Computers and Electrical Engineering 139 (2026) 111385

Contents lists available at ScienceDirect

Computers and Electrical Engineering

journal homepage: www.elsevier.com/locate/compeleceng

Fairness in vision-language text-to-image retrieval: Natural biases and adversarial threats☆

Dario Lazzaro [a,b], Antonio Emanuele Cinà [a], Fabio Roli [a,c], Davide Anguita [a], Battista Biggio [c], Giuseppe Laurita [d], Luca Oneto [a,*]

[a] University of Genoa, Via Opera Pia 11a, Genoa, 16145, Italy
[b] Sapienza University of Rome, Via Ariosto 25, Rome, 00185, Italy
[c] University of Cagliari, Via Marengo 3, Cagliari, 09100, Italy
[d] Sky Italia, Via Monte Penice 7, Milan, 20138, Italy

ARTICLE INFO

Dataset link: https://github.com/lazzd/FairLess

Keywords:
Vision-language pretrained models
Text-to-image retrieval systems
Fairness
Natural setting
Adversarial setting
Metrics
White-box attack
Mitigation

ABSTRACT

Vision-language pretrained models have substantially improved text-to-image retrieval by aligning visual and textual information in a shared semantic space. However, these systems may produce unfair outcomes across demographic groups while their vulnerability to adversarial manipulation remains largely unexplored.

This work studies fairness in text-to-image retrieval under both natural and adversarial settings. We address three limitations of prior research: its focus on neutral queries that do not explicitly mention sensitive attributes; its limited attention to demographic imbalance in the retrieval database; and its assumption that unfairness arises naturally from data rather than through malicious intervention.

We propose a unified framework for evaluating fairness in text-to-image retrieval. In the natural setting, we analyze how unfairness depends on both the pretrained model and the demographic composition of the retrieval database. In the adversarial setting, we define a realistic threat model in which an attacker injects a limited number of malicious images into the database. We introduce FairLess, a poisoning attack designed to amplify demographic bias while preserving semantic relevance, and evaluate a mitigation strategy to reduce its impact. We also extend fairness evaluation to attribute-labeled queries, where the target image is associated with a specific demographic group, and introduce metrics that jointly capture retrieval accuracy and fairness.

Experiments on two demographically annotated versions of MSCOCO and using CLIP-based models, fairness-mitigated variants, and BLIP-2, show that unfairness stems from both model bias and database polarization. Adversarial manipulation further amplifies demographic disparities, highlighting the need for robustness-aware fairness evaluation and mitigation in text-to-image retrieval systems.

☆ This article is part of a Special issue entitled: 'Multimodal Data Security' published in Computers and Electrical Engineering.

* Corresponding author.
E-mail addresses: dario.lazzaro@edu.unige.it (D. Lazzaro), antonio.cina@unige.it (A.E. Cinà), fabio.roli@unige.it (F. Roli), davide.anguita@unige.it (D. Anguita), battista.biggio@unica.it (B. Biggio), giuseppe.laurita@skytv.it (G. Laurita), luca.oneto@unige.it (L. Oneto).

https://doi.org/10.1016/j.compeleceng.2026.111385
Received 29 January 2026; Received in revised form 5 July 2026; Accepted 7 July 2026

0045-7906/© 2026 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

1. Introduction

Recent Vision-Language Pretrained Models (VLPs), trained on large-scale multimodal datasets, have substantially advanced cross-modal understanding [1–3]. These models capture complex interactions between visual and linguistic modalities, supporting downstream tasks such as zero-shot classification and Text-to-Image Retrieval (T2IR) [4]. Notably, models such as CLIP [4] and BLIP-2 [5] embed images and text into a shared semantic space, where their relevance can be directly measured through similarity functions. This line of research has driven significant progress in T2IR and related retrieval tasks, enabling users to retrieve images from large-scale Image Retrieval Databases (IRDs) by exploiting the semantic alignment between natural language queries and visual content [4–6].

