# Learning to Judge Motion: Likelihood-Based Evaluation with Normalizing Flows

Mahamat Issa Choueb*, Praveen Kumar Sekharamantry*†, Giulia Martinelli*, Alessandro Floris‡, Nicola Conci*

*Dept. of Information Engineering and Computer Science, University of Trento, Trento, Italy
†Dept. of Computer Science and Engineering, GITAM School of Technology, Visakhapatnam, India
‡Dept. of Electrical and Electronic Engineering, University of Cagliari, Sardinia, Italy

Email: mahamatissa.choueb@unitn.it,pk.sekharamantry@unitn.it,alessandro.floris84@unica.it,
giulia.martinelli-2@unitn.it,nicola.conci@unitn.it

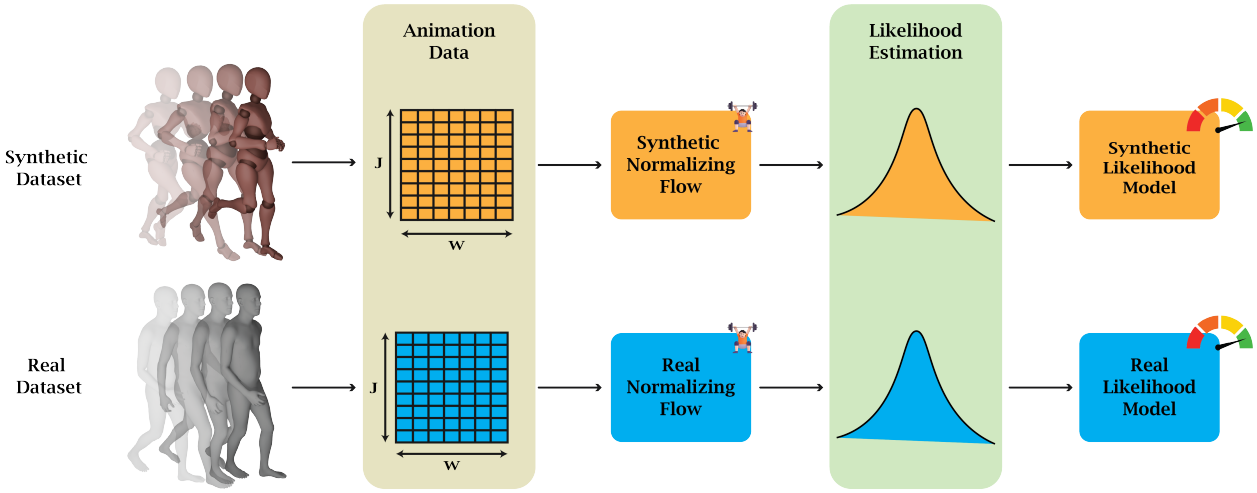


Fig. 1: Method overview. Animation data is processed through real and synthetic Conditional Glow models to estimate motion likelihood and assess alignment with real or synthetic distributions.

***Abstract*—Human motion generation aims to synthesize realistic and diverse body movements from inputs such as text, poses, or environmental cues. Effective evaluation of generated motion requires metrics that capture realism, accuracy, and diversity. Existing approaches tend to overemphasize similarity to ground truth, assume Gaussian distributions, ignore temporal structure, or rely heavily on subjective assessments. To overcome these limitations, we introduce a novel evaluation method based on Conditional Glow Normalizing Flows, which model the full data distribution without assuming gaussianity and without requiring paired reference samples.Trained on two contrasting datasets, AMASS, which contains high-fidelity motion capture performed by real actors, and Mixamo, a synthetic dataset with retargeted animations, our model provides exact likelihood estimates that quantify the degree to which a motion sequence aligns with real or synthetic distributions. This enables not just binary classification, but a nuanced evaluation of physical and kinematic plausibility. We conduct extensive experiments across several state-of-the-art generative models, demonstrating that our likelihood-based metric offers an interpretable tool for motion validation.**

## I. Introduction

Thanks to the rapid evolution of deep generative models and superior simulation technologies [1], the adoption of synthetic data in computer vision and computer graphics tasks has gradually seen a progressively increasing usage. In particular, generative models capable of replicating human motion have proven particularly effective. However, human observers are still able to spot the synthetic nature of the animation, because of subtle behavioral cues that prevent achieving a perfect level of realism [2]. To this aim, while human observation is key, the definition of a proper objective validation framework is highly desirable.

