directly, preserving its geometric structure at the cost of increased computational burden. The 3D-NSS model [79] extracts color and geometric statistics such as curvature, anisotropy, and LAB color components, applying natural scene statistics (NSS) principles and a support vector regressor (SVR) to predict quality. GPA-Net [64] introduces a graph convolutional framework with a novel Graph Point Attention (GPAConv) kernel that captures structure-aware distortions through learned spatial dependencies. ResSCNN [63] employs sparse convolutional operations to efficiently process geometry and color features encoded as sparse tensors, offering hierarchical feature extraction while reducing redundant computation. Similarly, GQI [80] utilizes CNNs trained on local geometric patches and integrates curvature and grayscale attributes to compute global perceptual scores.

These approaches typically achieve strong accuracy but are computationally demanding and require full 3D decoding, limiting their applicability in real-time streaming or embedded systems. They also involve significant preprocessing, including normalization and alignment of point coordinates, which increases latency and memory consumption.

5.2.2.3 | Hybrid Architectures

Hybrid PCQA models attempt to combine the advantages of both projection-based and model-based approaches by fusing 2D texture and 3D geometric cues. The MM-PCQA model [65] exemplifies this category by employing a dual-branch architecture that processes 2D projections and 3D point segments in parallel. A cross-modal attention mechanism merges features from both modalities, achieving improved accuracy but at the cost of considerable complexity. Similarly, Plain-PCQA [81] adopts a multi-branch deep learning framework that analyzes both visual and geometric attributes. It includes a no-reference branch for perceptual features extracted from projections and a degraded-reference branch that incorporates simplified geometric information. This hybrid configuration adapts to varying levels of reference availability, yielding high accuracy but remaining computationally heavy.

While hybrid and 3D model-based methods deliver competitive performance, their reliance on high-dimensional or multimodal data leads to significant computational overhead and preprocess-

ing costs, making them less practical for real-time or streaming applications. Projection-based NR-PCQA, in contrast, provides a more efficient pathway for perceptual assessment, as it can directly exploit the 2D projections already generated by V-PCC or DASH-based streaming pipelines. This eliminates the need for complex 3D reconstruction or feature synchronization, allowing for a lightweight yet effective assessment of perceptual quality.

The review above reveals a clear research trend toward projection-based NR-PCQA models, which balance performance and efficiency by leveraging 2D representations. However, existing methods often treat multiple projections equally, overlooking the fact that different views contribute unequally to the overall perceived quality. Moreover, most approaches rely on simple feature averaging or concatenation, which may dilute critical spatial cues. To address these limitations, the model proposed in this work introduces a Multi-View Adaptive Weighting PCQA (MVAW-PCQA) framework that processes six orthogonal 2D projections through a CNN backbone and adaptively fuses their deep features based on learned importance weights. This design not only reduces computational cost but also enhances interpretability by explicitly modeling view-level relevance in quality estimation. By eliminating the need for 3D reconstruction or multimodal integration, the proposed approach achieves high efficiency while maintaining strong correlation with human-perceived quality, paving the way for scalable and real-time NR-PCQA solutions in immersive communication systems.

5.2.3 | Research Gap and Contribution

From the reviewed literature, several gaps and open challenges can be identified in the field of objective PCQA:

1. High dependency on reference data. Most existing PCQA models are designed under the FR or RR paradigm, which requires access to either the complete or partial original point cloud. This dependency makes them impractical for deployment in real-time streaming scenarios, where the reference content is unavailable or too large to transmit.

2. Excessive computational complexity of 3D model-based methods. Deep 3D architectures, such as convolutional and graph-based networks, require dense point sampling, voxelization, or graph construction, resulting in high computational cost and long inference times. This significantly limits their scalability for real-time applications or integration into adaptive streaming pipelines.

3. Limited generalization and flexibility of hybrid models. Although hybrid models achieve strong accuracy by combining 2D and 3D modalities, they depend on complex preprocessing, cross-modal synchronization, and large memory footprints. These constraints hinder their deployment in dynamic or resource-constrained environments, such as mobile XR devices or live streaming systems.

4. Inefficient fusion in projection-based models. Most existing projection-based NR-PCQA methods process multiple 2D projections but treat all views equally, using either concatenation or uniform averaging to combine features. This uniform treatment ignores the fact that different projections contribute unequally to perceived quality, depending on the spatial structure, distortion type, and viewpoint relevance.

5. Lack of lightweight and adaptive NR-PCQA models. Despite their efficiency advantage, existing projection-based approaches often fail to balance model compactness and perceptual accuracy. There remains a need for a method that can provide both adaptive view importance estimation and computational efficiency while maintaining state-of-the-art performance.

This chapter aims to effectively bridge the identified gaps by presenting a precise and perceptually robust PCQA framework. This innovative approach is closely aligned with the overarching objectives of this dissertation, enhancing its overall contribution to the field. The objective is to address the gaps by introducing a Multi-View Adaptive Weighting Point Cloud Quality Assessment framework, a novel no-reference approach that relies solely on 2D projection views of the point cloud. The key idea is to replace costly 3D operations with multi-view 2D feature extraction and to integrate the resulting representations using an adaptive weighting fusion module that learns the relative contribution of each projection to the overall perceived quality. By operating exclusively on 2D projections, the proposed model drastically reduces computational overhead while maintaining sensitivity to geometric and color distortions.

5.3 | Proposed approach

We propose a projection-based no-reference NR-PCQA method called MVAW-PCQA, which predicts the QoE for distorted point clouds using only their 2D projection views as input.

MVAW-PCQA is built on a deep learning architecture (illustrated in Fig. 5.1) that extracts features from each projection view of the point cloud. It incorporates an adaptive weighting mechanism that prioritizes features derived from the most significant projection views, those that contain the most relevant information needed to predict the quality of the point cloud.

5.3.1 | MVAW-PCQA Architecture

Let us consider a colored 3D point cloud $\mathcal{PC}$, represented as a finite set of N points in Euclidean space:

$$\mathcal{PC} = \{p_n = (x_n, y_n, z_n) \in \mathbb{R}^3 \mid n = 1, 2, \ldots, N\}, \tag{5.1}$$

where each point p_n is defined by its 3D coordinates (x_n, y_n, z_n). Each point may also contain associated color attributes, which contribute to the perceptual representation of the object. To

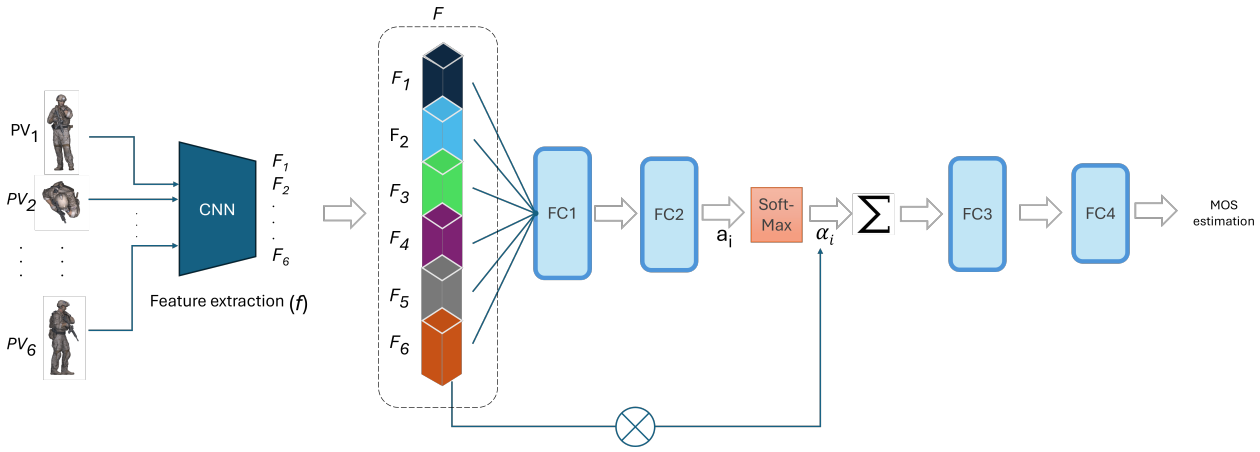

Figure 5.1: The architecture of the proposed MVAW-PCQA method.

reduce the complexity of operating directly on irregular 3D data, the proposed method converts the point cloud into a set of *2D projection views (PVs)*. Specifically, six orthogonal projections are generated, denoted as $\{PV_1, PV_2, \ldots, PV_6\}$, corresponding to the *front*, *back*, *left*, *right*, *top*, and *bottom* perspectives of the 3D point cloud, as illustrated in Fig. 5.2. These projections preserve both the *texture* and *color distribution* of the original $\mathcal{PC}$, while implicitly retaining geometric cues through perspective variations. This 2D representation provides a structured and dense format that is far easier to process using standard convolutional architectures, while still encapsulating the essential perceptual information of the original 3D scene.

The overall architecture of the proposed MVAW-PCQA model is shown in Fig. 5.1. The network takes as input the six PVs of each point cloud and follows a modular pipeline consisting of: (1) *multi-view feature extraction* using a convolutional backbone; (2) *adaptive weighting fusion* for view importance estimation; and (3) *quality regression* for predicting the MOS. Each of these components is described in the following subsections.

5.3.1.1 | Multi-View Feature Extraction

The first stage focuses on extracting high-level semantic and perceptual features from each of the six projection views. For this purpose, we adopt *ConvNeXt-T* [82] as the backbone CNN architecture, chosen for its excellent balance between *accuracy*, *computational efficiency*, and *scalability*. ConvNeXt-T builds upon the design principles of modern Transformers but retains the simplicity of CNNs, employing depthwise convolutions, layer normalization, and large receptive fields to achieve high representational capacity at a low inference cost.

The network was pre-trained on ImageNet to leverage strong generalization capabilities learned from large-scale natural image data. To adapt ConvNeXt-T to the PCQA task, its original classi-

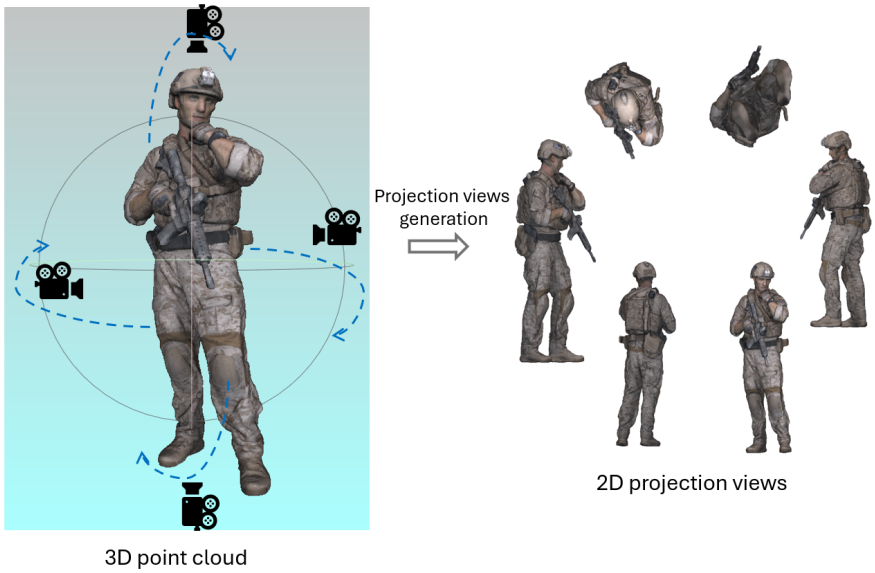

Figure 5.2: Generation of the six 2D projection views (front, back, left, right, top, and bottom) from the 3D point cloud.

fication head was replaced with a fully connected (FC) layer that outputs a 256-dimensional latent feature vector for each view. Formally, the CNN-based feature extraction function can be expressed as:

$$F_i = f(PV_i), \quad i \in \{1, 2, \ldots, P\}, \quad F_i \in \mathbb{R}^D, \tag{5.2}$$

where $f(\cdot)$ denotes the feature extractor, $P = 6$ is the number of projection views, and $D = 256$ is the feature dimensionality. Each PV is processed independently to obtain its corresponding feature vector F_i, ensuring that the network captures view-specific information without premature feature mixing. All extracted feature vectors are then vertically concatenated to form the global feature matrix:

$$F = [F_1, F_2, F_3, F_4, F_5, F_6]^T \in \mathbb{R}^{P \times D}, \tag{5.3}$$

This representation provides a compact yet informative description of the point cloud from multiple perspectives, preserving the statistical and perceptual diversity of the input data.

5.3.1.2 | Adaptive Weighting Fusion Mechanism

Traditional fusion strategies in multi-view PCQA, such as uniform averaging or heuristic weighting [83, 84, 85], assume that all projections contribute equally to the perceived quality. However, this assumption often leads to suboptimal performance, since certain viewpoints may contain more salient information about visual distortions or structural degradations. For example, occlusions, miss-

ing regions, or compression artifacts may be more evident from specific orientations, making them disproportionately important for accurate quality assessment.

To overcome this limitation, we propose a learnable adaptive weighting mechanism that dynamically determines the importance of each view. The mechanism operates as a lightweight attention-like module, implemented as a two-layer Multi-Layer Perceptron (MLP) with ReLU activation:

$$a_i = W_2\,\sigma(W_1 F_i + b_1) + b_2, \quad i = \{1, \ldots, P\}, \tag{5.4}$$

where:

- $W_1 \in \mathbb{R}^{256\times128}$ and $b_1 \in \mathbb{R}^{128}$ are trainable parameters of the first linear layer (FC1);
- σ denotes the ReLU activation function;
- $W_2 \in \mathbb{R}^{128\times1}$ and $b_2 \in \mathbb{R}^{1}$ are trainable parameters of the second linear layer (FC2).

The scalar outputs a_i represent preliminary importance scores assigned to each projection view. To ensure that the learned importance scores are normalized and comparable across views, a softmax function is applied along the projection dimension:

$$\alpha_i = \frac{\exp(a_i)}{\sum_{j=1}^{P}\exp(a_j)}, \quad \text{such that } \sum_{i=1}^{P}\alpha_i = 1. \tag{5.5}$$

The normalized weights α_i capture the relative contribution of each view, assigning higher values to projections that carry more perceptually significant information. The final fused feature representation is then obtained as a weighted combination of the six view-specific features:

$$F_{fused} = \sum_{i=1}^{P}\alpha_i F_i, \quad F_{fused} \in \mathbb{R}^{D}. \tag{5.6}$$

This adaptive fusion strategy enhances the model's sensitivity to view-dependent distortions while suppressing redundant or less informative projections. By learning these weights jointly with the rest of the network, the model can dynamically focus on the most relevant spatial cues during training, leading to more accurate quality predictions.

5.3.1.3 | Quality Regression and Output Prediction

The final stage of the MVAW-PCQA pipeline maps the fused feature vector F_{fused} to a predicted MOS value, which represents the estimated perceptual quality of the input point cloud. To achieve this, a two-layer MLP is applied, incorporating ReLU activation for non-linearity and a linear output for regression:

$$MOS_p = W_4\,\sigma(W_3 F_{fused} + b_3) + b_4, \tag{5.7}$$

where:

- $W_3 \in \mathbb{R}^{256\times128}$ and $b_3 \in \mathbb{R}^{128}$ are the trainable parameters of the first FC layer (FC3);
- $W_4 \in \mathbb{R}^{128\times1}$ and $b_4 \in \mathbb{R}^{1}$ are the trainable parameters of the second FC layer (FC4);
- σ again denotes the ReLU activation function.

This regression head learns a non-linear mapping from the high-dimensional fused feature space to a single scalar quality score. The model is trained end-to-end to minimize the difference between predicted and ground-truth MOS values, allowing the network to jointly optimize both the view-weighting mechanism and the quality regression function.

In summary, the proposed MVAW-PCQA architecture integrates multi-view feature extraction and adaptive feature fusion within a single lightweight framework. By processing orthogonal 2D projections of the point cloud, it avoids the computational burden of 3D convolutions and complex alignment procedures, while the adaptive weighting module ensures that the most informative views dominate the quality prediction. This combination of efficiency, adaptivity, and perceptual awareness allows MVAW-PCQA to outperform existing NR-PCQA methods both in terms of correlation with subjective scores and model compactness, making it suitable for real-time point cloud streaming and immersive communication scenarios.

5.4 | Experimental Results

To assess the effectiveness of the proposed MVAW-PCQA model, a comprehensive set of experiments was conducted on a well-established point cloud quality benchmark. The goal of this evaluation is to investigate the predictive performance of the proposed model in estimating the perceptual QoE and to compare it against existing state-of-the-art PCQA approaches. This section presents the dataset used for training and evaluation, followed by the experimental setup, evaluation metrics, and quantitative results.