Despite these advancements, recent studies have highlighted critical concerns regarding fairness [7–9] and safety, particularly in relation to data poisoning [10–14], in VLP-based T2IR. Fairness issues arise when retrieval outcomes systematically differ across sensitive attributes, such as disproportionately returning images of a particular gender or race, or exhibiting higher error rates for queries reflecting specific political viewpoints. By contrast, data poisoning refers to adversarial injection of malicious samples into the IRD with the intent to manipulate retrieval results.

In the context of T2IR system (un)fairness, a considerable body of research has already been conducted. Much of this work has focused on *neutral* queries, i.e., descriptions of people that do not explicitly specify sensitive attributes (e.g., "a person cooking" or "a doctor giving a lecture") [7–9]. In such settings, fairness is typically interpreted as requiring balanced representation across sensitive attribute classes in the retrieved results, rather than a predominance of any particular group [8]. Most prior studies have concentrated on intrinsic biases in T2IR systems, stemming from VLPs trained on biased IRDs [7]. To mitigate these issues, two main strategies have been explored. One line of work attempts to suppress the influence of sensitive attributes in embedding representations [7,15,16], while another seeks to produce fairer VLPs by training dedicated debiasing modules and integrating them into the model architecture [17].

In the context of T2IR system sensitivity to data poisoning attacks, prior research has demonstrated significant vulnerabilities arising from the injection of malicious samples into the IRD with the goal of manipulating retrieval results [10,11]. These attacks are commonly classified into three categories: *indiscriminate* attacks, which aim to increase the overall test error [18,19]; *targeted* attacks, which attempt to mislead the system into misclassifying specific examples [20,21]; and *backdoor* attacks, which introduce trigger-labeled samples into the training data so that, at test time, the presence of the trigger manipulates the output [22,23]. Such attacks have been shown to severely degrade system performance, disrupt business operations, and undermine user safety and trust [12–14]. They can be executed under a *white-box* threat model, where the adversary has full access to the system's architecture and parameters [24,25], or under a *black-box* model, where no internal knowledge is assumed [26,27].

In this work, we investigate three open research challenges related to the (un)fairness of T2IR systems.

The first issue stems from a limitation of prior studies, which have primarily concentrated on so-called *neutral* queries. This focus makes it challenging to design metrics that simultaneously account for both accuracy and sensitivity to protected attributes. As a result, establishing a clear and formal definition of fairness in T2IR remains an open research problem [8].

The second issue concerns the sources of bias underlying unfairness in T2IR systems. While earlier studies primarily examined intrinsic biases inherited from VLPs trained on skewed datasets [28], they largely neglected biases introduced by the *polarization* of the image retrieval database (IRD), i.e., the unequal representation of demographic groups [29]. When certain subgroups of the population are systematically over- or under-represented, T2IR systems face substantial challenges in ensuring fair retrieval [29]. These imbalances often reflect historical patterns; for instance, women may be underrepresented in specific professions, leading systems trained on such data to return predominantly male results when queried for those roles [30].

The third issue is more fundamental: prior research has largely assumed that unfairness arises from the *natural* presence of bias in training data or models, while neglecting the possibility of an *adversarial* setting in which a malicious actor deliberately injects biases to exacerbate unfair outcomes. To the best of our knowledge, no prior work has examined the sensitivity of VLP-based T2IR systems to poisoning attacks specifically designed to *induce bias* in retrieval results, despite such attacks already being studied in other tasks such as classification [31] or in fairness-aware re-ranking stages of image retrieval pipelines [32]. Existing (un)fairness mitigation strategies also exhibit important limitations. First, some methods require retraining large-scale models from scratch [7], which is often infeasible due to the substantial computational cost [4]. Second, most approaches are not designed to address *attribute-specific queries*, where the query explicitly references a particular class of a sensitive attribute. For example, in the case of *gender-specific queries* (e.g., "a man cooking" or "a woman driving"), fairness requires accurately preserving the gender specified in the query [30]. In such cases, suppressing attribute-related features may degrade retrieval accuracy or yield misrepresentative results [8]. Finally, even debiasing strategies that are effective in benign conditions remain vulnerable in adversarial scenarios, particularly under *white-box* attacks [24], where an adversary with full knowledge of the retrieval pipeline can craft targeted manipulations to circumvent fairness-preserving mechanisms.