Therefore, the research question we identified was whether it is possible to devise a metric capable of

discriminating between real human motion and synthetically generated video sequences. This proves to be particularly challenging, as human motion represents the final outcome of numerous complex articulations which, although constrained by body mechanics, can assume an infinite number of physically plausible forms. To tackle the problem, typical error metrics proposed in a deterministic form cannot be used to tackle the problem at hand, since they are not capable of handling multiple combined transformations (translations, rotations, scaling) evolving over time. Instead of relying solely on a binary discrimination between real and synthetic motion, we propose a probabilistic metric that estimates the degree of belonging of an animation to either real or synthetic distributions. In this perspective, the solution we propose, consists of investigating a Conditional Glow [3] Normalizing Flows, which is trained on motion data to describe the underlying distribution of actual human data and to capture this on a sense of likelihood estimation.

This approach estimates the likelihood of the input sequences through set of invertible transformations and convert them into a gaussian distribution. The advantage of this method lies in the fact that it does not rely on labeled validation criteria and solely focusing on a probabilistic discriminator to set apart synthetic and real motions regarding the fitted learned distribution.

The method we propose has been trained using two commonly adopted datasets used in human motion modeling and animation, namely the AMASS [4], and the Mixamo dataset [5]. The AMASS dataset serves as our real motion corpus. It is a large-scale collection of high-quality motion capture recordings captured with professional systems and performed by multiple human actors across diverse tasks. In contrast, Mixamo consists of synthetic animations generated via retargeting motion sequences to a variety of stylized characters. By training two separate Conditional Glow models on AMASS and Mixamo respectively, we obtain a dual-distribution representation (Fig.1). Motions with high likelihood under the AMASS model exhibit both physical and kinematic realism. Conversely, motions fitting the Mixamo distribution may show good joint coherence but lack physical plausibility. Motions that score poorly on both distributions often exhibit issues in both structure and realism.

The contributions of this paper are the following:

- A novel application of Glow Normalizing Flows to quantitatively evaluate human motion data.
- A likelihood-based approach that can be trained without paired data and without assuming Gaussianity of the data.
- The resilience and generalization ability on popular motion datasets.

The source code and the relevant data used in the experimental evaluation are available at https://github.com/mmlab-cv/AnimFlow.

## II. Related Work

### A. Evaluating Motion Generative Model

Human motion generation aims to synthesize realistic character movements through sequences of natural postures, often conditioned on external signals such as scene context, audio, or text [6]. Despite progress in modeling, evaluating generated motion remains a major challenge due to the complexity of human movement, the one-to-many nature of generation tasks, and the subjectivity in human perception.

Reliable evaluation metrics are essential but difficult to define. In this work, we focus on fidelity-based metrics, which assess the naturalness, smoothness, and plausibility of generated motion—key indicators of human-likeness. A common approach compares generated sequences to ground truth using metrics such as Mean Squared Error (MSE) or Normalized Power Spectrum [7], and pose estimation metrics like Percentage of Correct 3D Keypoints (PCK) [8]. However, these require paired data and fail when outputs deviate significantly from reference motions.

To overcome this, distribution-based metrics assess the overall quality of generated motion rather than individual sequences. Metrics like Fréchet Distance (FD) [9] and Average Variance Error (AVE) [10] operate directly in the motion space, analyzing geometric or statistical properties of pose sequences. FID-like approaches, on the other hand, compute distances in a learned feature space using standalone motion encoders [11]. A recent example is Fréchet Motion Distance (FMD) [12], which uses a Transformer-based autoencoder to capture latent distributions and assess motion artifacts such as foot skating and over-smoothing. However, FD-based metrics assume a gaussianity in the latent space, which may not hold in real-world motion data and can bias results.

Our method overcomes these limitations using Normalizing Flows, which model the full probability distribution of motion data without assuming Gaussianity and without requiring paired ground-truth sequences.

### B. Normalizing flows as evaluation metric.

Normalizing Flows (NFs) are a class of deep generative models that transform simple probability distributions (e.g., Gaussian) into complex, structured ones through a sequence of invertible and differentiable mappings. Among these, Conditional Glow is particularly suited for modeling structured data such as 3D human skeletons, thanks to its multi-scale architecture and fine-grained control over local and global dependencies.