5.4.1 | Dataset

The SJTU-PCQA dataset [86] is one of the most widely used and publicly available large-scale benchmarks for point cloud quality assessment. It was specifically designed to support the development and evaluation of both full-reference and no-reference PCQA models. The dataset includes nine distinct point cloud contents, each representing different levels of geometric complexity, texture richness, and color distribution. These contents span a wide range of perceptual characteristics, from simple static objects to complex human and environmental scenes, ensuring that the dataset covers various types of distortions and visual artifacts encountered in real-world applications.

Each original point cloud was subjected to seven distinct types of distortions, each applied at six different degradation levels, thus providing a diverse range of quality variations for learning-based models. The seven types of distortions include:

- **Octree-based compression**, simulating quality loss due to geometry quantization during 3D compression.
- **Color noise**, introducing random perturbations to color channels while preserving geometric structure.
- **Downscaling**, reducing spatial resolution by sub-sampling points to mimic acquisition or transmission constraints.
- **Geometry Gaussian noise**, adding Gaussian-distributed positional noise to points to degrade surface fidelity.
- **Downscaling + Color noise**, jointly introducing resolution and texture degradation.
- **Downscaling + Geometry Gaussian noise**, combining geometric and structural distortion effects.
- **Color noise + Geometry Gaussian noise**, representing complex mixed artifacts often present in compressed or streamed data.

Altogether, the SJTU-PCQA dataset contains a total of $9 \times 7 \times 6 = 378$ distorted samples, each associated with subjective quality scores collected through an extensive human study. The subjective tests were conducted in a controlled laboratory environment, following the ITU-R Recommendation BT.500-11 [87] guidelines for subjective video and visual quality assessment. A total of 64 participants (aged between 18 and 30 years) evaluated the visual quality of each point cloud sample. Participants rated the perceived quality using a 10-point scale, which was subsequently mapped to the conventional five-category ACR scale, including 1–2 (Bad), 3–4 (Poor), 5–6 (Fair), 7–8 (Good), and 9–10 (Excellent). The final subjective quality for each distorted point cloud was computed as the MOS by averaging the individual scores across all subjects. This provides a robust ground-truth measure of perceptual quality, which serves as the target for the regression task in our model training. The diversity of distortion types, combined with the wide MOS distribution, makes SJTU-PCQA a challenging and representative benchmark for testing no-reference PCQA algorithms.

To ensure a fair comparison, all point clouds were preprocessed following the official dataset protocol. Each 3D point cloud was first normalized within a unit cube to ensure geometric consistency, and its six orthogonal projection views were generated following the procedure described in Section 5.3.1. These PVs were then used as the input to the proposed MVAW-PCQA network for both training and evaluation.

5.4.2 | Implementation Details

The proposed MVAW-PCQA framework was implemented and trained on a high-performance workstation equipped with an Intel Core i9-14900K CPU, an NVIDIA RTX 4090 GPU, and 64 GB of RAM. The model was developed using the PyTorch deep learning framework, which provides the flexibility and efficiency required for large-scale experimentation. All experiments were conducted under the same computational environment to ensure reproducibility and consistency of results. To balance stability and convergence speed, different learning rates were assigned to the two main components of the network. The CNN backbone (ConvNeXt-T) was initialized with a smaller learning rate of 5×10^{-5}, as it was pre-trained on ImageNet, whereas the regression head was trained with a higher rate of 5×10^{-4} to allow faster adaptation to the PCQA task. This differential learning rate strategy stabilized optimization and prevented catastrophic forgetting in the pre-trained backbone. Training was carried out using the Adam optimizer [52], which combines adaptive learning rate adjustment with momentum to achieve fast and stable convergence. A weight decay term of 1×10^{-4} was applied every eight epochs to mitigate overfitting. The batch size was set to 4, which offered a good compromise between GPU memory efficiency and gradient estimation stability. A linear warm-up schedule was adopted for the first 15 epochs, followed by cosine learning rate decay over a maximum of 200 epochs. The final trained network consisted of approximately 29 million parameters, reflecting a compact yet expressive architecture. To enhance generalization and robustness to viewpoint variations, we applied data augmentation during training. Each projection view was randomly cropped to a spatial size of $224 \times 224 \times 3$ pixels and underwent random rotations in multiples of (90°). These augmentations introduced diversity into the training samples, helping the model learn orientation-invariant features and reducing overfitting to specific view configurations. Model performance was evaluated using K-fold cross-validation, a widely adopted approach for limited datasets. Following the protocol in [88], we set K=9 for the SJTU-PCQA dataset. In each fold, point clouds corresponding to eight contents (including all associated distortions and levels) were used for training, while the remaining one was reserved exclusively for validation. This process ensured that every content appeared once in the validation set, thereby providing a fair assessment of the model's generalization capability across unseen point cloud contents. The MVAW-PCQA model was trained to minimize the MSE loss between the predicted and subjective MOS values. To account for the inherent nonlinear relationship between objective predictions and subjective ratings, a logistic mapping function [74] was applied to the predicted scores before computing correlation metrics. The mapping is defined as:

$$Q_p = \frac{\lambda_1 - \lambda_2}{1 + \exp\left(-\frac{Q_f - \lambda_3}{|\lambda_4|}\right)} + \lambda_2, \tag{5.8}$$

where Q_p and Q_f represent the mapped and raw predicted MOS scores, respectively, and λ_i (with $i = 1, 2, 3, 4$) are curve-fitting parameters optimized through nonlinear regression. This post-

processing step improves the linearity of the relationship between objective and subjective scores, ensuring that correlation measures such as PCC and SCC reflect true prediction performance rather than scale distortions. Overall, these implementation settings were carefully chosen to balance accuracy, computational efficiency, and model stability, ensuring a fair and reproducible evaluation of the proposed MVAW-PCQA method.

5.4.3 | Comparing with the State-of-the-Art

Table 6.1 summarizes the performance of the proposed MVAW-PCQA model on the SJTU-PCQA dataset and compares it against several state-of-the-art FR and NR PCQA methods. The performance evaluation was conducted using four widely adopted objective quality metrics: the SCC, PCC, KCC, and RMSE. These metrics jointly capture both the accuracy and the monotonic relationship between the predicted and the ground-truth subjective MOS.

Table 5.1: Performance comparison with the state-of-the-art approaches on the SJTU-PCQA dataset.

Type	Approaches	SCC	PCC	KCC	RMSE
FR	GraphSIM [66]	0.8783	0.8449	0.6947	1.0321
	M-p2po [89]	0.7294	0.8123	0.5617	1.3613
	HD-p2po [90]	0.7157	0.7753	0.5447	1.4475
	M-p2pl [91]	0.6277	0.5940	0.4825	2.2815
	PB-PCQA [86]	0.6020	0.6076	-	1.8635
	PSNR-yuv [92]	0.7950	0.8170	0.6196	1.3151
	PCQM [68]	0.8644	0.8853	0.7086	1.0862
	MS-SSIM [93]	0.6888	0.4082	0.4995	2.2154
	PointSSIM [67]	0.6867	0.7136	0.4964	1.7001
NR	3D-NSS [79]	0.7144	0.7382	0.5174	1.7686
	ResSCNN [63]	0.8590	0.8931	0.6812	1.0373
	PQA-net [72]	0.8372	0.8586	0.6304	1.0719
	GPA-Net [64]	0.8750	0.8860	-	-
	Plain-PCQA [81]	0.9133	0.9302	0.7603	0.8607
	MM-PCQA [65]	0.9103	0.9226	0.7838	0.7716
	MS-PCQE [74]	0.9180	0.9326	0.7740	0.8241
	MVAW-PCQA	**0.9306**	**0.9466**	**0.7947**	**0.7219**

As reported in Table 6.1, the proposed MVAW-PCQA method consistently outperforms all competing approaches, including both FR and NR models. Notably, our model achieves an SCC of 0.9306 and a PCC of 0.9466, surpassing the top-performing NR model MS-PCQE (SCC = 0.9180, PCC = 0.9326). Similarly, MVAW-PCQA attains a KCC of 0.7947, outperforming the previous state-of-the-art MM-PCQA (KCC = 0.7838), while also achieving a lower RMSE of 0.7219 compared to 0.7716 for MM-PCQA. These results demonstrate that the proposed adaptive weighting fusion

mechanism allows the network to better capture perceptual cues from multiple projection views, improving correlation with subjective judgments and reducing prediction error.

The observed performance gain can be attributed to the multi-view feature learning and adaptive weighting mechanism integrated within MVAW-PCQA. By learning the relative importance of each projection view, the model can emphasize the most informative geometric and texture perspectives, leading to more discriminative feature representations. This allows the model to generalize better across different distortion types and intensity levels, as evidenced by the consistent improvement across all correlation metrics.

5.4.4 | Model Complexity and Parameter Comparison

To further contextualize the performance, Table 5.2 compares the number of parameters used in the neural networks of various PCQA methods. The proposed MVAW-PCQA model contains approximately 29 million parameters, which is higher than several lightweight architectures such as PQA-Net (0.29M), ResSCNN (1.23M), and MS-PCQE (14.27M). However, the parameter count of MVAW-PCQA is comparable to Plain-PCQA (28.5M) and significantly smaller than MM-PCQA (58.37M).

Table 5.2: Comparison of the number of parameters in PCQA Neural Networks.

Method	**Number of params**
PQA-Net [72]	0.29 M
ResSCNN [63]	1.23 M
MS-PCQE [74]	14.27 M
Plain-PCQA [81]	28.50 M
MM-PCQA [65]	58.37 M
MVAW-PCQA	29 M

Despite this moderately higher parameter count compared to some methods, MVAW-PCQA achieves substantial performance gains across all evaluation metrics, outperforming both compact and large models. This demonstrates that MVAW-PCQA effectively utilizes its additional representational capacity to learn richer perceptual quality features from multi-view projections without unnecessary redundancy. In practice, this means that the model is not only more accurate but also efficient in how it allocates its parameters for perceptual learning.

5.4.5 | Computational Efficiency and Inference Cost

Table 5.3 presents the results of a complementary experiment analyzing the computational efficiency of MVAW-PCQA in practical deployment scenarios. We evaluated three primary indicators: average

inference time per fold, peak GPU memory usage, and floating-point operations (FLOPs) per sample for the SJTU-PCQA dataset using the ConvNeXt-T backbone.

Table 5.3: Average inference cost per fold for SJTU-PCQA dataset.

Inference Time (s)	GPU Memory (GB)	FLOPs (G)
1.74	0.2	26.82

The model achieved an average total inference time of 1.74 seconds per fold, with a peak GPU memory usage of 0.2 GB and approximately 26.82 GFLOPs per sample. These figures indicate that MVAW-PCQA is both computationally efficient and lightweight, ensuring its suitability for near real-time or edge-level deployment, such as immersive video streaming or mobile XR applications.

The efficiency of MVAW-PCQA can be largely attributed to the ConvNeXt-T backbone, which is specifically optimized for parallel convolutional processing. While the model includes a relatively large number of parameters, it maintains low computational overhead due to its purely convolutional design. Unlike transformer-based architectures that rely on self-attention operations with irregular memory access patterns [94, 95], ConvNeXt-T employs structured convolutional blocks that exploit highly optimized GPU kernels. This leads to significantly lower FLOPs and faster inference time, as also reported in [82].

Therefore, MVAW-PCQA inherits the advantages of convolutional efficiency while incorporating an adaptive weighting fusion strategy that captures high-level perceptual dependencies among multiple projections. The result is a model that achieves state-of-the-art accuracy without sacrificing computational practicality. These findings confirm that MVAW-PCQA offers a high-capacity yet efficient solution for NR-PCQA tasks—balancing predictive precision, model compactness, and inference speed. Such characteristics make it a strong candidate for real-time quality monitoring in modern 3D streaming pipelines and immersive multimedia systems.

5.5 | Key findings and contributions

This chapter introduced the MVAW-PCQA model, a novel no-reference framework designed to assess the perceptual quality of point clouds using only 2D projection views. The proposed model was developed to overcome the computational and structural limitations of traditional FR and model-based 3D approaches by replacing direct 3D processing with efficient 2D feature extraction and adaptive fusion.

The main findings and contributions of this study are summarized as follows:

1. A novel projection-based NR-PCQA architecture. The chapter proposed a multi-view learning framework that operates exclusively on six orthogonal 2D projections of the input point

cloud. This approach eliminates the need for access to the original 3D reference model, drastically reducing both computational cost and data dependency while maintaining high prediction accuracy.

2. Adaptive weighting fusion for view importance learning. Unlike previous projection-based methods that rely on uniform or fixed fusion strategies, MVAW-PCQA introduces an adaptive weighting module that learns the relative importance of each projection. This mechanism enables the network to focus on the most informative perspectives, improving perceptual feature representation and overall quality prediction.

3. Strong predictive accuracy and generalization. Extensive experiments on the SJTU-PCQA dataset demonstrated that the proposed MVAW-PCQA significantly outperforms all existing FR and NR models. It achieved state-of-the-art results with SCC = 0.9306, PCC = 0.9466, KCC = 0.7947, and RMSE = 0.7219, indicating its ability to capture the nonlinear perceptual dependencies between point cloud distortions and human quality perception.

4. Balanced trade-off between accuracy and efficiency. Despite having a parameter count comparable to other high-capacity models (around 29 million), MVAW-PCQA maintains efficient inference with 26.82 GFLOPs per sample and a memory footprint of only 0.2 GB. This confirms that the model's design achieves an optimal balance between representational power and computational efficiency, making it well-suited for practical and real-time deployment.

5. Generalizable and scalable PCQA solution. The architecture's modular design, based on projection-level processing and adaptive fusion, enables easy extension to additional views or different backbones. This scalability makes the framework adaptable to diverse datasets, compression schemes, and streaming environments, contributing to the long-term goal of universal and reference-free 3D quality assessment.

In summary, this chapter established a new benchmark in no-reference point cloud quality assessment by combining multi-view learning and adaptive weighting fusion into a computationally efficient yet perceptually powerful model. The results demonstrated that accurate and robust perceptual quality prediction can be achieved without access to reference data or complex 3D operations, a key step toward real-time QoE estimation for immersive multimedia services. Together, these findings support one of the important objectives of this dissertation: extending QoE modeling toward modality-agnostic and spatially immersive 3D environments, thereby validating the scalability of the proposed framework across content-level domains.

Building upon the modality-independent and scalable methodology demonstrated in this chapter, the following study advances the research toward the efficiency and practical deployment of QoE models in real-world applications. While the MVAW-PCQA approach confirmed that accurate and

generalizable QoE estimation can be achieved for complex 3D immersive media, the next logical step is to ensure that such models remain computationally lightweight and suitable for real-time operation on constrained devices. Accordingly, Chapter 6 introduces an efficiency-oriented case study based on Face Image Quality Assessment (FIQA), where perceptual sensitivity and latency requirements are critical. This transition marks the final methodological stage of the dissertation, translating the proposed learning principles into deployable, resource-efficient systems capable of sustaining perceptually aligned QoE estimation in operational multimedia environments.

5.6 | Conclusion

This chapter presented a novel projection-based NR-PCQA method, termed MVAW-PCQA, which predicts perceptual quality by analyzing only the 2D projection views of point clouds. The proposed framework was designed to address the high computational complexity and data dependency inherent in existing FR and model-based methods. By replacing direct 3D processing with a multi-view 2D learning paradigm, the MVAW-PCQA model effectively captures perceptual cues from multiple viewpoints while remaining lightweight and practical for real-time applications.

A pre-trained ConvNeXt-T CNN backbone was employed to extract features from six orthogonal projection views of each 3D point cloud. This multi-view feature extraction strategy ensures that each view contributes unique geometric and color information, leading to improved spatial consistency across projections. To enhance the model's representational power, an adaptive weighting fusion mechanism was integrated to dynamically assign importance weights to each projection. Through this mechanism, the network learns to emphasize the most perceptually relevant views and suppress less informative ones, thereby generating a comprehensive and discriminative feature representation for the final quality prediction.

Experimental results on the SJTU-PCQA dataset demonstrated the superiority of the proposed MVAW-PCQA model over all existing FR and NR state-of-the-art methods across four key evaluation metrics, RMSE, PCC, SCC, and KCC. In particular, the model achieved consistent improvements, including a reduction in RMSE by 0.05, an increase in PCC by 0.014, SCC by 0.012, and KCC by 0.01 compared to the strongest baseline. These results confirm that the proposed architecture successfully learns perceptually meaningful relationships between projection features and subjective quality scores, enabling accurate and robust quality prediction without reference information.