To address these issues, we investigate the fairness of VLP-based T2IR systems under both *natural* and *adversarial* settings. In the *natural setting*, we analyze disparities arising from biases in VLPs or from imbalanced IRD data distributions. In the *adversarial setting*, we examine how a malicious actor can deliberately induce or amplify such disparities, defining a white-box threat model in which the attacker has full knowledge of the T2IR system and can inject a limited number of poisoned samples into the IRD. Within this framework, we introduce `FairLess`, a novel data poisoning attack designed to manipulate the system so that, for a given query or any of its semantically equivalent variants, the T2IR model returns images that are semantically appropriate in content but systematically biased along a sensitive attribute. For example, in response to queries such as "CEO of a big company" or "female CEO of an industry", the system may be manipulated to consistently retrieve images depicting male CEOs, thereby reinforcing biased

representations. To mitigate this problem, we also evaluate a possible defense strategy based on JPEG compression [33]. Unlike prior work, we also propose a new class of *attribute-labeled queries*, i.e., queries explicitly associated with sensitive attribute classes. This enables fairness evaluation on real-world benchmarks by assessing retrieval performance across demographic groups, thereby allowing disparities to be systematically measured. Finally, we introduce a set of novel accuracy and fairness metrics tailored to multimodal retrieval, designed to address gaps in existing evaluations by quantifying disparities and analyzing their underlying sources.

To empirically validate our proposal, we conduct a series of experiments on the MSCOCO dataset [34], focusing on both gender and race attributes. For gender, we adopt the labeling procedure introduced in [35], while for race, we rely on the demographic annotations provided in [36]. Our study employs a standard T2IR pipeline with several state-of-the-art VLPs. In particular, we evaluate two CLIP-based architectures [4] as well as BLIP-2 [5]. We further include two debiased variants of CLIP [17], each specifically trained to mitigate bias with respect to either gender or race. Our results reveal persistent fairness biases in T2IR systems, stemming both from the underlying VLPs and from imbalances in the IRD. Moreover, we demonstrate the effectiveness of `FairLess` in introducing or amplifying these biases, even when the VLPs have been fine-tuned with fairness mitigation strategies. Finally, the initial assessment of the mitigation strategy based on input preprocessing (i.e., JPEG compression) shows that it is possible to reduce the effectiveness of the attack under the considered setting, while not fully addressing the observed disparities. These findings underscore the urgent need for further research on fairness in T2IR and the development of robust, adversary-resistant mitigation approaches.

To summarize, this work provides a series of novel contributions:

- We analyze the limitations of current fairness evaluations in VLP-based T2IR systems, highlighting the role of attribute-labeled queries, dataset polarization, and the lack of adversarial analysis;
- We introduce a unified framework to study fairness under two settings: a natural setting, where disparities arise from model and dataset biases, and an adversarial setting, where a malicious actor injects samples into the IRD;
- In the adversarial setting, we define a realistic threat model and propose `FairLess`, a novel poisoning attack that induces demographic bias while preserving semantic relevance;
- We propose evaluation metrics that jointly capture retrieval accuracy and fairness across demographic groups, enabling evaluation on real-world benchmarks;
- We study fairness across demographic attributes, focusing on gender and race, and highlight how biases arise from both models and datasets and can be amplified under adversarial manipulation;
- We study a possible mitigation strategy for the proposed adversarial setting, showing that it can partially reduce the effectiveness of the attack.

The remainder of the paper is organized as follows. Section 2 reviews the current state of the art on (un)fairness in T2IR systems. Section 3 introduces preliminary definitions that are instrumental for understanding the technical contributions of the paper. In Section 4, we formally define the *natural* and *adversarial* (un)fairness settings along with the associated metrics. Section 5 then specifies the threat model considered for the *adversarial setting*. Building on this, Section 6 presents `FairLess`, a novel attack targeting fairness under the defined adversarial threat model. To empirically demonstrate the relevance of the proposed framework, Section 7 details the experimental setup, and Section 8 reports and discusses the results. Finally, Section 9 concludes the paper.