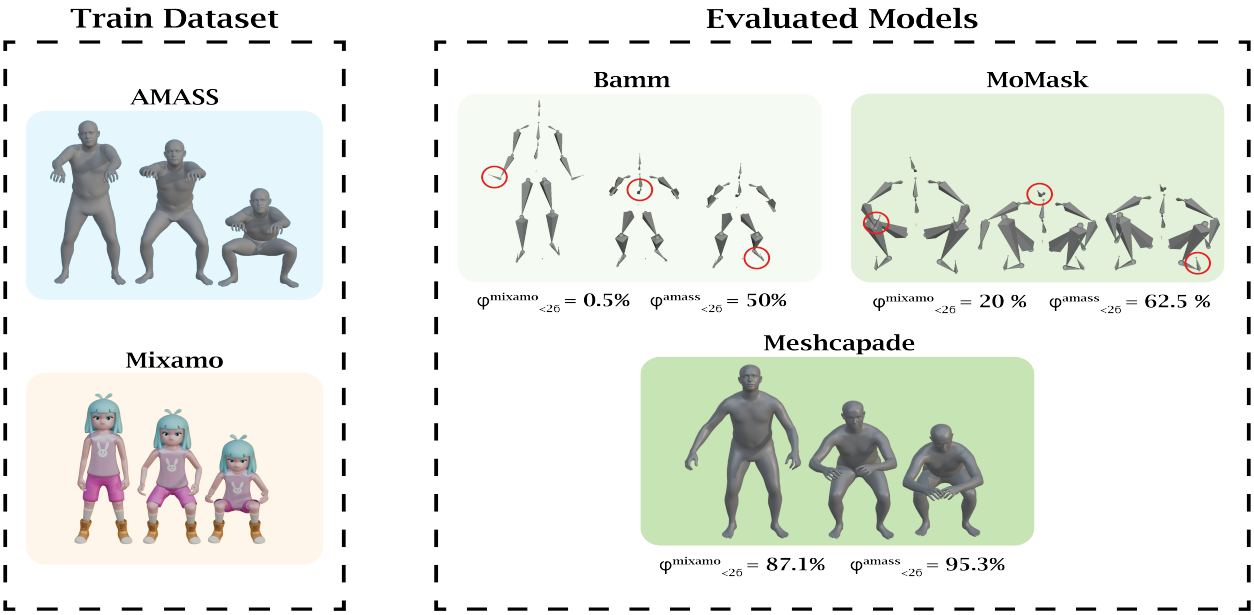


Fig. 2: Qualitative comparison of a squat animation generated by BAMM, MoMask, and Meshcapade. Red circles highlight joint rotation errors.

While Normalizing Flows have shown success in various domains for density estimation and anomaly detection, their use as evaluation metrics remains largely unexplored in human motion synthesis. For instance, Dias et al. [13] and Milković et al. [14] applied NFs to identify trajectory anomalies in robotics and surveillance systems. Similarly, compact flow-based approaches have been proposed for detecting synthetic patterns in industrial images [15] or localizing anomalies in crowd movement [16].

To the best of our knowledge, we are the first to employ Normalizing Flows as an evaluation metric in the context of human motion generation, using them to model the probability distributions of real and synthetic motions.

## III. Methodology

We propose a novel evaluation metric based on a *Conditional Glow* Normalizing Flow architecture, trained on both real and synthetic motion data—specifically, the AMASS and Mixamo datasets. All animations are provided in BVH format, a standard file format for motion capture data. This choice ensures that our method is compatible with a wide range of motion generation pipelines, as BVH is widely supported in computer graphics and animation software. The animation is represented by relative joint rotations in quaternions format, which are normalized to the range $[-1, 1]$. Each animation sequence is then structured as a 4D tensor of shape $(b, j, w, q)$, where $b$ is the batch size, $j$ is the number of joints, $w = 64$ is the fixed temporal window length, and $q = 4$ corresponds to the number of quaternion components. To ensure temporal consistency during training, each animation is divided into non-overlapping windows of fixed length. Only complete windows are retained, discarding any segments that do not meet the required frame count. This guarantees that all training samples are of uniform shape and length. The resulting windowed tensors are then flattened into vectors of dimension $(b, j \cdot w \cdot q)$ and passed to the model. By operating on temporally coherent, fixed-length motion segments, our approach captures both local and global motion dynamics, enabling a reliable, likelihood-based evaluation of motion.