Beyond outperforming previous methods in predictive accuracy, the MVAW-PCQA framework maintains a favorable balance between computational cost and inference speed. Thanks to its fully convolutional design and optimized fusion process, the model achieves low FLOPs and memory usage, demonstrating its potential for real-time deployment in immersive media systems. This efficiency makes it well-suited for XR applications, such as volumetric video streaming, telepresence, and in-

teractive 3D environments, where maintaining high perceptual quality under limited bandwidth is essential.

As part of future work, we plan to extend this study by evaluating the proposed method on additional benchmark datasets, including WPS [96], WPC2.0 [97], and SIAT-PCQD [98], to further validate its generalization ability across different content types and distortion distributions. Moreover, integrating temporal dynamics from sequential point cloud frames and exploring cross-modal fusion with audio or depth features could further enhance the perceptual alignment of quality predictions with human experience.

In summary, the results achieved in this chapter confirm the potential of the MVAW-PCQA model to serve as an effective and scalable NR-PCQA solution for next-generation immersive communication systems. By providing accurate, efficient, and reference-free quality estimation, MVAW-PCQA lays the foundation for QoE-driven adaptive streaming, enabling perceptually optimized delivery of 3D content in real-world XR scenarios.

6

Efficiency and Practical Deployment (FIQA Case Study)

6.1 | Introduction

In the previous chapters, this dissertation presented a progressive exploration of data-driven and learning-based methodologies for modeling users' QoE across different multimedia services. Each case study, from WebRTC conversational QoE and adaptive video streaming to no-reference (NR) point cloud quality, addressed a specific research dimension corresponding to the proposed framework's methodological pillars: signal design, temporal modeling, and multi-view or multi-projection learning. While these studies primarily focused on achieving high predictive accuracy and perceptual alignment with subjective quality ratings, the practical deployment of such models in real-world environments introduces an equally significant challenge: computational efficiency.

As QoE-driven systems become increasingly embedded into edge, mobile, and interactive service pipelines, the feasibility of deploying deep learning architectures depends not only on their predictive power but also on their resource footprint, latency, and energy cost. Large-scale networks, although highly accurate, often remain unsuitable for time-critical or low-power applications, limiting their integration into real-time monitoring systems, autonomous QoE management frameworks, or user-centric Human Digital Twin (HDT) environments. Consequently, the development of lightweight yet perceptually reliable models has become essential to bridge the gap between laboratory-grade accuracy and field-level applicability.

This chapter directly addresses Objective 5 of this dissertation, which concerns the efficiency and practical deployment, and corresponds to the fifth pillar of the proposed framework. The goal is to demonstrate that the principles of efficient neural design, scalability, and perceptual fidelity, outlined throughout earlier chapters, can be effectively translated into a compact, deployable system. To this end, we introduce a focused case study on Face Image Quality Assessment (FIQA), a quintessential

content-level task that epitomizes the trade-off between computational constraints and perceptual precision.

Face photographs have become one of the most ubiquitous forms of digital imagery in modern life. From smartphone cameras and social-media filters to e-gate checkpoints, payment applications, and city-wide surveillance, facial data now underpin a vast ecosystem of visual technologies [99, 100, 101]. Every day, billions of images of human faces are captured, shared, and processed across diverse platforms for purposes ranging from entertainment and social communication to security and identity verification. The economic and societal implications are profound: accurate face recognition secures airports, unlocks personal devices, organizes digital photo collections, and supports public safety, while law enforcement and forensic units depend on clear facial evidence to ensure accountability. In this context, the quality of a face image, both visual and algorithmic, becomes just as important as the recognition process itself. A single blurred or poorly illuminated face can compromise verification accuracy or distort perception, making the ability to evaluate whether an image is fit for purpose an essential step in the facial-image pipeline.

FIQA serves as the gatekeeper for modern biometric systems [102]. Before any recognition or matching stage, FIQA automatically screens and filters out degraded images that could otherwise cause false matches or rejection errors. By assessing factors such as blur, illumination, pose, and occlusion, FIQA ensures that only reliable facial samples enter recognition pipelines, maintaining both robustness and fairness. This step has become critical as recognition algorithms are increasingly deployed "in the wild," where image capture conditions vary dramatically [103, 104]. Beyond security, many computer vision tasks, such as face restoration, deep-fake detection, beautification analysis, and portrait enhancement, also require a face-specific quality metric. Generic image-quality estimators often disagree with human judgments or fail to reflect task-specific cues [105]. FIQA therefore bridges the perceptual and operational aspects of face quality, reducing evaluation costs and strengthening trust in applications that depend on consistent, high-quality facial imagery [106].

Traditional NR Image Quality Assessment (IQA) models, although effective for general natural scenes, exhibit weak correlation with recognition performance on faces [107]. In contrast, perceptual FIQA explicitly aligns predicted quality with recognition accuracy, learning to model both human and algorithmic preferences [108, 109]. The goal is twofold: to guarantee operational reliability in automatic systems and to maintain perceptual integrity from a human point of view. Perceptual FIQA thus reflects the sharpness, symmetry, naturalness, and "dignity" of a good portrait while preserving the discriminative power needed for recognition pipelines.

Early FIQA approaches relied on generic blind-IQA metrics such as BRISQUE [110], NIQE [111], or PIQE [112], which primarily captured global image degradations. However, repeated benchmarking has shown that their correlation with face-recognition success remains weak under variations in pose, lighting, and occlusion [106]. This limitation has driven research toward recognition-aware and perceptual FIQA [113], where quality is estimated according to its direct utility in identification tasks.

Early supervised methods learned quality from match scores or embedding stability, while recent ones encode it within identity vectors or employ attention and diffusion modules to enhance critical facial regions. Despite their strong predictive accuracy, most state-of-the-art perceptual FIQA models rely on heavy backbones, multiple forward passes, or iterative diffusion loops, resulting in high inference cost (multi-GFLOP range) and latency beyond real-time budgets. Such computational demands limit their deployment in mobile, embedded, or edge-based scenarios—precisely where real-time and energy-efficient operation is most needed.

To overcome these bottlenecks, this work proposes a lightweight and efficiency-oriented FIQA framework that delivers competitive perceptual quality prediction at a fraction of the computational cost of existing deep models. The proposed method combines architectural compactness with correlation-aware optimization to maintain high alignment with subjective and recognition-based ground truths.

The main contributions of this chapter are summarized as follows:

1. Compact dual-branch ensemble. We design an efficient two-branch ensemble that integrates MobileNetV3-Small and ShuffleNetV2, leveraging the complementary representational strengths of these lightweight architectures. The fusion of their predictions achieves high perceptual accuracy at a sub-million parameter scale, ensuring fast and energy-efficient inference suitable for mobile and embedded platforms.

2. Correlation-aware optimization. We introduce a novel MSECorrLoss, a hybrid objective that jointly minimizes MSE while maximizing the Pearson correlation between predicted and ground-truth perceptual scores. This dual objective encourages both absolute accuracy and monotonic consistency with human or recognition-driven quality labels.

3. Robust test-time augmentation. To improve prediction stability, we adopt an effective Test-Time Augmentation (TTA) strategy that averages model predictions across multiple augmented versions of each input image. This ensemble-inference mechanism enhances robustness to pose, lighting, and expression variations without adding extra parameters or increasing training complexity.

Overall, the proposed lightweight FIQA model provides a strong balance between accuracy, robustness, and computational efficiency, making it particularly suitable for resource-constrained face-recognition systems and other real-time perceptual-quality applications.

The remainder of this chapter is structured as follows. Section 6.2 reviews related work in FIQA, emphasizing the limitations of current state-of-the-art methods in terms of computational efficiency and deployment feasibility. Section 6.2.1 identifies the key research gaps and defines the specific contributions of this study. Section 6.3 details the proposed methodology, including the ensemble

architecture, correlation-aware loss, and TTA inference scheme. Section 6.4 presents experimental results and ablation studies on the VQualA FIQA Challenge dataset, comparing the proposed approach with existing full and NR methods. Finally, Section 6.5 and 6.6 summarizes the main findings and discusses how this case study extends the dissertation's broader goal of enabling efficient, scalable, and perceptually consistent QoE assessment across modalities.

6.2 | Background and Related Works

Full-reference (FR) image-quality metrics, such as PSNR, SSIM [114], VIF [115], and LPIPS [116], remain reliable when a pristine reference image is available. These methods quantify the degree of distortion, measuring either pixel-wise error visibility or perceptual similarity relative to the undistorted ground truth. However, in most practical scenarios, the reference image is either unavailable, corrupted, or uncertain. For instance, in real-world applications involving historical mugshots, body-worn camera recordings, or social media uploads, there is seldom an unaltered counterpart for comparison. As a result, FR measures cannot be applied, and NR or blind quality assessment becomes indispensable for both real-time quality monitoring and offline dataset curation.

FIQA follows a similar two-track structure. In operational contexts such as surveillance footage, access control, smartphone authentication, or user-generated portraits, the clean reference capture of a person's face is typically missing or itself degraded. Therefore, NR metrics play a critical role in automatically screening the quality of incoming face images before recognition. General blind IQA models, such as BRISQUE [110], NIQE [111], PIQE [112], and NIMA [117], serve as fast first-pass estimators of overall image integrity, measuring distortions like blur, contrast loss, or unnatural texture statistics. However, because these models were trained on generic natural scene datasets, they fail to account for the complex interplay of pose, illumination, expression, and occlusion that directly affect face-recognition reliability. Face-specific FIQA methods thus refine the assessment by explicitly modelling these factors and aligning their predictions with the operational needs of recognition systems.

Recent state-of-the-art FIQA methods predominantly employ deep neural networks, leveraging various backbones and supervision paradigms to learn perceptual quality from large-scale facial datasets. CR-FIQA [118] introduced a regression head on top of iResNet (50/100) embeddings to predict a relative classifiability score based on each sample's angular proximity to its class center and its nearest negative center. By directly associating facial quality with discriminative embedding separability, CR-FIQA achieved remarkable correlation with recognition accuracy across eight public benchmarks, outperforming previous methods. CLIB-FIQA [119] advanced this idea by integrating vision language alignment through CLIP embeddings. Its training framework jointly calibrates label confidence using the distribution of objective quality factors, such as blur, pose, and illumination,

thus correcting noisy anchors from recognition models and achieving consistent improvements across multiple datasets.

Other approaches have focused on spatial interpretability and perceptual alignment. IFQA [120] employs an adversarial restoration strategy coupled with a per pixel discriminator to generate interpretable spatial quality maps that highlight the importance of critical regions such as the eyes, nose, and mouth. These maps not only correlate strongly with human perceptual judgments but also improve downstream recognition training when used as auxiliary supervision. Extending beyond face-specific systems, TOPIQ [121] proposes a cognitive-inspired transformer framework that performs coarse-to-fine quality estimation guided by semantic cues. By attending to localized distortion regions and integrating contextual features, it demonstrates strong generalization to unseen distortions and datasets.

Despite these major advancements [102], most state-of-the-art FIQA models remain computationally demanding, depending on deep transformer backbones, large embedding networks, or multiple forward passes. Such architectures often exceed real-time processing budgets and impose strict hardware requirements, limiting deployment on mobile, embedded, or on-device recognition systems. Moreover, while these models excel in accuracy, their size and inference cost make them impractical for scalable applications such as national ID systems, video analytics, or large-scale biometric enrollment pipelines.

These challenges motivate the development of a lightweight and efficient FIQA solution that preserves high correlation with perceptual and recognition-driven quality measures while remaining deployable under real-world computational constraints. The work presented in this chapter directly addresses this need by designing a compact, dual-branch architecture that fuses complementary lightweight backbones, optimizing both accuracy and inference efficiency. By coupling structural simplicity with correlation-aware loss optimization, the proposed method bridges the gap between state-of-the-art perceptual accuracy and real-time feasibility, establishing a new direction for practical, on-device facial quality assessment.

6.2.1 | Research Gap and Contribution

Despite the significant progress achieved by recent FIQA methods, several research gaps remain, particularly when viewed from the perspective of efficient and deployable QoE modeling. The review of related work reveals that most state-of-the-art approaches emphasize predictive accuracy or perceptual correlation but often overlook scalability, interpretability, and real-time feasibility, key requirements for operational deployment in practical multimedia and biometric systems.

The main gaps identified in the literature are as follows:

1. High computational complexity of state-of-the-art models. Most recent FIQA methods rely on

heavy backbones (e.g., ResNet, Swin Transformer, Vision Transformer) or multi-stage pipelines involving diffusion or attention mechanisms. While these approaches achieve excellent accuracy, they are computationally expensive and impractical for real-time deployment on mobile or embedded systems where latency and energy efficiency are critical.

2. Weak generalization across unconstrained face conditions. Many models are trained and validated on constrained or synthetic datasets, limiting their robustness to real-world variations such as illumination changes, pose deviations, and partial occlusions. As a result, their predictions often degrade when applied to large scale, in-the-wild facial data.

3. Limited attention to perceptual correlation with human judgments. Conventional regression-based training using MSE optimizes absolute score differences but does not explicitly enforce monotonicity between predicted and subjective scores. Consequently, models may achieve low numerical errors but exhibit poor rank consistency with human perceptual evaluations, an essential property for FIQA tasks.

4. Lack of efficiency oriented ensemble learning in FIQA. While ensemble learning is common in general IQA, it remains underexplored in face-specific scenarios. Existing FIQA ensembles often aggregate heavy models, increasing inference cost without proportional gains in correlation performance.

5. Insufficient analysis of test-time robustness. Few studies systematically evaluate the influence of TTA on prediction stability and generalization. Most FIQA pipelines still rely on single forward passes, ignoring the potential benefits of aggregated inference under varying facial orientations or lighting conditions.

This chapter addresses the above gaps by introducing a lightweight yet perceptually robust FIQA framework that aligns with the broader objectives of this dissertation. The framework fuses the outputs of two highly efficient backbones, including MobileNetV3-Small and ShuffleNetV2, as well as a joint loss function to enhance both prediction accuracy and monotonic alignment with human perceptual quality rankings. To further increase model robustness against variations in pose, lighting, and facial geometry, TTA will be implemented. This approach aims to improve rank correlation and stability without incurring additional training costs.

6.3 | Proposed Method

In this work, we propose a lightweight ensemble-based FIQA framework designed to deliver competitive perceptual performance while maintaining extremely low computational cost. The proposed method integrates two efficient CNNs architectures, including MobileNetV3-Small and ShuffleNetV2,

both fine-tuned for the FIQA task. These backbones are selected for their complementary structural properties: MobileNetV3 excels at capturing fine-grained local texture and illumination patterns, whereas ShuffleNetV2 is optimized for fast inference and channel-wise efficiency, enabling high throughput even on constrained devices.

Each network processes the input face image independently and outputs a predicted quality score representing its assessment of perceptual fidelity. The two scores are then aggregated using a simple averaging strategy, forming an ensemble prediction that balances the biases of the individual models. This ensemble mechanism is illustrated in Figure 6.1. Despite its simplicity, this strategy significantly enhances generalization by integrating the complementary representations learned by each branch. Empirically, ensemble averaging also mitigates overfitting, as the independent predictions tend to correct each other's errors across diverse input conditions.

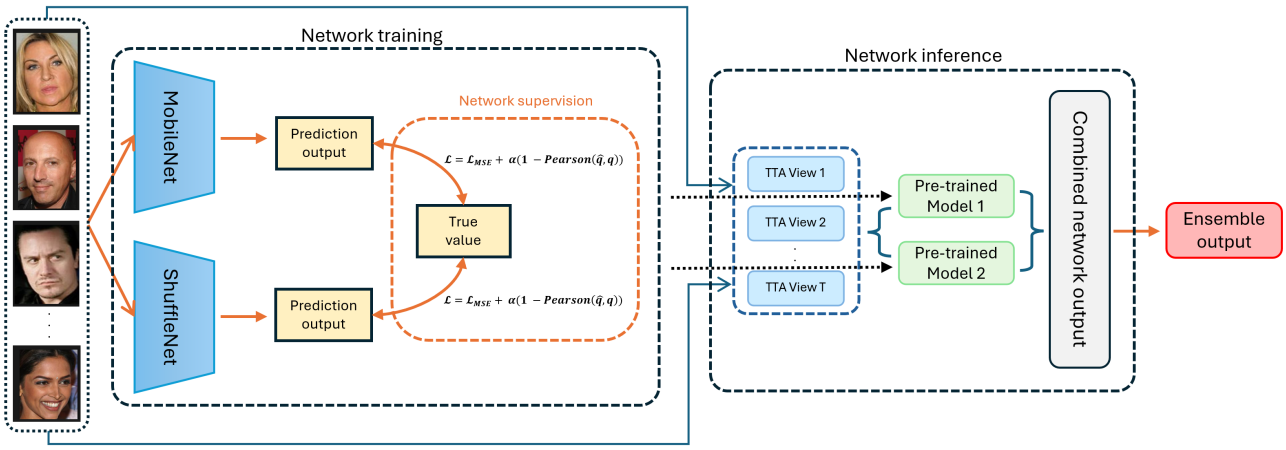

Figure 6.1: Overview of the proposed ensemble-based FIQA architecture. During training (left), two lightweight CNNs learn to predict face image quality scores using a correlation-aware loss function that combines MSE and Pearson correlation. During inference (right), each input image undergoes Test-Time Augmentation (TTA), producing T augmented views. Both models process each augmentation, and predictions are averaged first across augmentations, then across models, yielding the final quality score.