2. Related work

In this section, we review the state of the art in relation to the contribution of our work, highlighting three main open issues concerning (un)fairness in T2IR systems. First, prior studies have mostly considered *neutral* queries, thereby limiting the scope of fairness metrics. Second, they have primarily emphasized biases inherited from VLP training data, while neglecting those introduced by *polarization* in the IRD. Third, they have largely focused on the *natural setting* (see Section 2.1), overlooking the *adversarial setting* (see Section 2.2), where classical (un)fairness mitigation strategies often prove ineffective. Finally, Section 2.3 consolidates these limitations into specific research gaps and positions our work accordingly.

2.1. Natural setting

In the *natural setting*, several prominent works [7–9,17,37] have been published, all of which exhibit the aforementioned limitations.

Wang et al. [7] focus on the case of gender bias, analyzing the behavior of several VLPs under *neutral* queries and reporting that, for example, CLIP retrieves on average 6.4 male images among the top-10 results. To address this issue, they propose two debiasing strategies: *FairSample*, an in-processing method that incorporates fairness constraints during training, and *CLIP-clip*, a post-processing approach applied after training to adjust the learned representations. Specifically, *FairSample* is designed for models trained from scratch and modifies the contrastive loss used to align image and text embeddings. For queries identified as gender-*neutral*, it introduces gender-balanced sampling of positive and negative pairs, thereby mitigating demographic bias in the learned representations. The method is evaluated using SCAN [38], trained on MS-COCO [34] and Flickr30k [39]. However, since it requires training the model from scratch, *FairSample* cannot be applied to frozen VLPs such as CLIP. To overcome this limitation, the authors propose *CLIP-clip*, a model-agnostic debiasing method applicable post-hoc to pretrained models. It reduces gender bias by identifying and removing visual embedding dimensions most associated with gender. This is achieved offline by computing the mutual information between each dimension and binary gender variables, using the estimator proposed in [40]. At retrieval time,

the top dimensions are removed from both visual and textual embeddings to maintain alignment. While *FairSample* and *CLIP-clip* improve demographic balance for *neutral* queries, they present several important limitations. First, the methods overlook *attribute-labeled* queries, focusing exclusively on *neutral* ones. Second, no analysis is conducted to assess whether retrieval performance differs across demographic groups associated with the query, potentially revealing disparities across attributes. Third, the methods address only VLP-induced bias, ignoring bias arising from the IRD. Finally, because the debiasing is integrated into the retrieval pipeline, the methods are vulnerable in white-box adversarial settings: an attacker with full knowledge of the retrieval process can craft poisoned inputs that concentrate adversarial perturbations in the preserved dimensions or distribute them across all components to evade the dimension-removal mechanism.

Although the work of Wang et al. [7] provides valuable insights into bias in VLPs, Ali et al. [8] highlight the absence of a fairness evaluation based on established metrics. To address this limitation, they introduce a framework for evaluating fairness in VLPs across tasks such as zero-shot classification, image retrieval, and captioning, focusing on human-centric queries. They consider both binary gender and multi-class race attributes. Ali et al. [8] evaluate a series of non-mitigated and mitigated CLIP-based variants, including *CLIP-clip* [7], *FairPCA* [15], and *prompt learning* [16]. *FairPCA* projects CLIP embeddings into a subspace designed to remove sensitive information, such as gender or race, by minimizing the statistical dependence between the projected representations and the protected attributes. *Prompt learning* aims to reduce correlations between retrieval outputs and protected attributes by optimizing learnable text prompts within an adversarial framework. Once trained, these prompts are appended to the text embeddings of the queries to mitigate bias. Both *FairPCA* and *prompt learning* demonstrate improvements in fairness across gender and race attributes in retrieval tasks when evaluated on the FairFace dataset. To establish a simple baseline, the authors also test a query-augmentation strategy. Given a *neutral* query such as "a photo of a doctor", attribute-specific variants are generated by appending explicit references to classes of a sensitive attribute (e.g., "male" and "female", or "light" and "dark"). These augmented queries are submitted independently, and their top-k retrieval results are merged to produce a demographically balanced output. Surprisingly, this naive method outperforms several of the more sophisticated mitigation strategies. Note, however, that on real-world benchmarks such as Flickr30K, their evaluation is limited to methods for mitigating gender bias. Nevertheless, as with [7], Ali et al. [8] do not distinguish among different *attribute-specific* attributes, focus only on VLP-induced bias while ignoring bias arising from the IRD, and neglect the *adversarial setting*, where all mitigation strategies may still fail.