Our approach is built upon the *Conditional Glow* model, a type of normalizing flow that transforms complex data distributions into simpler ones—typically a multivariate Gaussian—via a sequence of invertible and differentiable transformations [3]. The model is trained to minimize the negative log-likelihood over a dataset $\mathcal{D}$:

$$\mathcal{L}(\mathcal{D}) = \frac{1}{N} \sum_{i=1}^{N} -\log p_\theta(\mathbf{x}^{(i)}) \tag{1}$$

where $\mathbf{x}^{(i)} \in \mathbb{R}^n$ denotes the flattened quaternion representation of joint rotations over a temporal window, and $p_\theta$ is the learned likelihood function.

Glow consists of three main components: ActNorm layers, which normalize activations per channel and eliminate the need for batch normalization; invertible $1 \times 1$ convolutions, which increase model expressiv-

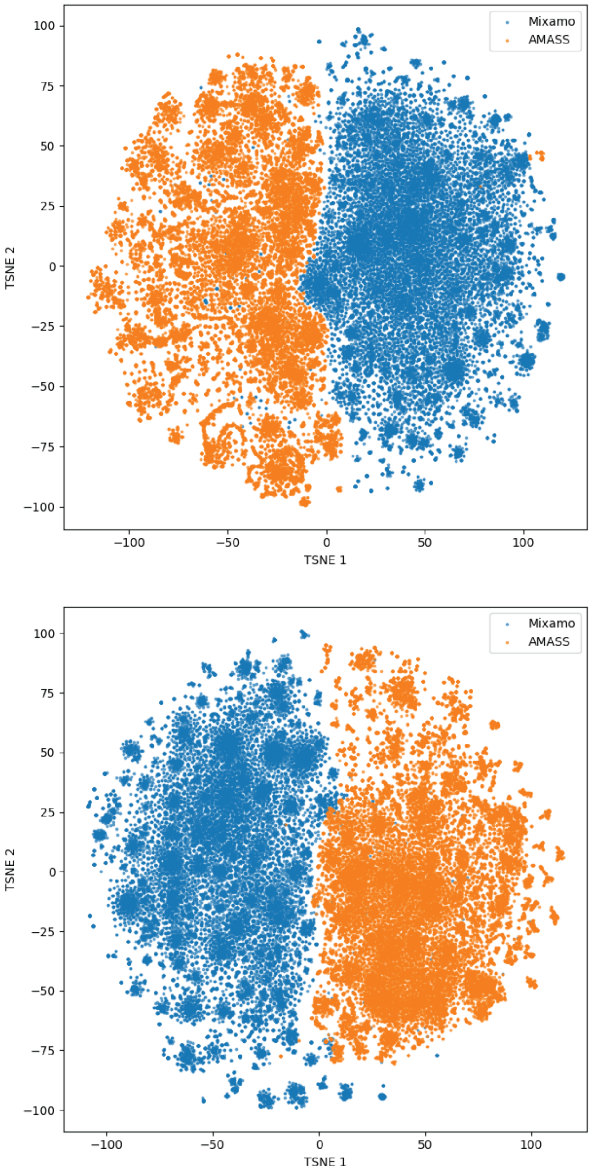


Fig. 3: t-SNE plots of latent spaces from Glow models trained on AMASS (top) and Mixamo (bottom). Real (orange) and synthetic (blue) motions are clearly separated in both cases.

ity by learning cross-channel dependencies; and affine coupling layers, which apply complex but invertible transformations conditioned on subsets of the input. Our architecture comprises three flow levels, each with eight flow steps.

To enable likelihood-based evaluation across real and synthetic data domains, we train two separate Conditional Glow models: synthetic (Mixamo) and real (AMASS). Training on these two contrasting datasets yields two distinct probabilistic motion distributions. The AMASS-trained model assigns higher likelihoods to physically and kinematically plausible motions, while the Mixamo-trained model captures motions that are structurally coherent but potentially physically unrealistic. During inference, we compute the log-likelihood of a given motion under both models. A high score under the AMASS model indicates a motion consistent with real human movement. A high Mixamo likelihood but low AMASS likelihood suggests synthetic quality with good joint coherence but poor physical grounding. Low likelihood across both distributions typically signals unrealistic or corrupted motion.