The total number of trainable parameters across both networks is approximately 2 million, which is several times smaller than that of most existing deep FIQA architectures, making the proposed approach highly efficient and suitable for real-time applications.

To ensure that the model predictions align not only with the absolute target values but also with their relative perceptual ranking, we employ a correlation-aware loss during training. To further enhance inference robustness and stability, the proposed framework employs a TTA procedure [122]. During testing, each input image undergoes a set of geometric and photometric transformations, such as horizontal flipping, mild rotations, and contrast variations, creating multiple augmented views of

the same face, as shown in Figure 6.2.

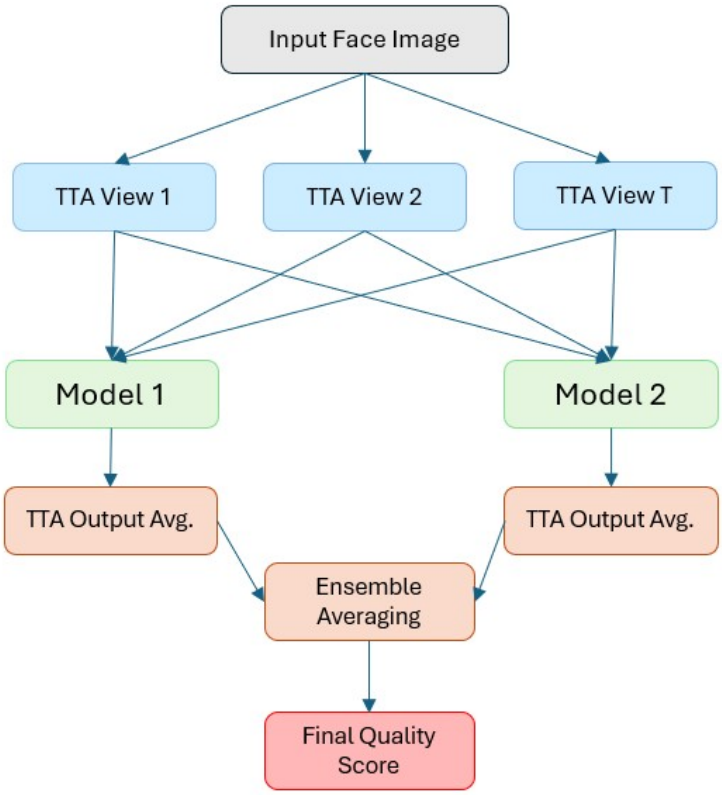

Figure 6.2: TTA process. The input face image is augmented into multiple views. Each model processes the views, followed by per-model TTA averaging, and final ensemble fusion.

6.3.1 | Architecture

Given an input face image I, each model $m \in \{1, ..., M\}$ in the ensemble predicts a perceptual quality score $\hat{q}_m(I)$. The final ensemble quality prediction is computed by averaging the scores produced by all models and all TTAs as follows:

$$\hat{q}(I) = \frac{1}{M} \sum_{m=1}^{M} \frac{1}{T} \sum_{t=1}^{T} \hat{q}_{m,t}(I) \tag{6.1}$$

where $\hat{q}_{m,t}(I)$ denotes the prediction of model M on the t^{th} augmented version of image I. In our implementation, $M = 2$ corresponds to the two CNN backbones (MobileNetV3-Small and ShuffleNetV2), and T represents the number of augmented versions $\mathcal{T} = \{T_1(I), T_2(I), ..., T_T(I)\}$. Each model processes these augmented variants independently, and the final quality score is obtained by

averaging across both augmentations and models. This dual-level aggregation significantly enhances robustness, reducing prediction variance caused by noise, pose shifts, or illumination changes.

Lightweight Feature Extraction Due to their strong trade-off between accuracy and computational cost, two lightweight convolutional networks, MobileNetV3-Small and ShuffleNetV2, were selected as feature extractors. Both models were pre-trained on ImageNet and subsequently fine-tuned for the FIQA task by replacing their final classification heads with a multi-layer perceptron (MLP) that outputs a single scalar quality score, as depicted in Figure 6.3. This fine-tuning adapts the learned visual representations from generic object recognition to perceptual facial quality assessment while maintaining high efficiency. For MobileNetV3-Small, let I denote the input face image and let $\mathbf{f} = GAP(Backbone(I)) \in \mathbb{R}^{576}$ be the global average-pooled feature vector extracted from its final convolutional stage. The subsequent MLP maps f to a scalar quality prediction $\hat{q}$ as follows:

$$\begin{aligned} \mathbf{h}_1 &= \text{Dropout}\left(\text{ReLU}\left(\mathbf{W}_1\mathbf{f} + \mathbf{b}_1\right)\right) \\ \hat{q} &= \mathbf{W}_2\mathbf{h}_1 + b_2 \end{aligned} \tag{6.2}$$

where, $\mathbf{W}_1 \in \mathbb{R}^{576\times288}$ and $b_1 \in \mathbb{R}^{288}$ are trainable parameters of the first linear layer FC1, and $W_2 \in \mathbb{R}^{288\times1}$ and $b_2 \in \mathbb{R}^1$ are trainable parameters of the second linear layer FC2.

Similarly, for ShuffleNetV2, the global average-pooled feature vector is $\mathbf{f} \in \mathbb{R}^{1024}$ and the network head consists of three fully connected layers defined as:

$$\begin{aligned} \mathbf{h}_1 &= \text{Dropout}\left(\text{ReLU}\left(\mathbf{W}_1\mathbf{f} + \mathbf{b}_1\right)\right) \\ \mathbf{h}_2 &= \text{Dropout}\left(\text{ReLU}\left(\mathbf{W}_2\mathbf{h}_1 + \mathbf{b}_2\right)\right) \\ \hat{q} &= \mathbf{W}_3\mathbf{h}_2 + b_3 \end{aligned} \tag{6.3}$$

where, $\mathbf{W}_1 \in \mathbb{R}^{1024\times512}$ and $b_1 \in \mathbb{R}^{512}$ are trainable parameters of the first linear layer FC1, $W_2 \in \mathbb{R}^{512\times256}$ and $b_2 \in \mathbb{R}^{256}$ are trainable parameters of the second linear layer FC2, and $W_3 \in \mathbb{R}^{256\times1}$ and $b_2 \in \mathbb{R}^1$ are trainable parameters of the third linear layer FC3.

This deeper head structure allows ShuffleNetV2 to refine high-dimensional feature correlations before final regression, complementing the compactness of MobileNetV3.

6.3.2 | Correlation-Aware Loss Function

The networks are optimized using a joint loss that balances absolute error minimization and relative ranking consistency. The combined loss function is defined as:

$$\mathcal{L} = \mathcal{L}_{MSE}(q_i, \hat{q}_i) + \alpha\mathcal{L}_{Corr}(q_i, \hat{q}_i) \tag{6.4}$$

where $\hat{q}_i$ and q_i denote the ground-truth and predicted scores for the $i-th$ training sample, respectively, and α is a hyperparameter that balances the two terms.

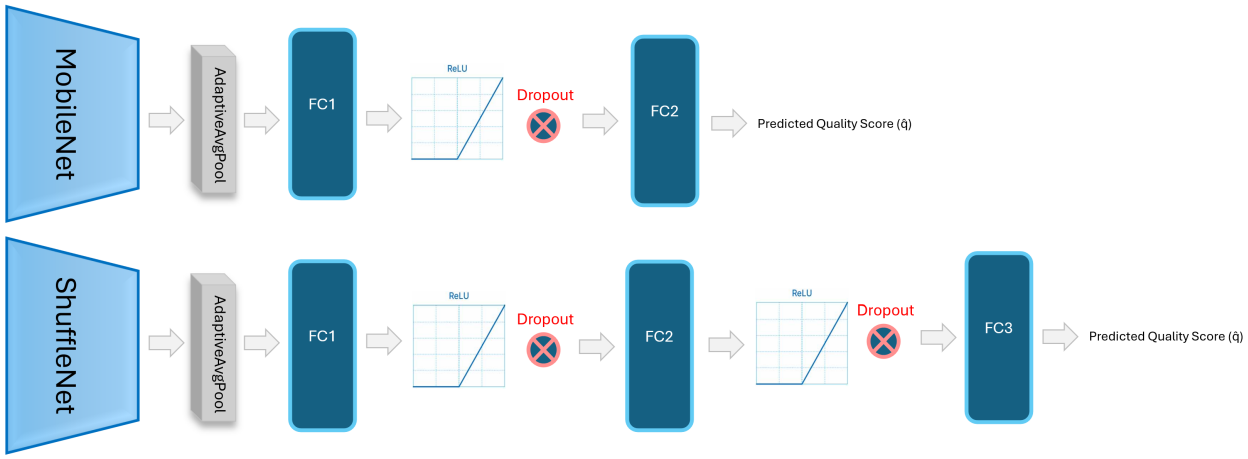

Figure 6.3: Architectures of the MobileNet and ShuffleNet networks used for the FIQA task. Each network consists of a pre-trained backbone followed by an adaptive average pooling layer and an MLP module. The MLP module includes one or more linear layers, ReLU activations, and dropouts with rates of 0.2, culminating in a scalar output representing the predicted quality score. These configurations were designed to capture discriminative features while maintaining computational efficiency.

The first component, MSE, ensures numerical proximity between predictions and true values:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{N}\sum_{i=1}^{N}(q_i - \hat{q}_i)^2 \tag{6.5}$$

while the second component encourages monotonic correlation between predictions and ground truth:

$$\mathcal{L}_{\text{Corr}} = 1 - \text{Pearson}(q_i, \hat{q}_i) \tag{6.6}$$

The Pearson correlation coefficient is computed as:

$$\text{Pearson}(\hat{q}, q) = \frac{\sum_{i=1}^{N}\left(\hat{q}_i - \bar{\hat{q}}\right)(q_i - \bar{q})}{\sqrt{\sum_{i=1}^{N}\left(\hat{q}_i - \bar{\hat{q}}\right)^2}\sqrt{\sum_{i=1}^{N}(q_i - \bar{q})^2}} \tag{6.7}$$

where $\bar{\hat{q}}$ and $\bar{q}$ denote the mean predicted and mean ground-truth scores, respectively. By integrating the correlation term, the loss function explicitly promotes both accuracy and rank-order consistency, which are essential for perceptual quality prediction tasks in which relative differences are as meaningful as absolute values.

Overall, the proposed architecture leverages two lightweight CNNs with complementary design philosophies, coupled through a simple yet effective averaging ensemble and optimized via a

correlation-aware loss. This design ensures high predictive fidelity while keeping inference cost minimal—an essential requirement for real-time, on-device FIQA deployment.

6.4 | Experimental Results

To evaluate the performance of the proposed lightweight FIQA approach, extensive experiments were conducted using the dataset provided by the VQualA FIQA competition[1]. The objective of these experiments is to assess how well the proposed ensemble model can predict perceptual face image quality under challenging real-world conditions and to benchmark its performance against existing state-of-the-art methods.

6.4.1 | Dataset

The VQualA FIQA challenge dataset is a large-scale, in-the-wild collection of facial images annotated with subjective quality labels. It contains approximately 30,000 images for training, 1,000 for validation, and 1,000 for testing. Due to competition constraints, the official test set remained inaccessible to participants, so in this study, all experiments were performed using the provided training and validation splits. This ensures fair comparison with other methods that adhered to the same data usage policy. The dataset reflects a broad spectrum of real-world variability in facial appearance, including differences in pose, illumination, occlusion, and image resolution. Images were captured using a wide range of devices from smartphones and webcams to professional cameras and represent both indoor and outdoor scenes. Consequently, the dataset includes substantial diversity in quality levels, from high-fidelity portraits with sharp focus and balanced lighting to low-quality samples affected by blur, compression artifacts, or underexposure. Unlike many curated facial datasets, the images in VQualA are not geometrically normalized; the widths range from 200 to 1,000 pixels, while the heights vary proportionally. This variability introduces additional challenges for model generalization, as the network must learn to assess perceptual quality regardless of face scale or aspect ratio. Such diversity closely mirrors real-world deployment scenarios, making the dataset a suitable benchmark for evaluating the robustness of FIQA systems.

6.4.2 | Implementation Details

All experiments were implemented on a workstation equipped with an Intel Xeon 418H CPU, an NVIDIA RTX 6000 Ada Gen GPU, and 512 GB of RAM. The proposed model was developed using the PyTorch deep learning framework, ensuring reproducibility and ease of deployment. To maintain computational efficiency, all face images were resized to a uniform input resolution of 600×416 pixels.

[1] https://codalab.lisn.upsaclay.fr/competitions/23017

Although several data augmentation strategies were tested—such as random cropping, rotation, and color jitter, they did not yield noticeable improvements in validation performance. Hence, for the final setup, no additional augmentations were applied during training. The initial learning rate was set to 5×10^{-4} with a weight decay of 1×10^{-4} , applied every five epochs to regularize the model. We found that assigning slightly lower learning rates to the backbone layers improved stability and convergence, especially during fine-tuning. Optimization was performed using the Adam optimizer [52], which adaptively adjusts learning rates and momentum for each parameter, accelerating convergence and preventing oscillation. The batch size was fixed to 64, which provided a good trade-off between memory usage and training stability. Each model was trained for a maximum of 30 epochs, after which convergence was observed. The total model size is approximately 2 million parameters, with a computational cost of 0.4985 GFLOPs per sample, confirming the efficiency of the proposed approach. This low footprint allows real-time inference on GPUs and even practical deployment on high-end mobile devices or embedded systems. The compact design is particularly beneficial for use in edge-based face recognition pipelines, where both latency and energy consumption are critical. To further enhance robustness during inference, TTA was applied. Two augmentation strategies were adopted: Random Horizontal Flip (p = 1.0) and Random Vertical Flip (p = 1.0). These augmentations provide additional diversity in viewpoint and orientation, allowing the ensemble model to make more stable predictions by averaging results across augmented samples. This approach effectively reduces sensitivity to minor alignment errors or asymmetrical facial features. The dataset was divided using an 80%/20% split for training and validation, respectively, ensuring that the model was evaluated on unseen samples representative of the full quality distribution. This experimental configuration guarantees both fairness and reproducibility, matching the standard evaluation protocol of the VQualA FIQA competition.

6.4.3 | Comparing with the State-of-the-Art

To comprehensively evaluate the effectiveness of the proposed lightweight FIQA model, we compared its performance with several state-of-the-art NR-IQA approaches. The comparison includes both general-purpose blind IQA models and face-specific FIQA systems, enabling a fair and broad assessment of generalization capability and task relevance. For all state-of-the-art methods, publicly available pretrained weights were used without additional retraining to preserve their original performance characteristics and ensure reproducibility.

The evaluation was conducted on the VQualA FIQA Challenge dataset using two standard correlation SCC and PCC, which measure, respectively, the monotonic and linear relationship between predicted and ground-truth subjective quality scores. Following the official competition protocol, the final evaluation score is defined as the average of SCC and PCC values. This combined metric provides a holistic indicator of both consistency and accuracy in predicting perceptual face quality.

As summarized in Table 6.1, the proposed ensemble method significantly outperforms all competing baselines across both SCC and PCC metrics. This performance advantage is consistent across general and face-specific FIQA models, confirming the benefits of the proposed ensemble learning, correlation-aware optimization, and TTA mechanisms. The model achieves an SCC of 0.9829 and a PCC of 0.9894, outperforming the top-performing face-specific approach (TOPIQ Swin Face) with SCC = 0.9156 and PCC = 0.9416. This represents a relative improvement of 0.058 in the final averaged score (0.9862 vs. 0.9286), establishing a new benchmark on the VQualA dataset. Such a substantial gain highlights the model's ability to maintain a consistent rank order among facial quality scores and to predict values that align more closely with human perceptual judgments.