Wang et al. [9] highlight a limitation of methods such as *CLIP-clip*, which operate solely on the visual modality and ignore the interaction between vision and language. To address this issue, they propose *FairCLIP*, which leverages learnable word vector prefixes to extract attribute concepts and guide the training of a projection matrix. This matrix is designed to suppress features associated with sensitive attributes (e.g., gender categories) while preserving those related to task-relevant concepts, referred to as target attributes (e.g., "happy", "glasses"). At inference time, the matrix is applied to the visual embeddings to produce debiased representations. The method demonstrates a favorable trade-off between fairness and retrieval accuracy when evaluated on sensitive attributes such as gender and age. However, as with [7,8], Wang et al. [9] do not distinguish among different *attribute-specific* attributes, focus exclusively on VLP-induced bias while overlooking bias arising from the IRD, and neglect the *adversarial setting*, where the mitigation strategy may still fail.

The recent work of Zhang et al. [17] proposes a mitigation strategy that explicitly addresses a critical failure mode of prior debiasing techniques, a phenomenon the authors refer to as *over-debiasing*. This phenomenon arises when methods designed to remove social bias from image or text embeddings also degrade the cross-modal alignment necessary for downstream tasks, such as accurate retrieval. The authors argue that this effect stems from the assumption that visual and textual embeddings encode bias in structurally similar ways. In contrast, they show that the nature and magnitude of bias differ significantly across modalities, which challenges earlier methods that apply the same debiasing procedure to both [7]. To address this discrepancy, the authors introduce a two-stage debiasing pipeline. The first stage involves training a *bias alignment module*, which maps image and text embeddings into a shared bias subspace, making their respective bias components comparable. The second stage applies a *counterfactual debiasing* procedure that removes these aligned components while preserving neutral semantic content and modality alignment. The method is evaluated on T2IR using Flickr30K and MSCOCO, showing reductions in demographic bias across multiple protected attributes (e.g., gender and race) with limited degradation in retrieval performance. However, the same limitations observed in previous works remain.

Finally, Luo et al. [37] propose an optimal-transport-based mitigation strategy for improving fairness in vision-language models. Their method reduces the Sinkhorn distance between the overall sample distribution and the distributions associated with each demographic group, thereby encouraging better alignment across protected groups. Their analysis shows that VLPs can exhibit substantial disparities across demographic subgroups and that such a mitigation strategy can improve the performance-fairness trade-off. However, their work does not investigate how fairness-mitigated VLPs behave when deployed in T2IR systems, including scenarios involving IRD polarization or adversarial poisoning attacks that directly manipulate the retrieval database.

Overall, our experimental analysis focuses on representative VLP backbones, namely CLIP [4] and BLIP-2 [5], as well as fairness-mitigated variants, specifically CLIP-DEB [17] and the FairCLIP variants by Luo et al. [37]. This perspective is complementary to pipeline-level mitigation approaches, since improving the fairness and robustness of the underlying retriever may provide a stronger basis for additional interventions such as *CLIP-clip* [7], *FairPCA* [15], and *prompt learning* [16], which may be applied on top of VLP-based retrievers but can introduce extra overhead or require pipeline modifications.