## IV. Results and Discussion

### A. Dataset

We train and evaluate our model using two datasets: Mixamo[1] and AMASS [4]. Mixamo is a large-scale collection of synthetic 3D character animations provided by Adobe, containing thousands of motion clips covering a wide variety of human actions. These motions are generated from a limited set of base animations and are automatically retargeted to a diverse set of stylized, often cartoon-like, 3D characters. Due to this automated retargeting process, Mixamo animations exhibit consistent joint kinematics across characters but may suffer from physical artifacts such as foot sliding, lack of contact realism, and limited stylistic diversity.

In contrast, AMASS is a large-scale dataset that aggregates motion capture data from 24 professional, optical marker-based MoCap datasets and fits them into a unified representation using the SMPL body model [17]. AMASS provides temporally coherent pose and shape parameters across a wide range of subjects, actions, and recording conditions. As all motions are performed by real human actors and captured in controlled environments, AMASS offers high-quality, physically plausible, and naturalistic motion data, making it a reliable benchmark for learning-based motion models.

### B. Implementation Details

Our model is implemented using PyTorch Lightning and trained on a single NVIDIA RTX 4090 GPU. The input consists of animation sequences represented as tensors of shape $(b, j, w, q)$, where $b$ is the number of animation samples (batch size), $j$ is the number of joints, $w = 64$ is the temporal window length (i.e., number of frames), and $q$ corresponds to the quaternion representation of joint rotations. Each animation tensor is flattened into a vector of shape $(b, j \cdot w \cdot q)$ before being passed to the model.

We adopt a Conditional Glow-based normalizing flow architecture with four steps per block, two blocks, and

[1]https://www.mixamo.com/

TABLE I: In-distribution ($\phi_{<2\sigma}$) and out-of-distribution ($\phi_{>3\sigma}$) scores for models and datasets under Glow models trained on Mixamo (top) and AMASS (bottom). Higher $\phi_{<2\sigma}$ indicates better alignment with the training distribution.

| | Reference | Category | Conference | $\phi_{<\mathbf{2}\sigma}$ | $\phi_{>\mathbf{3}\sigma}$ |
|---|---|---|---|---|---|
| **Train on Synthetic (Mixamo)** | Momask [18] | Model | CVPR2024 | 3.13% | 94.06% |
| | Meshcapade[2] | Model | Company | 70.56% | 21.77% |
| | Mixamo[1] | Dataset | Company | 95.82% | 2.33% |
| | AMASS [4] | Dataset | ICCV2019 | 62.66% | 28.82% |
| | Bamm [19] | Model | ECCV2024 | 1.42% | 96.16% |
| **Train on Real (AMASS)** | Momask [18] | Model | CVPR2024 | 8.03% | 91.11% |
| | Meshcapade[3] | Model | Company | 99.19% | 0.00% |
| | Mixamo[1] | Dataset | Company | 66.81% | 26.00% |
| | AMASS [4] | Dataset | ICCV2019 | 99.85% | 0.00% |
| | Bamm [19] | Model | ECCV2024 | 19.73% | 73.01% |

1024 hidden units in each linear layer. The model is trained for 100 epochs using the Adam optimizer with a learning rate of $2\times10^{-4}$, a weight decay of $1\times10^{-4}$, and a ReduceLROnPlateau scheduler that halves the learning rate when the validation loss plateaus. We use a batch size of 64 during training.

### *C. Experiments*

We design our experimental setup in two main stages. First, we validate whether the Conditional Glow models trained on AMASS and Mixamo can distinguish real from synthetic motion. Then, we apply our learned metric to evaluate the output of several state-of-the-art motion generative models. Specifically, we assess motions generated by MoMask [18] and BAMM [19], as well as those produced by a proprietary model developed by Meshcapade. These evaluations are conducted without retraining the metric, allowing us to demonstrate its generalizability across generative pipelines.

**Evaluation Metrics**. To evaluate the quality of generated motion sequences, we adopt the Out-of-Distribution (OOD) Fraction metric, which measures how well motion samples conform to the distribution learned by the normalizing flow. The OOD fraction quantifies the proportion of samples whose latent codes fall outside a defined confidence interval around the mean of the flow's prior distribution.