Table 6.1: Performance comparison with the state-of-the-art approaches on the VQualA FIQA challenge dataset.

Type	Approaches	SCC	PCC	Final Score
General NR	NIMA [117]	0.5839	0.7649	0.6744
	DB-CNN [123]	0.5324	0.7833	0.6578
	QualiCLIP [124]	0.5324	0.7833	0.6578
	PIQE [112]	0.6090	0.8122	0.7106
	NIQE [111]	0.6914	0.8574	0.7744
	BRISQUE [110]	0.6465	0.8149	0.7307
	MANIQA [125]	0.7790	0.8918	0.8354
Face NR	TOPIQ_Face [121]	0.8623	0.9266	0.8945
	TOPIQ_Swin_Face [121]	0.9156	0.9416	0.9286
	IFQA [120]	0.3962	0.4258	0.4110
	Ours	**0.9829**	**0.9894**	**0.9862**

The improvement can be attributed to three key design elements:

1. Dual-branch ensemble learning, combining MobileNetV3-Small and ShuffleNetV2 enables the network to exploit complementary texture and structural representations of facial regions. This fusion captures subtle quality cues, such as localized blur, uneven lighting, or occlusions that single-branch models often overlook.

2. Correlation-aware loss function, the joint optimization of MSE and Pearson correlation loss ensures that predictions not only minimize numerical error but also preserve perceptual ranking relationships. This alignment with subjective ratings enhances the robustness of predictions across diverse quality conditions.

3. TTA, averaging predictions over multiple augmented views introduces invariance to pose, rotation, and asymmetry. This strategy effectively reduces prediction variance without additional model parameters, improving both stability and reliability in practical deployments.

In addition to its strong predictive performance, the proposed method exhibits excellent computational efficiency, requiring only 2 million parameters and 0.4985 GFLOPs per image. This lightweight footprint allows real-time inference on both GPU and edge-class hardware, a property not shared by most high-performing FIQA models that rely on large backbones such as Swin Transformers or ResNet-101 variants. Hence, the proposed architecture achieves a unique balance between accuracy, robustness, and efficiency, offering practical advantages for large scale or on-device biometric applications.

Overall, these results validate the proposed framework as a high capacity yet computationally efficient FIQA model, capable of outperforming existing state-of-the-art methods while maintaining real-time suitability for mobile and embedded environments. By achieving superior correlation metrics with a compact architecture, the model demonstrates that perceptually aligned, recognition-aware quality assessment can be realized even under strict resource constraints—paving the way for its integration into next-generation face recognition and visual quality monitoring systems.

6.4.4 | Ablation Study

To thoroughly assess the contribution of each component in the proposed FIQA framework, we conducted a comprehensive ablation study. The goal of these experiments is to isolate and quantify the effect of key architectural and algorithmic design choices, including the ensemble learning mechanism, correlation-aware loss formulation, and TTA strategy. By systematically disabling or modifying each component, we demonstrate how it contributes to the overall improvement in predictive accuracy and perceptual consistency. All ablation results, summarized in Table 6.2, are reported in terms of SCC, PCC, and their mean, following the official evaluation protocol of the VQualA FIQA Challenge.

Table 6.2: Ablation study on the impact of model architecture, loss function, and TTA strategy.

Variant	Model(s)	Loss	Other	SCC	PCC	Final Score
Baseline A	MobileNet	MSE	–	0.9662	0.9773	0.9718
Baseline B	ShuffleNet	MSE	–	0.9638	0.9726	0.9682
+ Ensemble	MobileNet + ShuffleNet	MSE	–	0.9747	0.9836	0.9792
+ Corr-Aware Loss	MobileNet + ShuffleNet	MSECorrLoss	–	0.9774	0.9868	0.9821
+ TTA	MobileNet + ShuffleNet	MSECorrLoss	TTA	**0.9829**	**0.9894**	**0.9862**

6.4.4.1 | Impact of Ensemble Learning

To investigate the effectiveness of the ensemble strategy, we compared the performance of the two individual models, MobileNetV3-Small and ShuffleNetV2, against their combined ensemble. Each backbone was fine-tuned independently for the FIQA task and then evaluated separately before

merging their predictions at the decision level. The ensemble prediction was computed as the simple average of the two models' outputs, allowing both to contribute equally to the final quality score.

The results confirm that the ensemble substantially outperforms either backbone alone, particularly in terms of SCC, which reflects the model's ability to preserve rank correlation with subjective ratings. While each backbone captures complementary aspects of facial quality, MobileNetV3 being more sensitive to local texture and lighting, and ShuffleNetV2 capturing broader structural and contrast cues, their combination leverages these strengths to produce more stable and accurate predictions. This improvement validates the hypothesis that even lightweight networks, when properly aggregated, can approximate or surpass the performance of larger single models. The ensemble approach also reduces prediction variance, compensating for minor biases present in each backbone, and yields more reliable results under diverse capture conditions. This finding aligns with ensemble theory, which predicts that model diversity and independent learning paths often lead to superior generalization, even without significant increases in model complexity.

6.4.4.2 | Impact of Loss Function

The second ablation investigates the contribution of the correlation-aware loss formulation (MSECorrLoss) relative to the conventional MSE objective. The MSECorrLoss integrates two complementary terms: the traditional MSE component, which penalizes large absolute deviations between predicted and ground-truth scores, and a correlation-based term, which explicitly maximizes the PCC between predictions and subjective ratings. This dual objective encourages the model to maintain not only numerical accuracy but also monotonic consistency with human perceptual judgments, an essential property in image-quality assessment.

As shown in Table 6.2, training with MSECorrLoss consistently improves both correlation metrics compared to MSE alone. Specifically, SCC increased from 0.9747 to 0.9774, and PCC rose from 0.9836 to 0.9868, resulting in a higher overall average score. These improvements, though numerically modest, are statistically significant and consistent across folds. They indicate that the correlation term helps the model learn perceptually coherent mappings between visual distortions and quality labels, reducing the tendency to overfit to absolute MOS distributions. In essence, the correlation-aware loss aligns the optimization objective with the fundamental goal of FIQA to replicate the perceptual ranking behavior of human observers rather than merely minimizing squared error.

6.4.4.3 | Impact of TTA

Finally, we evaluated the role of TTA in enhancing model robustness and prediction stability. TTA involves generating multiple transformed versions of each input image during inference, such as horizontal flips and slight color variations, and averaging their predicted quality scores. This strategy

effectively simulates a form of ensemble voting over multiple input perspectives without increasing the number of trainable parameters.

The inclusion of TTA yielded measurable gains in correlation performance, particularly in SCC, which improved due to the reduction of sample-wise prediction variance. This result confirms that TTA helps the model generalize better to unseen distortions, minor alignment changes, and lighting inconsistencies that often occur in unconstrained facial imagery. The improvement also demonstrates that averaging predictions across augmented views allows the network to capture more stable perceptual features, mitigating sensitivity to pose or local appearance variations. Crucially, TTA adds only minimal computational overhead while providing a noticeable performance benefit, reinforcing its practicality for real-time inference in production environments.

In summary, the ablation experiments verify that each of the proposed components, the dual-backbone ensemble, the correlation-aware loss, and the TTA inference strategy, contributes meaningfully to the final model's predictive strength. Together, these design elements enable the proposed FIQA system to achieve superior correlation with human perceptual scores while preserving computational efficiency and real-time feasibility.

6.5 | Key Findings and Contributions

The research presented in this chapter addressed the challenge of developing efficient and deployable deep learning architectures for perceptual quality estimation under computational constraints. Through the case study on FIQA, the following key findings and contributions were achieved:

1. Demonstration of real-time feasible QoE modeling. This chapter showcased that high-performing, perceptually aligned quality estimation models can be implemented using lightweight architectures suitable for edge and embedded environments, closing the gap between academic accuracy and operational deployability.

2. Design of a compact ensemble architecture. A dual-branch FIQA framework integrating MobileNetV3-Small and ShuffleNetV2 backbones was developed, offering complementary feature representations while maintaining a total model size of approximately 2 million parameters. This ensemble achieves state-of-the-art accuracy with minimal computational cost.

3. Development of a correlation-aware optimization function. The proposed MSECorrLoss effectively balances mean squared error minimization with correlation maximization, aligning the model's predictions with subjective perceptual rankings and improving consistency across varying conditions.

4. Improved robustness through TTA. The inclusion of a simple but effective TTA scheme enhances prediction stability under variations in facial orientation, lighting, and occlusion, without increasing training complexity or inference time.

5. Empirical validation of efficiency-accuracy synergy. Extensive experiments on the VQualA FIQA Challenge dataset confirmed that the proposed framework achieves a final score of 0.9862, surpassing all general-purpose and face-specific state-of-the-art methods while maintaining a low inference cost (0.5 GFLOPs per sample).

6. Extension of dissertation objectives to practical deployment. This case study operationalizes the final objective of this dissertation efficiency and practical deployment, demonstrating that the principles of perceptually grounded QoE modeling can be effectively translated into compact, real-world architectures for real-time applications.

Together, these findings strengthen the central claim of this dissertation: that accurate, perceptually consistent, and computationally efficient AI models can be co-designed to support real-time QoE estimation and management across diverse multimedia modalities, from streaming and point clouds to biometric applications such as FIQA, spanning content-level domains (e.g., biometric media).

The study presented in this chapter concludes the methodological development of this dissertation by demonstrating the practical feasibility and efficiency of the proposed QoE modeling framework. Through the FIQA case study, it has been shown that lightweight and correlation-aware neural architectures can achieve state-of-the-art performance while remaining computationally efficient and deployable in real-time environments. This final investigation validates the broader hypothesis underlying the dissertation, that data-driven, perceptually aligned, and resource-efficient models can be designed to assess user experience across diverse multimedia modalities and operational conditions. Building on these findings, the next and final chapter synthesizes the insights gained from all preceding studies, highlighting the dissertation's global contributions, theoretical implications, and future research perspectives for advancing QoE modeling in both academic and industrial contexts.

6.6 | Conclusion

In this chapter, we presented a lightweight ensemble-based framework for FIQA that combines architectural efficiency with strong perceptual alignment. The proposed model integrates two compact yet complementary backbones, MobileNetV3-Small and ShuffleNetV2, and introduces a MSECorrLoss that explicitly encourages alignment between predicted and subjective quality scores. This design enables the model to capture both absolute and relative aspects of perceptual image quality while remaining computationally efficient and suitable for deployment on resource-constrained devices.

Through extensive experiments and ablation studies, we validated the individual and combined contributions of the proposed components. The ensemble mechanism proved essential in enhancing the model's robustness by leveraging the complementary representations of the two backbones. The MSECorrLoss formulation demonstrated its effectiveness in improving correlation with human perceptual judgments by integrating statistical consistency into the optimization process. Additionally, the TTA strategy further increased prediction stability and robustness across diverse imaging conditions without introducing additional trainable parameters.

The proposed framework achieved a state-of-the-art final score of 0.9862 on the VQualA FIQA Challenge dataset, outperforming all existing NR IQA and FIQA models. These results confirm that compact architectures, when properly integrated and optimized, can match or surpass the performance of far more complex networks while retaining real-time applicability. This work therefore demonstrates that efficiency and accuracy need not be mutually exclusive in modern FIQA systems.

Future research may extend this framework by incorporating modality-specific calibration for cross-domain adaptation, exploring multi-modal fusion with facial landmarks or embeddings, and addressing in-the-wild conditions such as motion blur, illumination imbalance, and sensor noise. Such extensions would further enhance the generalization and usability of lightweight FIQA models in real-world, large-scale biometric, and multimedia applications.

7

Research Insights and Future Perspectives

7.1 | Overview and Summary of the Dissertation

This dissertation has presented a comprehensive research framework for learning-based QoE modeling across diverse interactive and immersive media contexts. The central motivation stems from the increasing complexity of multimedia ecosystems, which range from real-time communication and adaptive streaming to emerging volumetric and biometric modalities, where user-perceived quality results from the intricate interplay between system behavior, human perception, and contextual dynamics.

Traditional QoE estimation methods, often based on handcrafted metrics or static configurations, struggle to generalize across such heterogeneous environments. Hence, the overarching aim of this work has been to design data-driven, perceptually aligned, and computationally efficient QoE prediction models that bridge the gap between human perception and machine inference, while remaining deployable in real-world multimedia systems.

Building on the conceptual foundation introduced in Chapter 1, the dissertation develops a two-layer learning framework that jointly addresses both **service-level** and **content-level** QoE. These layers represent complementary dimensions of perceived quality:

- **Service-Level QoE** models user experience as a function of network and application behavior, addressing services such as WebRTC, streaming, and web browsing.

- **Content-Level QoE** assesses perceptual fidelity in visual or spatial media, including 3D point clouds and face images, where quality depends on intrinsic content attributes rather than transmission conditions.

Across both layers, the research is unified by five methodological pillars: (i) *signal design*, (ii) *temporal modeling*, (iii) *multi-view fusion*, (iv) *multi-projection fusion*, and (v) *computational efficiency*. Figure 7.1 illustrates how these pillars connect conceptually and operationally across the two layers of the proposed framework.

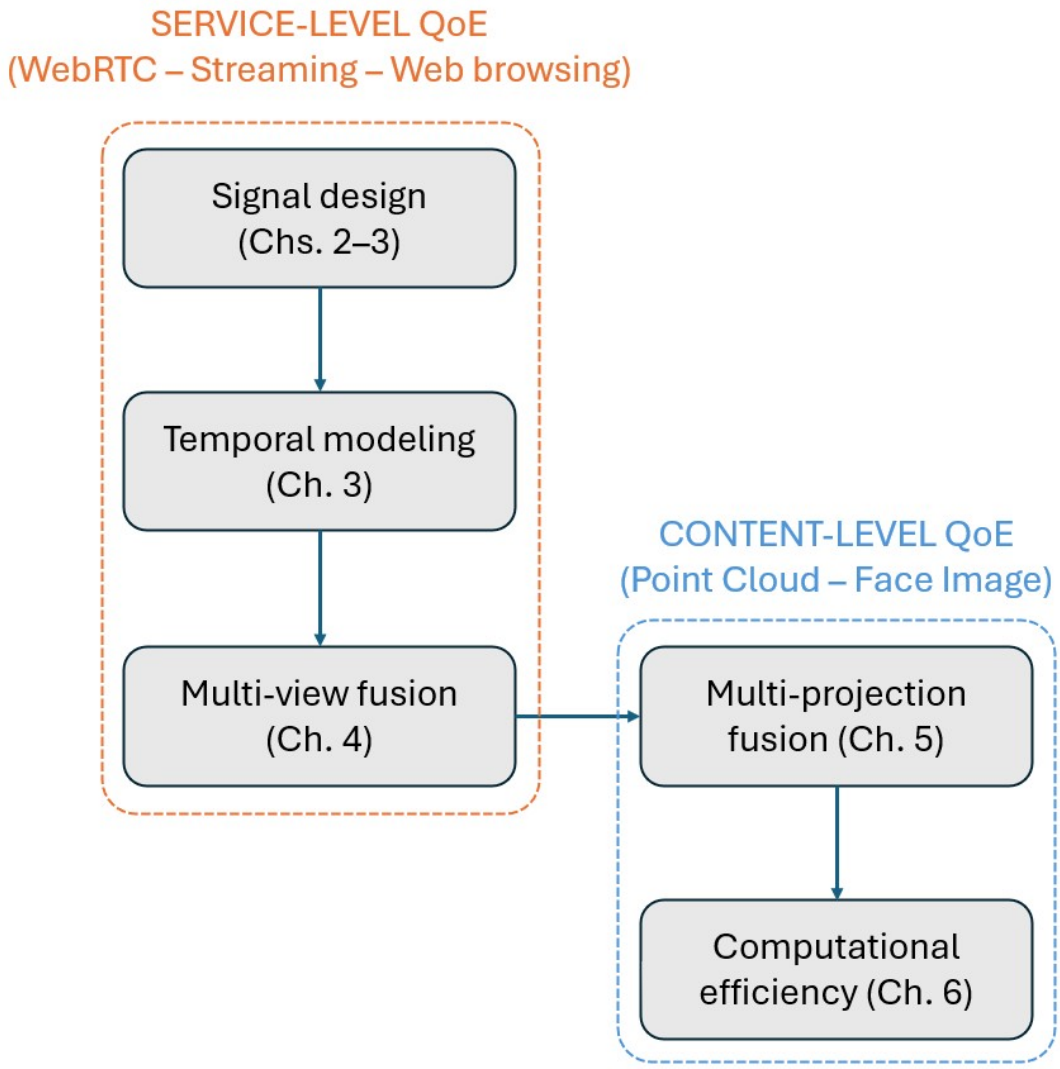

Figure 7.1: Unified two-layer conceptual framework for learning-based QoE modeling. The framework integrates **Service-Level QoE** (WebRTC, streaming, web browsing) and **Content-Level QoE** (point cloud, face image) within a unified methodological structure encompassing five key pillars: signal design, temporal modeling, multi-view fusion, multi-projection fusion, and computational efficiency. Each pillar corresponds to the contributions presented in Chapters 2–6.