2.2. Adversarial setting

While adversarial techniques have been leveraged to mitigate fairness issues [41], the use of adversarial methods to actively induce bias has received considerably less attention. In this sense, Solans et al. [31] investigates how fairness-aware machine learning

models can be deliberately compromised through poisoning attacks. Their study focuses on a binary classification scenario in which fairness algorithms aim to enforce parity between two demographic groups, labeled as privileged and unprivileged, based on a sensitive attribute such as gender. The authors propose a poisoning strategy that degrades demographic parity by injecting a small number of malicious training points. These points are optimized to reduce the likelihood of positive outcomes for the unprivileged group, while preserving high classification accuracy to maintain stealth. The poisoning strategy is optimized in both white-box and black-box scenarios, where the latter relies on surrogate models to generalize the attack to unknown classifiers. The attack is validated on both synthetic and real-world datasets, including COMPAS [42]. Results show that the method can significantly increase disparities between demographic groups without substantially harming overall predictive performance. While the task considered differs from ours, classification rather than T2IR, this work is relevant as it demonstrates that models can be adversarially manipulated to increase demographic disparity without substantially harming accuracy.

Another related yet orthogonal line of work examines how adversarial perturbations influence fairness-aware components within retrieval pipelines. In particular, Ghosh et al. [32] propose an attack that targets the fairness re-ranking module of an image search engine, rather than the retrieval model itself. Their method trains a Generative Adversarial Perturbation (GAP) [43] model that modifies visual features in a way that influences the fairness criteria used during re-ranking. Once these perturbed images are indexed, fairness-aware re-ranking algorithms such as FMMR [44] may assign them to underrepresented demographic groups within the ranking process, thereby promoting them in the final returned list. However, this method differs fundamentally from our setting. While Ghosh et al. manipulate the demographic information consumed by fairness re-rankers, our objective is to study the robustness of the retriever itself and of fairness-oriented mitigation strategies applied at the retrieval. In addition, their attack requires training a GAP model using surrogate demographic inference networks (e.g., DeepFace [45] or FairFace [46]), whereas our method is training-free and operates directly on the similarity function underlying T2IR. Furthermore, their evaluation focuses on broad person-centric queries (i.e., a limited set of search queries containing a generic "person" descriptor), and does not consider attribute-labeled queries that explicitly define demographic attributes. Finally, their experiments do not consider different dataset polarizations nor the interaction between intrinsic and dataset-induced biases, factors that could play a central role in demographic disparities under adversarial retrieval manipulation.

These observations indicate that the adversarial manipulation of fairness in T2IR remains only partially explored, especially in scenarios involving the interaction between retrieval mechanisms, dataset composition, and sensitive attributes. Moreover, existing works mainly highlight the need for mitigation strategies, while systematic evaluations of defenses in this specific setting remain largely unexplored.

2.3. Positioning this work within existing research

The works discussed in Sections 2.1 and 2.2 provide important foundations for studying fairness in VLP-based T2IR systems, while also revealing several key limitations. Overall, three main research gaps emerge.

First, existing studies on fairness in T2IR have primarily focused on *neutral* queries, where the sensitive attribute is not explicitly mentioned. Although this setting is important for evaluating demographic balance in generic retrieval outputs, it does not capture cases in which a query is associated with a specific demographic group. As a result, it limits the ability to assess whether retrieval performance varies across sensitive attributes and to jointly evaluate accuracy and fairness on real-world benchmarks.

Second, unfairness has mainly been attributed to biases encoded in the VLP model, typically inherited from large-scale pretraining data. By contrast, the role of the IRD has received limited attention. In practical T2IR systems, however, the demographic composition of the IRD can directly affect retrieval outcomes, potentially leading to unfair behavior even when the underlying VLP model remains unchanged. This motivates the need to distinguish between model-induced and dataset-induced bias, as well as to analyze their interaction.