It is formally defined as:

$$\phi_{>k\sigma} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\left(|z_i| > k\sigma\right) \tag{2}$$

where $N$ is the total number of motion samples (e.g., frames or windows), $z_i$ is the latent code of sample $i$, $\sigma$ is the standard deviation of the standard normal prior (typically 1), and $\mathbf{1}(\cdot)$ is the indicator function.We compute this metric both for out-of-distribution samples (e.g., $\|z_i\| > 3\sigma$) to detect unlikely or implausible motions, and for in-distribution samples (e.g., $\|z_i\| < 2\sigma$) to assess how much of the motion data lies near the peak of the learned distribution. A higher in-distribution fraction indicates better alignment with the learned motion manifold, while a higher OOD fraction suggests greater deviation from physically plausible human motion

**Real vs. Synthetic Validation.** In the first stage of our evaluation, we assess the ability of the two Conditional Glow models—trained independently on AMASS and Mixamo—to distinguish between real and synthetic motion data. Table I reports the quantitative results in terms of in-distribution and out-of-distribution (OOD) fractions. When evaluating the model trained on real motion (AMASS), we observe that 99.85% of the real test samples fall within the high-likelihood region. Conversely, for the model trained on synthetic motion (Mixamo), 95.8% of Mixamo samples fall within $[-2\sigma, 2\sigma]$, while only 62.66% of AMASS samples do—demonstrating the model's capacity to capture the structure of synthetic motion and reject real data as out-of-distribution. These findings are further supported by the t-SNE visualization in Figure 3, where the latent embeddings of motion sequences reveal a clear separation between real and synthetic samples. The clustering behavior in the latent space confirms that both models successfully learn distinct representations for their respective data distributions.

**Evaluation on Generative Models.** In the second stage of our evaluation, we apply our proposed likelihood-based metric to motion sequences generated by three different human motion generation models: MoMask, BAMM, and Meshcapade. These models vary widely in terms of output quality and design philosophy, allowing us to test the robustness of our evaluation method across generative settings. As shown in Table I, Meshcapade achieves the highest scores for both the AMASS and Mixamo-trained models, with in-distribution fractions of 99.19% and 70.56%, respectively. This reflects the high physical and kinematic quality of Meshcapade's output. The motions it generates are temporally coherent, anatomically plausible, and free from common artifacts like foot sliding or sudden discontinuities.

MoMask and BAMM yield lower scores, highlighting common failure cases in generative motion. MoMask, in

particular, struggles with long-term consistency: while it produces locally plausible poses, its sequences lack fluid transitions and often exhibit flickering and physical inconsistencies. These issues are especially apparent in the squat animation shown in Figure 2, where MoMask's motion fails to include a realistic downward trajectory of the pelvis. Instead, the animation begins close to the final squat position, with minimal global motion across frames. Additionally, several joints display clearly incorrect quaternion rotations: the hands and feet twist unnaturally, and the head bends at an unrealistic angle—nearly 90 degrees relative to the neck. These joint-level artifacts are highlighted in red in the figure and explain the drop in in-distribution score (62.5% for the AMASS model and 20% for the Mixamo model).

BAMM demonstrates better global motion, capturing the expected descent and ascent in the squat. However, it suffers from inconsistent or distorted joint behavior. As shown in Figure 2, hands bend unnaturally, the spine appears overextended, and the feet fail to maintain consistent contact with the ground. These errors are reflected in the likelihoods: 50% for AMASS and only 0.5% for Mixamo, suggesting that while the motion is globally structured, it lacks fine-grained physical realism.

In Figure 2, report also the per-method likelihoods for this specific animation. These values are consistent with the qualitative observations: Meshcapade performs best overall, MoMask suffers from temporal and structural errors, and BAMM strikes a balance between plausible structure and flawed local articulation. Together, the results demonstrate that our evaluation metric effectively captures both global motion quality and local joint fidelity, aligning closely with visual perception.

## V. Conclusion

In this work, we introduced a novel, likelihood-based evaluation metric for human motion quality using Conditional Glow normalizing flows. By training separate models on real (AMASS) and synthetic (Mixamo) motion datasets, we learned two distinct probabilistic distributions that enable fine-grained assessment of motion plausibility without requiring paired ground-truth data. Our experiments show that the dual-model approach reliably distinguishes real from synthetic motion, capturing both physical and kinematic fidelity. The metric aligns well with visual assessments and successfully identifies common generative artifacts such as flickering and joint errors. Unlike traditional methods, it offers interpretable, data-driven scores without requiring ground-truth pairing or manual thresholds.