The research follows a progressive methodological trajectory, with each chapter addressing a distinct yet interconnected component of the two-layer QoE framework:

- **Chapter 2 (WebRTC QoE)** established the foundation for *signal design* by introducing an application-layer QoE estimation framework for Web-based conversational systems. Leveraging telemetry from *webrtc-internals*, it demonstrated how interpretable machine learning regression could accurately predict subjective Mean Opinion Scores (MOS), forming the basis of data-driven QoE modeling from real-world communication signals.

- **Chapter 3 (Streaming QoE)** advanced toward *temporal modeling* using transformer architectures that capture sequential dependencies in adaptive video streaming. By encoding per-segment session dynamics and salient events such as start-up delays and stalls, this work highlighted the importance of temporal awareness in continuous QoE prediction.

- **Chapter 4 (Multi-View QoE)** expanded the framework toward *collaborative learning* by proposing a multi-view neural architecture that enables distributed entities (e.g., ISPs and OTTs) to jointly enhance QoE prediction without data sharing. This work introduced the notion of *model-level fusion*, ensuring privacy preservation while improving generalization across heterogeneous datasets.

- **Chapter 5 (NR-PCQA)** transitioned from the *service-level* to the *content-level* QoE layer. It introduced an NR point cloud quality assessment model that employs multi-projection fusion and adaptive view weighting to predict the perceptual quality of 3D content without access to references. This chapter validated the scalability of the proposed methodology to immersive, spatially complex media.

- **Chapter 6 (FIQA)** focused on *computational efficiency and deployment*, presenting a lightweight ensemble-based framework for face image quality assessment (FIQA). By combining MobileNet and ShuffleNet backbones with correlation-aware learning, it achieved state-of-the-art perceptual correlation under strict computational constraints, demonstrating real-world deployability.

Collectively, these chapters trace a coherent evolution from *signal-driven modeling* to *temporal, collaborative, modality-agnostic, and deployment-oriented* QoE frameworks. This cumulative progression supports the dissertation's central contribution: a unified, scalable, and perceptually consistent methodology for learning-based QoE estimation applicable across both service- and content-level domains.

7.1.1 | Main Scientific Contributions

The dissertation has made a series of interrelated scientific contributions that advance the field of QoE modeling across diverse multimedia modalities, from conversational communications to immersive 3D and biometric content. Each contribution has been developed as a self-contained study corresponding to one research chapter, while collectively forming a coherent methodological and conceptual framework.

1. WebRTC Conversational QoE from Application-Layer Telemetry (Chapter 2)

   The first major contribution is the development of an application-layer QoE estimation framework for real-time audiovisual communications. Unlike prior approaches relying on network-layer parameters, this work exploits telemetry data from webrtc-internals to model user-perceived conversational quality directly at the application level. Through rigorous correlation analysis, feature selection, and regression modeling, the framework offers a reproducible, interpretable, and deployable solution for predicting subjective MOS without intrusive instrumentation. This contribution represents a methodological foundation for real-world, data-driven QoE analysis in browser-based communication systems.

2. Temporal Modeling for Adaptive Video Streaming QoE (Chapter 3)

   The second contribution introduces a transformer-based deep learning model for adaptive video streaming, capable of learning long-term temporal dependencies between quality variations and playback events. The model encodes per-second sequences of bitrate, resolution, and stalling patterns into high-dimensional embeddings, enabling accurate and robust prediction of continuous QoE over time. This work contributes a novel sequential QoE modeling pipeline that surpasses traditional frame or segment-based methods and establishes a new direction toward temporal attention architectures in streaming QoE prediction.

3. Collaborative Multi-View Learning for QoE Prediction (Chapter 4)

   The third contribution extends QoE modeling to a collaborative, privacy-preserving paradigm using multi-view learning. This framework enables independent entities, such as ISPs and OTT providers, to train separate models on their own datasets while sharing latent-level representations rather than raw data. A fusion layer aggregates learned features to enhance prediction accuracy while maintaining confidentiality. This study introduces the concept of model-level collaboration in QoE prediction — a step toward cross-domain integration of quality intelligence without violating data ownership or privacy regulations.

4. No-Reference Point Cloud Quality Assessment (NR-PCQA) (Chapter 5)

The fourth contribution expands the scope of QoE research from conventional 2D video to immersive 3D media by proposing the Multi-View Adaptive Weighting (MVAW-PCQA) framework. This projection-based NR model processes six orthogonal views of a point cloud through a shared convolutional backbone and fuses them via adaptive weighting. The approach achieves state-of-the-art results on the SJTU-PCQA dataset while remaining computationally efficient. This contribution demonstrates the modality-generalizability and scalability of QoE modeling techniques to next-generation media such as volumetric video and point cloud streaming.

5. Efficiency and Practical Deployment (FIQA Case Study) (Chapter 6)

   The fifth contribution focuses on the efficiency and real-world deployability of QoE-related assessment systems. Through the FIQA case study, a compact two-branch ensemble was designed, combining MobileNet and ShuffleNet architectures with a correlation-aware loss function and a Test-Time Augmentation (TTA) strategy. The resulting model achieved state-of-the-art correlation with human perceptual ratings on the VQualA challenge FIQA dataset while operating under strict computational constraints. This final contribution highlights the feasibility of real-time, resource-efficient perceptual models, bridging the gap between theoretical QoE estimation and practical implementation in operational environments.

6. Two-Layer Framework for Objective QoE Modeling

   Beyond the individual studies, this dissertation establishes a unified, two-layer conceptual framework that connects service-level and content-level QoE modeling through five methodological pillars: signal design, temporal modeling, multi-view fusion, multi-projection fusion, and computational efficiency. At the service level, the framework integrates insights from WebRTC and adaptive streaming studies, showing how temporal and contextual dependencies can be captured through telemetry-based signal design and sequence-aware learning. At the content level, it extends these principles to perceptual and spatial domains—demonstrating that the same data-driven mechanisms governing communication quality can predict visual fidelity in point clouds and biometric imagery.

   The framework thus provides a coherent theoretical basis for QoE modeling that spans both system-driven and perception-driven dimensions of multimedia experience. It formalizes a progression from feature curation and temporal understanding to multi-modal integration and efficiency optimization, illustrating how learning-based models can evolve from descriptive analysis toward deployable, perceptually aligned intelligence. Collectively, these advancements contribute a generalizable methodology for learning-based QoE estimation, offering conceptual unity across heterogeneous media and practical guidance for designing scalable, interpretable, and resource-aware models in future interactive and immersive environments.

7.1.2 | Scientific and Practical Implications

The scientific and practical implications of this dissertation extend beyond the development of isolated QoE models, contributing instead to a systematic redefinition of how human perception can be quantified, learned, and operationalized across modern multimedia ecosystems. By integrating temporal modeling, collaborative learning, multimodal assessment, and computational efficiency within a framework, the research advances both the theoretical and applied frontiers of QoE modeling across service-level and content-level domains.

1. Advancing the Scientific Understanding of QoE

   From a scientific standpoint, the dissertation redefines QoE estimation as a data-driven perceptual inference problem, bridging psychometric principles with machine learning representations. Earlier QoE research often focused on either network metrics or static feature-based regression. The presented models, from transformer-based temporal prediction to adaptive fusion of multi-view projections, move beyond this paradigm by embedding perceptual dependencies directly into learned latent spaces. This enables the automatic discovery of cross-modal correlations between system parameters, content characteristics, and subjective experience. The findings collectively contribute to a generalized QoE modeling theory capable of spanning different service categories (communication, streaming, immersive, and biometric), demonstrating that perceptual quality can be learned consistently through data-driven signal design and hierarchical attention.

2. Methodological Innovation and Reproducibility

   The dissertation also provides a reproducible methodological pipeline for QoE estimation, encompassing data preprocessing, feature encoding, sequentialization, model design, and evaluation metrics. Each chapter builds upon this structured methodology, adapting it to distinct media types while preserving consistency in training, validation, and interpretability. This ensures that the resulting models are not only accurate but also scientifically transparent and transferable, enabling future researchers and practitioners to replicate and extend the work across new datasets or modalities. Furthermore, the integration of interpretable learning mechanisms (e.g., feature correlation analysis, adaptive weighting) enhances the understanding of which factors most influence perceived quality, thus contributing to explainable AI in QoE research.

3. Practical Relevance and Real-World Deployment

   From a practical perspective, the research provides actionable solutions for deployable, real-time QoE management in modern and immersive multimedia environments.

- The WebRTC and streaming QoE models can support online adaptation and troubleshooting in browser-based or cloud-based media services.
- The multi-view collaborative learning framework enables cross-organizational quality estimation while preserving user data privacy, a crucial requirement in multi-stakeholder ecosystems such as ISPs, OTT providers, and service regulators.
- The NR-PCQA model directly supports immersive media pipelines (e.g., volumetric video, XR rendering) by enabling perceptually guided rate adaptation without 3D reconstruction overhead.
- The FIQA case study validates the practical feasibility of high-performance perceptual modeling within constrained hardware conditions, paving the way for integration into edge devices and embedded systems.

Together, these developments contribute to the engineering foundation of future immersive and Human Digital Twin (HDT) systems, where users' emotional, perceptual, and experiential states can be continuously monitored, predicted, and optimized through scalable, multimodal QoE assessment models.

4. Societal and Industrial Impact

   Finally, the dissertation has direct implications for the design of user-centric digital ecosystems. Accurate, real-time QoE modeling promotes fairness and inclusivity by ensuring that adaptive systems optimize service quality based on perceptual outcomes rather than raw technical performance. Industries such as telecommunications, media streaming, autonomous systems, and extended reality can directly benefit from these models to enhance resource allocation, reduce latency, and deliver improved user experiences. Moreover, the methodological emphasis on efficiency and interpretability aligns with sustainable AI principles, reducing energy and computation costs while maintaining high perceptual accuracy.

7.1.3 | Limitations and Scope of the Proposed Framework

While this dissertation proposes a comprehensive and validated framework for learning-based QoE modeling, it is important to clearly articulate its scope and limitations.

First, the notion of QoE adopted throughout this work is primarily operationalized through subjective MOS and closely related perceptual ratings. Although MOS remains the dominant and standardized proxy for perceived quality in multimedia research, it represents an aggregated and context-dependent measure that cannot fully capture all dimensions of user experience, such as emotional response, long-term satisfaction, task success, or social context. Consequently, the proposed

models should be interpreted as predictors of perceived technical quality rather than complete representations of human experience in its broader psychological or sociological sense.

Second, the models developed in this dissertation focus predominantly on technical and content-related influence factors, including application-layer telemetry, network events, visual distortions, and geometric fidelity. Non-technical factors, such as user expectations, cultural background, mood, attention, social setting, or task importance, are largely outside the scope of the current modeling pipeline. While these factors are known to influence QoE, their reliable measurement and integration remain challenging and highly context-dependent. As a result, the presented framework prioritizes measurable, system-level signals that can be collected automatically and deployed at scale.

Third, the proposed learning-based approaches rely on data-driven generalization, which inherently depends on the representativeness of the available datasets. Although cross-dataset and cross-scenario evaluations are conducted where possible, performance may degrade under extreme distribution shifts, unseen content types, or novel interaction paradigms. The framework mitigates, but does not eliminate this risk through careful signal design, temporal modeling, and view-level fusion.

Fourth, ethical considerations arise when deploying QoE models in real-world systems, particularly in scenarios involving biometric data (e.g., FIQA) or large-scale user monitoring. While this dissertation emphasizes privacy-preserving learning strategies and avoids raw-data sharing, it does not explicitly address broader ethical questions such as informed consent, algorithmic bias, or downstream misuse of perceptual predictions. These aspects should be carefully considered when translating the proposed models into operational systems.

Finally, the framework is not intended as a monolithic or universal QoE solution, but rather as a methodological paradigm. Each chapter instantiates a subset of the framework tailored to a specific service or content modality. The generality of the approach lies in the design principles signal relevance, temporal awareness, collaborative learning, and efficiency rather than in a single unified model applicable to all scenarios.

7.1.4 | Future Work and Outlook

While this dissertation has proposed and validated a cohesive conceptual two-layer framework for data-driven QoE estimation across multiple service domains, several open challenges and research opportunities remain. Future research should also explore how subjective, contextual, and behavioral factors, such as user intent, affective state, and social context can be integrated into QoE models without compromising scalability or privacy. Moreover, efforts should aim to extend this foundation toward generalizable, adaptive, and human-centered perceptual intelligence systems, capable of bridging service-level and content-level analysis through continuous learning and cross-modal integration.

1. Toward Multimodal and Cross-Domain QoE Fusion

   A natural progression of this research lies in developing multimodal QoE fusion models that combine information from heterogeneous data sources, such as facial expressions, speech, body movements, and contextual features, to provide a holistic assessment of user experience. Although the dissertation demonstrated modality transferability through separate models for WebRTC, streaming, point cloud, and face-quality scenarios, future work could integrate these modalities under a unified perceptual framework. Such models could leverage cross-modal attention and transformer-based fusion strategies to capture how users' perceptual states emerge from the interplay between visual, auditory, and contextual stimuli. This step would also align with the broader goal of perceptually aware HDTs.

2. Continual and Federated Learning for QoE Adaptation

   Another promising direction is the adoption of continual and federated learning paradigms to enable QoE models that evolve over time and across distributed environments. Continual learning would allow QoE estimators to adapt to new content types, devices, or user groups without catastrophic forgetting, ensuring sustained performance in dynamic multimedia ecosystems. Meanwhile, federated learning aligns with the collaborative spirit introduced in Chapter 4, allowing multiple service providers to train shared QoE models without exchanging raw data, preserving both scalability and privacy. Integrating these paradigms would create self-improving QoE systems capable of adapting to evolving service conditions.

3. Human-in-the-Loop and Explainable QoE Models

   To enhance trust and interpretability, future work should explore human-in-the-loop QoE modeling, where subjective feedback or behavioral data continuously calibrates the objective estimators. Explainable AI techniques could be used to visualize which perceptual features drive model predictions, clarifying how system-level variations (e.g., bitrate, delay, or viewpoint) influence perceived quality. Such explainable and interactive QoE models would not only improve scientific transparency but also facilitate ethical and fair decision-making in applications where user experience has social or safety implications.

4. Real-Time Deployment and HDT Integration

   Finally, an important frontier concerns the deployment of QoE models in real-time HDT ecosystems. The proposed frameworks, particularly the lightweight designs in Chapters 5 and 6, provide a solid foundation for integrating perceptual intelligence into digital twins that continuously monitor users' affective and perceptual states. Future implementations may involve coupling QoE estimation modules with edge computing architectures or immersive XR platforms, allowing adaptive service management and perceptual optimization in real time. Such

integration would represent a step toward autonomous, perceptually aware digital twins, capable of anticipating user needs and optimizing service delivery dynamically.

In summary, the future of QoE research lies in bridging perception, intelligence, and deployment, moving from isolated prediction models toward adaptive, multimodal, and human-centered ecosystems. By combining interpretability, scalability, and efficiency, the next generation of QoE frameworks will not only measure experience but also enable intelligent systems that learn from, respond to, and ultimately enhance the human experience itself.

7.2 | Closing Remarks

This dissertation has presented a comprehensive exploration of learning-based QoE estimation across diverse service categories and perceptual domains. Following a progressive research trajectory, from conversational WebRTC analysis and temporal streaming models to collaborative multi-view learning, immersive point cloud assessment, and efficient face image quality prediction, the work has demonstrated how QoE can be systematically learned, interpreted, and operationalized within intelligent multimedia systems.

Each study contributed a distinct methodological advancement corresponding to one of the dissertation's five pillars: signal design, temporal modeling, multi-view fusion, multi-projection fusion, and computational efficiency. Together, these studies formed a coherent two-layer framework that bridges service-level QoE, governed by system and network behavior, with content-level QoE, defined by perceptual and spatial fidelity. The resulting framework not only unifies these complementary dimensions but also establishes the methodological and computational foundations for scalable, privacy-preserving, and deployable QoE models.