Third, most existing analyses consider a *natural setting*, where disparities arise from pre-existing biases in models or data. Consequently, the impact of poisoning attacks on demographic fairness in T2IR systems remains insufficiently understood. Although prior work has shown that fairness mitigation strategies may degrade or fail under adversarial conditions, existing adversarial approaches are either formulated for different tasks, such as classification, or target fairness-aware re-ranking components rather than the VLP-based retrieval process itself. In particular, the possibility that a malicious actor deliberately injects samples into the IRD to induce or amplify demographic disparities remains largely unexplored.

Our work addresses these gaps by proposing a unified framework for studying fairness in VLP-based T2IR under both natural and adversarial settings.

In particular, we introduce *attribute-labeled queries*, which enable fairness evaluation on realistic retrieval benchmarks by associating each query with the sensitive attribute of its ground-truth image. This makes it possible to move beyond neutral queries and systematically analyze disparities across demographic groups.

Within this framework, we define evaluation metrics that jointly capture retrieval accuracy and fairness. Unlike prior approaches, which either require knowledge of the IRD distribution or ignore the demographic attribute associated with the query, the proposed metrics directly measure group-wise performance and quantify disparities through gap-based formulations. This enables a unified evaluation of relevance and fairness across groups. Moreover, we analyze the role of IRD polarization, disentangling biases induced by the retrieval model from those introduced by the composition of the IRD.

Finally, we define a realistic adversarial threat model and introduce `FairLess`, a poisoning attack that manipulates the IRD to induce demographic bias while preserving semantic relevance in the retrieved results. In this way, our work positions fairness in T2IR not only as a problem of bias measurement under benign conditions, but also as a vulnerability that can be actively exploited through adversarial manipulation. Notably, we also evaluate a possible mitigation strategy for this threat model, aiming to reduce the impact of this kind of attack.

3. Preliminaries

In this section, we revisit key preliminary concepts related to VLP-based T2IR systems (see Section 3.1) and (un)fairness (see Section 3.2), which lay the groundwork for the contribution presented later in this paper. In particular, Section 3.3 discusses the problem of (un)fairness in T2IR, highlighting the existing gaps and challenges that our work aims to address.

3.1. VLP-based T2IR systems

T2IR aims to retrieve the most relevant images from an IRD given a query expressed in natural language [47]. To achieve this, modern systems leverage Vision-Language Pretrained models (VLPs), which embed both images and text into a shared semantic space, allowing relevance to be measured through similarity functions. These models are trained on large-scale datasets to capture rich semantic representations across modalities [48]. Images are typically encoded using architectures such as ViT [49] or ResNet [50], while text is processed through transformer-based encoders [51]. The resulting embeddings encapsulate the semantic content of each input. VLPs like CLIP [4] and BLIP-2 [5] process the two modalities independently and align them within a shared embedding space, where similarity metrics (e.g., cosine similarity) are employed to compare and rank results. This cross-modal alignment underpins their strong performance in zero-shot classification and retrieval tasks [4].

More formally, let $\mathcal{D}$ denote the IRD and let q be a natural language query. Each image $v \in \mathcal{D}$ is encoded by an image encoder f_I into a feature vector $f_I(v)$, while the query is encoded by a text encoder f_T into $f_T(q)$. Both encoders map their respective inputs into a shared embedding space, where similarity $\phi(f_I(v), f_T(q))$, typically measured using cosine similarity [4], is used to rank images according to their relevance to the query. Selecting the top-k images enables the system to return multiple plausible results, thereby increasing robustness to ambiguity and variation in natural language queries [52,53].

3.2. Algorithmic (Un)fairness

Algorithmic (un)fairness arises when an algorithm's behavior is influenced by sensitive attributes such as gender, race, or ethnicity, leading to disparate outcomes across demographic groups [54]. For instance, a hiring algorithm deployed by a major technology company was found to systematically disadvantage female candidates, particularly for software engineering positions [55]. Such disparities are concerning because they can perpetuate or amplify existing societal biases, restrict opportunities for underrepresented groups, and erode trust in automated decision-making systems.