## VI. Acknowledgment

Funded by the European Union under NextGenerationEU , Mission 4 Component 2 - CUP C69J24000180004.

## References

[1] J. Xu, H. Li, and S. Z. and, "An overview of deep generative models," *IETE Technical Review*, vol. 32, no. 2, pp. 131–139, 2015.

[2] U. Zabala, I. Rodriguez, J. M. Martínez-Otzeta, and E. Lazkano, "Modeling and evaluating beat gestures for social robots," *Multimedia Tools and Applications*, vol. 81, p. 3421–3438, Aug. 2021.

[3] D. P. Kingma and P. Dhariwal, "Glow: Generative flow with invertible 1x1 convolutions," *Advances in neural information processing systems*, vol. 31, 2018.

[4] N. Mahmood, N. Ghorbani, N. F. Troje, G. Pons-Moll, and M. J. Black, "Amass: Archive of motion capture as surface shapes," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5442–5451, 2019.

[5] Z. Guo, J. Xiang, K. Ma, W. Zhou, H. Li, and R. Zhang, "Make-it-animatable: An efficient framework for authoring animation-ready 3d characters," *arXiv preprint arXiv:2411.18197*, 2024.

[6] W. Zhu, X. Ma, D. Ro, H. Ci, J. Zhang, J. Shi, F. Gao, Q. Tian, and Y. Wang, "Human motion generation: A survey," 2023.

[7] B. Li, Y. Zhao, S. Zhelun, and L. Sheng, "Danceformer: Music conditioned 3d dance generation with parametric motion transformer," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, pp. 1272–1279, 2022.

[8] Y. Yang and D. Ramanan, "Articulated human detection with flexible mixtures of parts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 12, pp. 2878–2890, 2013.

[9] J. Wang, Y. Rong, J. Liu, S. Yan, D. Lin, and B. Dai, "Towards diverse and natural scene-aware 3d human motion synthesis," 2022.

[10] Z. Zhou and B. Wang, "Ude: A unified driving engine for human motion generation," 2022.

[11] A. Ghosh, N. Cheema, C. Oguz, C. Theobalt, and P. Slusallek, "Synthesis of compositional animations from textual descriptions," *CoRR*, vol. abs/2103.14675, 2021.

[12] A. Maiorca, H. Bohy, Y. Yoon, and T. Dutoit, "Objective evaluation metric for motion generative models: Validating fréchet motion distance on foot skating and over-smoothing artifacts.," in *Proceedings of the 16th ACM SIGGRAPH Conference on Motion, Interaction and Games*, MIG '23, (New York, NY, USA), Association for Computing Machinery, 2023.

[13] M. L. D. Dias, C. L. C. Mattos, T. L. C. da Silva, J. A. F. de Macêdo, and W. C. P. Silva, "Anomaly detection in trajectory data with normalizing flows," in *2020 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, 2020.

[14] F. Milkovic, L. Posilović, D. Medak, M. Subasic, S. Loncaric, and M. Budimir, "Franomaly: flow-based rapid anomaly detection from images," *Applied Intelligence*, vol. 54, pp. 1–14, 02 2024.

[15] I. Pathirannahalage, V. Jayasooriya, J. Samarabandu, and A. Subasinghe, "A comprehensive analysis of real-time video anomaly detection methods for human and vehicular movement," *Multimedia Tools and Applications*, vol. 84, pp. 7519–7564, 04 2024.

[16] T. Ganokratanaa, S. Aramvith, and N. Sebe, "Unsupervised anomaly detection and localization based on deep spatiotemporal translation network," *IEEE Access*, vol. 8, pp. 50312–50329, 2020.

[17] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "SMPL: A skinned multi-person linear model," *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, vol. 34, pp. 248:1–248:16, Oct. 2015.

[18] C. Guo, Y. Mu, M. G. Javed, S. Wang, and L. Cheng, "Momask: Generative masked modeling of 3d human motions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1900–1910, 2024.

[19] E. Pinyoanuntapong, M. U. Saleem, P. Wang, M. Lee, S. Das, and C. Chen, "Bamm: bidirectional autoregressive motion model," in *European Conference on Computer Vision*, pp. 172–190, Springer, 2024.