By bridging human perception with computational intelligence, the research lays the groundwork for perceptually aware digital ecosystems and HDT systems that can understand, predict, and optimize user experience in real-time. Ultimately, this dissertation reinforces the view that QoE is not merely a numerical indicator but a connective layer between cognition and computation—one that, when modeled responsibly and efficiently, can drive the next generation of adaptive, user-centered, and intelligent multimedia services.

References

[1] Haopeng Wang, Haiwei Dong, and Abdulmotaleb El Saddik, "Immersive multimedia communication: State-of-the-art on extended reality streaming," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 21, no. 7, July 2025.

[2] Xue Yao, Simeon C. Calvert, and Serge P. Hoogendoorn, "Driving heterogeneity identification using machine learning: A review and framework for analysis," *Transportation Research Interdisciplinary Perspectives*, vol. 32, pp. 101511, 2025.

[3] Alejandro S Martínez Sala, Lucio Hernando-Cánovas, Juan-Carlos Sanchez-Aarnoutse, and Juan J. Alcaraz, "Resource-efficient fog computing vision system for occupancy monitoring: A real-world deployment in university libraries," *Internet of Things*, vol. 34, pp. 101748, nov 2025.

[4] MohammadAli Hamidi, Gülnaziye Bingöl, Alessandro Floris, Simone Porcu, and Luigi Atzori, "Analysis of Application-layer Data to Estimate the QoE of WebRTC-based Audiovisual Conversations," in *2023 IEEE Globecom Workshops (GC Wkshps)*. IEEE, 2023, pp. 365–370.

[5] Mohammad Ali Hamidi, Simone Porcu, Alessandro Floris, and Luigi Atzori, "A Transformer-Based Modelling Approach for Robust QoE Estimation in Video Streaming," in *2025 23rd Mediterranean Communication and Computer Networking Conference (MedComNet)*. IEEE, 2025, pp. 1–6.

[6] MohammadAli Hamidi, Simone Porcu, Alessandro Floris, and Luigi Atzori, "Towards the Application of Multi-view Learning in Quality of Experience Collaborative Modelling," in *2024 16th International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 2024, pp. 286–292.

[7] MohammadAli Hamidi, Simone Porcu, Alessandro Floris, and Luigi Atzori, "MVAW-PCQA: A No-reference Point Cloud Quality Assessment via Multi-View Adaptive Weighting," in *2025 17th International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 2025.

[8] MohammadAli Hamidi, Hadi Amirpour, Luigi Atzori, and Christian Timmerer, "A lightweight ensemble-based face image quality assessment method with correlation-aware loss," *arXiv preprint arXiv:2509.10114*, 2025.

[9] Boni Garcia, Francisco Gortazar, Luis Lopez-Fernandez, Micael Gallego, and Miguel Paris, "WebRTC Testing: Challenges and Practical Solutions," *IEEE Communications Standards Magazine*, vol. 1, no. 2, pp. 36–42, 2017.

[10] IT Union, "Methods for the subjective assessment of video quality audio quality and audiovisual quality of internet video and distribution quality television in any environment," *SERIES P: Terminals And Subjective And Objective Assessment Methods*, 2016.

[11] ITU-T Subjective Video Quality Assessment, "Methods for multimedia applications," *Telecommuniation Standardization Sector of ITU Recommendation P*, vol. 911, 1998.

[12] Jasmina Baraković Husić, Adna Alić, Sabina Baraković, and Mladen Mrkaja, "QoE Prediction of WebRTC Video Calls Using Google Chrome Statistics," in *2021 20th Int. Symposium INFOTEH-JAHORINA (INFOTEH)*, 2021, pp. 1–6.

[13] Doreid Ammar, Katrien De Moor, Lea Skorin-Kapov, Markus Fiedler, and Poul E. Heegaard, "Exploring the Usefulness of Machine Learning in the Context of WebRTC Performance Estimation," in *2019 IEEE 44th Conference on Local Computer Networks (LCN)*, 2019, pp. 406–413.

[14] Elena Cipressi and Maria Luisa Merani, "An effective machine learning (ml) approach to quality assessment of voice over ip (voip) calls," *IEEE Networking Letters*, vol. 2, no. 2, pp. 90–94, 2020.

[15] Markus Fiedler, Tobias Hossfeld, and Phuoc Tran-Gia, "A generic quantitative relationship between quality of experience and quality of service," *Ieee Network*, vol. 24, no. 2, pp. 36–41, 2010.

[16] Maria Torres Vega, Cristian Perra, and Antonio Liotta, "Resilience of video streaming services to network impairments," *IEEE Transactions on Broadcasting*, vol. 64, no. 2, pp. 220–234, 2018.

[17] M.-N. Garcia, P. List, S. Argyropoulos, D. Lindegren, M. Pettersson, B. Feiten, J. Gustafsson, and A. Raake, "Parametric model for audiovisual quality assessment in iptv: Itu-t rec. p.1201.2," in *2013 IEEE 15th International Workshop on Multimedia Signal Processing (MMSP)*, 2013, pp. 482–487.

[18] Junchen Jiang, Vyas Sekar, and Hui Zhang, "Improving fairness, efficiency, and stability in http-based adaptive video streaming with festive," in *Proceedings of the 8th international conference on Emerging networking experiments and technologies*, 2012, pp. 97–108.

[19] Xiaoqi Yin, Abhishek Jindal, Vyas Sekar, and Bruno Sinopoli, "A control-theoretic approach for dynamic adaptive video streaming over http," in *Proceedings of the 2015 ACM conference on special interest group on data communication*, 2015, pp. 325–338.

[20] Michael Seufert, Sebastian Egger, Martin Slanina, Thomas Zinner, Tobias Hoßfeld, and Phuoc Tran-Gia, "A survey on quality of experience of http adaptive streaming," *IEEE Communications Surveys Tutorials*, vol. 17, no. 1, pp. 469–492, 2015.

[21] Orlewilson Bentes Maia, Hani Camille Yehia, and Luciano de Errico, "A concise review of the quality of experience assessment for video streaming," *Computer communications*, vol. 57, pp. 1–12, 2015.

[22] Doreid Ammar, Katrien De Moor, Min Xie, Markus Fiedler, and Poul Heegaard, "Video qoe killer and performance statistics in webrtc-based video communication," in *2016 IEEE Sixth International Conference on Communications and Electronics (ICCE)*. IEEE, 2016, pp. 429–436.

[23] Nikita Smirnov and Sven Tomforde, "Real-time rate control of webrtc video streams in 5g networks: Improving quality of experience with deep reinforcement learning," *Journal of Systems Architecture*, vol. 148, pp. 103066, 2024.

[24] Jasmina Baraković Husić, Adna Alić, Sabina Baraković, and Mladen Mrkaja, "Qoe prediction of webrtc video calls using google chrome statistics," in *2021 20th International Symposium INFOTEH-JAHORINA (INFOTEH)*. IEEE, 2021, pp. 1–6.

[25] Gaetano Carlucci, Luca De Cicco, Stefan Holmer, and Saverio Mascolo, "Analysis and Design of the Google Congestion Control for Web Real-Time Communication (WebRTC)," in *Proceedings of the 7th International Conference on Multimedia Systems*, 2016, MMSys '16.

[26] Katrien De Moor, Sebastian Arndt, Doreid Ammar, Jan-Niklas Voigt-Antons, Andrew Perkis, and Poul E. Heegaard, "Exploring diverse measures for evaluating QoE in the context of WebRTC," in *2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX)*, 2017, pp. 1–3.

[27] Christos Tsiaras, Manuel Rösch, and Burkhard Stiller, "VoIP-based calibration of the DQX model," in *2015 IFIP Netw. Conf. (IFIP Networking)*, 2015, pp. 1–9.

[28] ITU, "Methods for subjective determination of transmission quality," Recommendation ITU-T P.800, 1996.

[29] Andrew Perkis, Christian Timmerer, and et al., "QUALINET White Paper on Definitions of Immersive Media Experience (IMEx)," in *European Network on Quality of Experience in Multimedia Systems and Services, 14th QUALINET meeting (online)*, 2020.

[30] Sara Vlahovic, Mirko Suznjevic, and Lea Skorin-Kapov, "A survey of challenges and methods for Quality of Experience assessment of interactive VR applications," *Journal on Multimodal User Interfaces*, pp. 1–35, 04 2022.

[31] Christos G Bampis, Zhi Li, and Alan C Bovik, "Continuous prediction of streaming video QoE using dynamic networks," *IEEE Signal Processing Letters*, vol. 24, no. 7, pp. 1083–1087, 2017.

[32] Monalisa Ghosh, Dr Chetna Singhal, and Rushikesh Wayal, "DeSVQ: Deep learning based streaming video QoE estimation," in *Proc. of the 23rd Int. Conf. on Distributed Computing and Networking*, 2022, pp. 19–25.

[33] Tasnim Abar, Asma Ben Letaifa, and Sadok El Asmi, "Chapter Five - User behavior-ensemble learning based improving QoE fairness in HTTP adaptive streaming over SDN approach," vol. 123 of *Advances in Computers*, pp. 245–269. Elsevier, 2021.

[34] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[35] Ms D Deepa et al., "Bidirectional encoder representations from transformers (BERT) language model for sentiment analysis task," *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, vol. 12, no. 7, pp. 1708–1721, 2021.

[36] Kazuki Irie, Albert Zeyer, Ralf Schlüter, and Hermann Ney, "Language Modeling with Deep Transformers," in *Proc. Interspeech 2019*, 2019, pp. 3905–3909.

[37] ITU, "Parametric bitstream-based quality assessment of progressive download and adaptive audiovisual streaming services over reliable transport," Recommendation ITU-T P.1203, 2017.

[38] Dimitris Tsolkas, Eirini Liotou, Nikos Passas, and Lazaros Merakos, "A survey on parametric QoE estimation for popular services," *Journal of Network and Computer Applications*, vol. 77, no. 1, pp. 1–17, 2017.

[39] Nabajeet Barman and Maria G. Martini, "QoE Modeling for HTTP Adaptive Video Streaming–A Survey and Open Challenges," *IEEE Access*, vol. 7, pp. 30831–30859, 2019.

[40] Georgios Kougioumtzidis, Vladimir Poulkov, Zaharias D. Zaharis, and Pavlos I. Lazaridis, "A Survey on Multimedia Services QoE Assessment and Machine Learning-Based Prediction," *IEEE Access*, vol. 10, pp. 19507–19538, 2022.

[41] Michael Seufert and Irena Orsolic, "Improving the Transfer of Machine Learning-Based Video QoE Estimation Across Diverse Networks," *IEEE Transactions on Network and Service Management*, pp. 1–1, 2023.

[42] Huaizheng Zhang, Linsen Dong, Guanyu Gao, Han Hu, Yonggang Wen, and Kyle Guan, "DeepQoE: A Multimodal Learning Framework for Video Quality of Experience (QoE) Prediction," *IEEE Transactions on Multimedia*, vol. 22, no. 12, pp. 3210–3223, 2020.

[43] Rushi Babaria, Sharat Chandra Madanapalli, Himal Kumar, and Vijay Sivaraman, "FlowFormers: Transformer-based Models for Real-time Network Flow Classification," in *2021 17th International Conference on Mobility, Sensing and Networking (MSN)*, 2021, pp. 231–238.

[44] Wei Shen, Xiaonan He, Chuheng Zhang, Xuyun Zhang, and Jian Xie, "A transformer-based user satisfaction prediction for proactive interaction mechanism in DuerOS," in *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 2022, pp. 1777–1786.

[45] Zutong Li and Lei Yang, "DCVQE: A Hierarchical Transformer for Video Quality Assessment," in *Proceedings of the Asian Conference on Computer Vision*, 2022, pp. 2562–2579.

[46] Alexander Raake, Marie-Neige Garcia, Werner Robitza, Peter List, Steve Göring, and Bernhard Feiten, "A bitstream-based, scalable video-quality model for HTTP adaptive streaming: ITU-T P.1203.1," in *2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX)*, 2017, pp. 1–6.

[47] Werner Robitza, Steve Göring, Alexander Raake, David Lindegren, Gunnar Heikkilä, Jörgen Gustafsson, Peter List, Bernhard Feiten, Ulf Wüstenhagen, Marie-Neige Garcia, Kazuhisa Yamagishi, and Simon Broom, "HTTP Adaptive Streaming QoE Estimation with ITU-T Rec. P.1203 – Open Databases and Software," in *9th ACM Multimedia Systems Conference*, 2018.

[48] Paula Branco, Luís Torgo, and Rita P. Ribeiro, "SMOGN: a pre-processing approach for imbalanced regression," in *Proc. of the First Int. Workshop on Learning with Imbalanced Domains: Theory and Applications*, Paula Branco Luís Torgo and Nuno Moniz, Eds. 22 Sep 2017, vol. 74 of *Proc.s of Machine Learning Research*, pp. 36–50, PMLR.

[49] Luís Torgo, Paula Branco, Rita P. Ribeiro, and Bernhard Pfahringer, "Resampling strategies for regression," *Expert Systems*, vol. 32, pp. 465 – 476, 2015.

[50] Fenglin Liu, Xuancheng Ren, Zhiyuan Zhang, Xu Sun, and Yuexian Zou, "Rethinking Skip Connection with Layer Normalization," in *International Conference on Computational Linguistics*, 2020.

[51] James Bergstra and Yoshua Bengio, "Random search for hyper-parameter optimization.," *Journal of machine learning research*, vol. 13, no. 2, 2012.

[52] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[53] Xiaoqiang Yan, Shizhe Hu, Yiqiao Mao, Yangdong Ye, and Hui Yu, "Deep multi-view learning methods: A review," *Neurocomputing*, vol. 448, pp. 106–129, 2021.

[54] D.N. da Hora, A.S. Asrese, V. Christophides, R. Teixeira, and D. Rossi, "Narrowing the gap between QoS metrics and Web QoE using Above-the-fold metrics," in *Passive and Active Measurement (PAM)*, R. Beverly, G. Smaragdakis, and A. Feldmann, Eds. 2018, pp. 31–43, Springer International Publishing.

[55] Jun Yu, Jing Li, Zhou Yu, and Qingming Huang, "Multimodal Transformer With Multi-View Visual Representation for Image Captioning," *IEEE Trans on Circ. and Syst. for Video Tech.*, vol. 30, no. 12, pp. 4467–4480, 2020.

[56] Soumyadeep Bhattacharjee and Wenyao Xu, "VoiceLens: A multi-view multi-class disease classification model through daily-life speech data," *Smart Health*, vol. 23, pp. 100233, 2022.

[57] Jingying Chen, Lei Yang, Lei Tan, and Ruyi Xu, "Orthogonal channel attention-based multi-task learning for multi-view facial expression recognition," *Pattern Recognition*, vol. 129, pp. 108753, 2022.

[58] Yalan Qin, Chuan Qin, Xinpeng Zhang, Donglian Qi, and Guorui Feng, "NIM-Nets: Noise-Aware Incomplete Multi-View Learning Networks," *IEEE Transactions on Image Processing*, vol. 32, pp. 175–189, 2023.

[59] Tara N. Sainath, Oriol Vinyals, Andrew Senior, and Haşim Sak, "Convolutional, Long Short-Term Memory, fully connected Deep Neural Networks," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 4580–4584.

[60] Mohammad Hosseini and Christian Timmerer, "Dynamic Adaptive Point Cloud Streaming," in *Proceedings of the 23rd Packet Video Workshop*, New York, NY, USA, 2018, PV '18, p. 25–30, Association for Computing Machinery.

[61] Simone Porcu, Claudio Marche, and Alessandro Floris, "No-Reference Objective Quality Metrics for 3D Point Clouds: A Review," *Sensors*, vol. 24, no. 22, 2024.

[62] O Nakagami and PCCWD G-PCC, "PCC WD G-PCC (Geometry-Based PCC)," *ISO/IEC JTC1/SC29/WG11 MPEG, Standard*, , no. 17771, 2018.

[63] Yipeng Liu, Qi Yang, Yiling Xu, and Le Yang, "Point Cloud Quality Assessment: Dataset Construction and Learning-based No-reference Metric," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 19, no. 2s, Feb. 2023.

[64] Ziyu Shan, Qi Yang, Rui Ye, Yujie Zhang, Yiling Xu, Xiaozhong Xu, and Shan Liu, "GPA-Net:No-Reference Point Cloud Quality Assessment With Multi-Task Graph Convolutional Network," *IEEE Transactions on Visualization and Computer Graphics*, vol. 30, no. 8, pp. 4955–4967, 2024.

[65] Zicheng Zhang, Wei Sun, Xiongkuo Min, Quan Zhou, Jun He, Qiyuan Wang, and Guangtao Zhai, "MM-PCQA: Multi-Modal Learning for No-reference Point Cloud Quality Assessment," in *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI-23)*, 2023.

[66] Qi Yang, Zhan Ma, Yiling Xu, Zhu Li, and Jun Sun, "Inferring point cloud quality via graph similarity," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 6, pp. 3015–3029, 2022.

[67] Evangelos Alexiou and Touradj Ebrahimi, "Towards a point cloud structural similarity metric," in *2020 IEEE International Conference on Multimedia Expo Workshops (ICMEW)*, 2020, pp. 1–6.

[68] Gabriel Meynet, Yana Nehmé, Julie Digne, and Guillaume Lavoué, "Pcqm: A full-reference quality metric for colored 3d point clouds," in *2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX)*, 2020, pp. 1–6.

[69] Irene Viola and Pablo Cesar, "A reduced reference metric for visual quality evaluation of point cloud contents," *IEEE Signal Processing Letters*, vol. 27, pp. 1660–1664, 2020.

[70] Sebastian Schwarz, Marius Preda, Vittorio Baroncini, Madhukar Budagavi, Pablo Cesar, Philip A. Chou, Robert A. Cohen, Maja Krivokuća, Sébastien Lasserre, Zhu Li, Joan Llach, Khaled Mammou, Rufael Mekuria,

Ohji Nakagami, Ernestasia Siahaan, Ali Tabatabai, Alexis M. Tourapis, and Vladyslav Zakharchenko, "Emerging MPEG Standards for Point Cloud Compression," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 9, no. 1, pp. 133–148, 2019.

[71] Qi Yang, Yipeng Liu, Siheng Chen, Yiling Xu, and Jun Sun, "No-reference point cloud quality assessment via domain adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 21179–21188.

[72] Qi Liu, Hui Yuan, Honglei Su, Hao Liu, Yu Wang, Huan Yang, and Junhui Hou, "PQA-Net: Deep no reference point cloud quality assessment via multi-view projection," *IEEE transactions on circuits and systems for video technology*, vol. 31, no. 12, pp. 4645–4660, 2021.

[73] Zicheng Zhang, Wei Sun, Yucheng Zhu, Xiongkuo Min, Wei Wu, Ying Chen, and Guangtao Zhai, "Evaluating Point Cloud from Moving Camera Videos: A No-Reference Metric," *IEEE Transactions on Multimedia*, pp. 1–13, 2023.

[74] Xiongli Chai and Feng Shao, "MS-PCQE: Efficient No-Reference Point Cloud Quality Evaluation via Multi-Scale Interaction Module in Immersive Communications," *IEEE Transactions on Consumer Electronics*, pp. 1–1, 2024.

[75] Sam Van Damme, Maria Torres Vega, Jeroen van der Hooft, and Filip De Turck, "Clustering-based Psychometric No-Reference Quality Model for Point Cloud Video," in *Proceedings - International Conference on Image Processing, ICIP*. 2022, pp. 1866–1870, IEEE Computer Society.

[76] Jannis Weil, Yassin Alkhalili, Anam Tahir, Thomas Gruczyk, Tobias Meuser, Mu Mu, Heinz Koeppl, and Andreas Mauthe, "Modeling Quality of Experience for Compressed Point Cloud Sequences based on a Subjective Study," in *2023 15th International Conference on Quality of Multimedia Experience (QoMEX)*, 2023, pp. 135–140.

[77] Minh Nguyen, Shivi Vats, and Hermann Hellwagner, "No-Reference Quality of Experience Model for Dynamic Point Clouds in Augmented Reality," in *MHV*. 2 2024, pp. 90–91, Association for Computing Machinery (ACM).

[78] Qi Liu, Honglei Su, Tianxin Chen, Hui Yuan, and Raouf Hamzaoui, "No-Reference Bitstream-Layer Model for Perceptual Quality Assessment of V-PCC Encoded Point Clouds," *IEEE Transactions on Multimedia*, vol. 25, pp. 4533–4546, 2023.

[79] Zicheng Zhang, Wei Sun, Xiongkuo Min, Wenhan Zhu, Tao Wang, Wei Lu, and Guangtao Zhai, "A No-Reference Evaluation Metric for Low-Light Image Enhancement," in *2021 IEEE International Conference on Multimedia and Expo (ICME)*, 2021, pp. 1–6.

[80] Aladine Chetouani, Maurice Quach, Giuseppe Valenzise, and Frédéric Dufaux, "Deep Learning-Based Quality Assessment Of 3d Point Clouds Without Reference," in *2021 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, 2021, pp. 1–6.

[81] Xiongli Chai, Feng Shao, Baoyang Mu, Hangwei Chen, Qiuping Jiang, and Yo-Sung Ho, "Plain-PCQA: No-Reference Point Cloud Quality Assessment by Analysis of Plain Visual and Geometrical Components," *IEEE*

Transactions on Circuits and Systems for Video Technology, vol. 34, no. 7, pp. 6207–6223, 2024.

[82] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie, "A convnet for the 2020s," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 11976–11986.

[83] S Alireza Golestaneh and Kris Kitani, "No-reference image quality assessment via feature fusion and multi-task learning," *arXiv preprint arXiv:2006.03783*, 2020.

[84] Wen-xu Tao, Gang-yi Jiang, Zhi-di Jiang, and Mei Yu, "Point cloud projection and multi-scale feature fusion network based blind quality assessment for colored point clouds," in *Proceedings of the 29th ACM international conference on multimedia*, 2021, pp. 5266–5272.

[85] Sebastian Bosse, Dominique Maniry, Klaus-Robert Müller, Thomas Wiegand, and Wojciech Samek, "Deep neural networks for no-reference and full-reference image quality assessment," *IEEE Transactions on image processing*, vol. 27, no. 1, pp. 206–219, 2017.

[86] Qi Yang, Hao Chen, Zhan Ma, Yiling Xu, Rongjun Tang, and Jun Sun, "Predicting the Perceptual Quality of Point Cloud: A 3D-to-2D Projection-Based Exploration," *IEEE Transactions on Multimedia*, vol. 23, pp. 3877–3891, 2021.

[87] Int. Telecommun. Union, "Methodology for the subjective assessment of the quality of television pictures itu-r recommendation bt.500-11, tech. rep.," 2000.

[88] Yu Fan, Zicheng Zhang, Wei Sun, Xiongkuo Min, Ning Liu, Quan Zhou, Jun He, Qiyuan Wang, and Guangtao Zhai, "A no-reference quality assessment metric for point cloud based on captured video sequences," in *2022 IEEE 24th international workshop on Multimedia signal processing (MMSP)*. IEEE, 2022, pp. 1–5.

[89] Rufael Mekuria, Zhu Li, Christian Tulvan, and Phil Chou, "Evaluation criteria for pcc (point cloud compression)," *ISO/IEC JTC*, vol. 1, pp. N16332, 2016.

[90] Rufael Mekuria, Kees Blom, and Pablo Cesar, "Design, implementation, and evaluation of a point cloud codec for tele-immersive video," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 4, pp. 828–842, 2017.

[91] Dong Tian, Hideaki Ochimizu, Chen Feng, Robert Cohen, and Anthony Vetro, "Geometric distortion metrics for point cloud compression," in *2017 IEEE International Conference on Image Processing (ICIP)*, 2017, pp. 3460–3464.

[92] Eric M. Torlig, Evangelos Alexiou, Tiago A. Fonseca, Ricardo L. de Queiroz, and Touradj Ebrahimi, "A novel methodology for quality assessment of voxelized point clouds," in *Optical Engineering + Applications*, 2018.

[93] Z. Wang, E.P. Simoncelli, and A.C. Bovik, "Multiscale structural similarity for image quality assessment," in *The Thrity-Seventh Asilomar Conference on Signals, Systems Computers, 2003*, 2003, vol. 2, pp. 1398–1402 Vol.2.

[94] Prajit Ramachandran, Niki Parmar, Ashish Vaswani, Irwan Bello, Anselm Levskaya, and Jon Shlens, "Stand-alone self-attention in vision models," *Advances in neural information processing systems*, vol. 32, 2019.

[95] Shaibal Saha and Lanyu Xu, "Vision transformers on the edge: A comprehensive survey of model compression and acceleration strategies," *arXiv preprint arXiv:2503.02891*, 2025.

[96] Qi Liu, Honglei Su, Zhengfang Duanmu, Wentao Liu, and Zhou Wang, "Perceptual quality assessment of colored 3d point clouds," *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–1, 2022.

[97] Qi Liu, Hui Yuan, Raouf Hamzaoui, Honglei Su, Junhui Hou, and Huan Yang, "Reduced reference perceptual quality model with application to rate control for video-based point cloud compression," *IEEE Transactions on Image Processing*, 2021.

[98] Xinju Wu, Yun Zhang, Chunling Fan, Junhui Hou, and Sam Kwong, "Subjective Quality Database and Objective Study of Compressed Point Clouds With 6DoF Head-Mounted Display," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 12, pp. 4630–4644, 2021.

[99] Xiao Yang, Wenhan Luo, Linchao Bao, Yuan Gao, Dihong Gong, Shibao Zheng, Zhifeng Li, and Wei Liu, "Face Anti-Spoofing: Model Matters, so Does Data," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, June 2019, pp. 3502–3511, IEEE.

[100] Sungeun Hong and Jongbin Ryu, "Unsupervised Face Domain Transfer for Low-Resolution Face Recognition," *IEEE Signal Processing Letters*, vol. 27, pp. 156–160, 2020.

[101] Woobin Im, Sungeun Hong, Sung-Eui Yoon, and Hyun S. Yang, "Scale-Varying Triplet Ranking with Classification Loss for Facial Age Estimation," in *Computer Vision – ACCV 2018*, C.V. Jawahar, Hongdong Li, Greg Mori, and Konrad Schindler, Eds., vol. 11365, pp. 247–259. Springer International Publishing, Cham, 2019, Series Title: Lecture Notes in Computer Science.

[102] Torsten Schlett, Christian Rathgeb, Olaf Henniger, Javier Galbally, Julian Fierrez, and Christoph Busch, "Face Image Quality Assessment: A Literature Survey," *ACM Computing Surveys*, vol. 54, no. 10s, pp. 1–49, Jan. 2022.

[103] Na Zhang, "A Study on the Impact of Face Image Quality on Face Recognition in the Wild," July 2023, arXiv:2307.02679 [cs].

[104] Maria Cuellar, Hon Kiu, To, and Arush Mehrotra, "Accuracy and Fairness of Facial Recognition Technology in Low-Quality Police Images: An Experiment With Synthetic Faces," May 2025, arXiv:2505.14320 [cs].

[105] Puspita Majumdar, Akshay Agarwal, Mayank Vatsa, and Richa Singh, "Facial Retouching and Alteration Detection," in *Handbook of Digital Face Manipulation and Detection*, Christian Rathgeb, Ruben Tolosana, Ruben Vera-Rodriguez, and Christoph Busch, Eds., pp. 367–387. Springer International Publishing, Cham, 2022, Series Title: Advances in Computer Vision and Pattern Recognition.

[106] Biying Fu, Cong Chen, Olaf Henniger, and Naser Damer, "A Deep Insight into Measuring Face Image Utility with General and Face-specific Image Quality Metrics," in *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, Waikoloa, HI, USA, Jan. 2022, pp. 1121–1130, IEEE.

[107] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-Reference Image Quality Assessment in the Spatial Domain," *IEEE Transactions on Image Processing*, vol. 21, no. 12, pp. 4695–4708, Dec. 2012.

[108] Lacey Best-Rowden and Anil K. Jain, "Learning Face Image Quality From Human Assessments," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 12, pp. 3064–3077, Dec. 2018.

[109] Philipp Terhorst, Jan Niklas Kolf, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper, "Face Quality Estimation and Its Correlation to Demographic and Non-Demographic Bias in Face Recognition," in *2020 IEEE International Joint Conference on Biometrics (IJCB)*, Houston, TX, USA, Sept. 2020, pp. 1–11, IEEE.

[110] Anish Mittal, Anush K. Moorthy, and Alan C. Bovik, "Blind/Referenceless Image Spatial Quality Evaluator," in *2011 Conference Record of the Forty Fifth Asilomar Conference on Signals, Systems and Computers (ASILOMAR)*, Pacific Grove, CA, USA, Nov. 2011, pp. 723–727, IEEE.

[111] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a "Completely Blind" Image Quality Analyzer," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 209–212, Mar. 2013.

[112] Venkatanath N, Praneeth D, Maruthi Chandrasekhar Bh, Sumohana S. Channappayya, and Swarup S. Medasani, "Blind Image Quality Evaluation using Perception based Features," in *2015 Twenty First National Conference on Communications (NCC)*, Mumbai, India, Feb. 2015, pp. 1–6, IEEE.

[113] Xuekai Wei, Mingliang Zhou, Heqiang Wang, Haoyan Yang, Lei Chen, and Sam Kwong, "Recent Advances in Rate Control: From Optimization to Implementation and Beyond," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 1, pp. 17–33, Jan. 2024.

[114] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli, "Image Quality Assessment: From Error Visibility to Structural Similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, Apr. 2004, Conference Name: IEEE Transactions on Image Processing.

[115] H.R. Sheikh and A.C. Bovik, "Image Information and Visual Quality," *IEEE Transactions on Image Processing*, vol. 15, no. 2, pp. 430–444, Feb. 2006.

[116] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang, "The Unreasonable Effectiveness of Deep Features as a Perceptual Metric," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2018, pp. 586–595, ISSN: 2575-7075.

[117] Hossein Talebi and Peyman Milanfar, "NIMA: Neural Image Assessment," *IEEE Transactions on Image Processing*, vol. 27, no. 8, pp. 3998–4011, Aug. 2018.

[118] Fadi Boutros, Meiling Fang, Marcel Klemt, Biying Fu, and Naser Damer, "CR-FIQA: Face Image Quality Assessment by Learning Sample Relative Classifiability," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Vancouver, BC, Canada, June 2023, pp. 5836–5845, IEEE.

[119] Fu-Zhao Ou, Chongyi Li, Shiqi Wang, and Sam Kwong, "CLIB-FIQA: Face Image Quality Assessment with Confidence Calibration," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 1694–1704, ISSN: 2575-7075.

[120] Byungho Jo, Donghyeon Cho, In Kyu Park, and Sungeun Hong, "IFQA: Interpretable FACE Quality Assessment," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2023, pp. 3444–3453.

[121] Chaofeng Chen, Jiadi Mo, Jingwen Hou, Haoning Wu, Liang Liao, Wenxiu Sun, Qiong Yan, and Weisi Lin, "TOPIQ: A Top-Down Approach from Semantics to Distortions for Image Quality Assessment," *IEEE Transactions on Image Processing*, 2024.

[122] Jing Ma, "Improved Self-Training for Test-Time Adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 23701–23710.

[123] A Deep Bilinear Convolutional Neural Network, "Blind Image Quality Assessment Using A Deep Bilinear Convolutional Neural Network," *Deep Bilinear Convolutional Neural*, 2022.

[124] Lorenzo Agnolucci Leonardo Galteri Marco Bertini, "Quality-Aware Image-Text Alignment for Opinion-Unaware Image Quality Assessment," *arXiv preprint arXiv:2403.11176*, 2024.

[125] Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang, "MANIQA: Multi-Dimension Attention Network for No-Reference Image Quality Assessment," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 1191–1200.

Bio

MohammadAli Hamidi is a PhD candidate at the University of Cagliari in the Department of Electrical and Electronic Engineering (DIEE/UdR CNIT), where he is under the supervision of Prof. L. Atzori and A. Floris. He earned his Bachelor of Science degree in Computer Engineering – Software in 2013, followed by a Master of Science in Computer Engineering – Artificial Intelligence in 2015. Since 2022, he has been a member of the Net4U Laboratory (https://sites.unica.it/net4u/) within the DIEE, as well as a member of the Italian University Consortium for Telecommunications (CNIT). His research interests include machine learning, computer vision, image processing, Multimedia Quality of Experience (QoE), Image/Video and Point Cloud Quality Assessment (PCQA).

Acknowledges

This thesis was produced while attending the PhD programme in Electronic and Computer Engineering at the University of Cagliari, Cycle XXXVIII, with the support of a scholarship financed by the Ministerial Decree no. 351 of 9th April 2022, based on the NRRP - funded by the European Union - NextGenerationEU - Mission 4 "Education and Research", Component 1 "Enhancement of the offer of educational services: from nurseries to universities" - Investment 4.1 "Extension of the number of research doctorates and innovative doctorates for public administration and cultural heritage".