To address such disparities, the problem of (un)fairness has been extensively studied in the literature, particularly in the context of classification tasks. In this setting, dedicated metrics have been proposed to quantify differences in model behavior across demographic groups [56–58]. Among these metrics, two of the most widely adopted are *Demographic Parity* (DP) and *Equal Opportunity* (EO). We formally define DP and EO in the setting of binary classification with a binary sensitive attribute, though the definitions can be extended to multiclass or regression tasks, as well as to general categorical and numerical sensitive attributes. Let $S(X) \in \{a, b\}$ denote the value of a binary sensitive attribute[1] (e.g., gender or race), where a and b represent the two demographic groups under consideration. Let $X \in \mathcal{X}$ be the observable features of an individual, $Y \in \mathcal{Y} = \{\pm 1\}$ the true outcome, and $\hat{Y} = \text{sign}(f(X))$ the predicted label produced a function (or model) $f : \mathcal{X} \to \mathbb{R}$. DP requires the probability of a positive prediction to be equal across groups, regardless of the true outcome:

$$\mathbb{P}\{\hat{Y} = 1 \mid S(X) = a\} = \mathbb{P}\{\hat{Y} = 1 \mid S(X) = b\}. \tag{1}$$

This criterion aims to ensure equal treatment across demographic groups and to prevent disparate impact, where one group is systematically favored over another. EO, on the other hand, focuses on individuals for whom the positive outcome is appropriate. It requires the probability of receiving a positive prediction, conditioned on the true outcome being positive, to be equal across groups:

$$\mathbb{P}\{\hat{Y} = 1 \mid Y = 1, S(X) = a\} = \mathbb{P}\{\hat{Y} = 1 \mid Y = 1, S(X) = b\}. \tag{2}$$

This ensures that all qualified individuals, regardless of group membership, have equal access to positive outcomes. In recent years, a substantial body of work has focused on analyzing the unfairness of prediction models, leading to the development of a wide range of mitigation techniques. Unfairness in classification can stem from biased data, biased learning processes, or the amplification of existing biases by the model itself. As a result, significant effort has been devoted to understanding these phenomena and developing appropriate metrics to evaluate fairness [56–58]. Then, numerous mitigation strategies have been proposed, which are typically categorized into three main families: pre-processing, in-processing, and post-processing methods [56–58]. Pre-processing techniques aim to mitigate historical biases by transforming the training data prior to learning, after which standard ML models are applied. In-processing techniques introduce fairness constraints directly into the learning phase, modifying the model's internal mechanisms to ensure fair outcomes. Post-processing techniques adjust the outputs of already trained models to improve fairness without altering the underlying training process. Furthermore, researchers have begun to explore the intersection between fairness and other trustworthiness criteria [56], including robustness [59] and regressivity [54].

[1] Note that this is a simplification, imposed by data availability constraints, as we will discuss in Section 7.2. It allows us to address the problem of algorithmic fairness mathematically without overcomplicating the notation. In a real-world application, a minimally realistic assumption would be to consider a multivalued sensitive attribute, as we will discuss later in the paper.

3.3. (Un)fairness in T2IR

When dealing with T2IR systems, adapting classical (un)fairness metrics such as DP or EO is not straightforward. Recent studies on fairness in VLP-based T2IR systems have proposed retrieval-specific metrics to assess disparities across demographic groups, often by adapting definitions originally developed for classification tasks. One such adaptation is the *DDP* metric [8], based on the fairness formulation in [60], which quantifies group imbalances by comparing the proportion of retrieved images across sensitive attributes with the corresponding distribution in the IRD. However, these approaches typically assume full knowledge of the IRD's demographic composition, i.e., the number of images belonging to each group. This assumption rarely holds in real-world scenarios, where IRDs are often distributed, proprietary, or opaque. Moreover, even if the IRD is biased, T2IR systems are expected to behave fairly regardless of the underlying group distribution. In addition to *DDP*, the authors of [8] employ the *Skew@k* metric introduced in [61], which relaxes the requirement of knowing the full dataset composition. Instead, it evaluates fairness by comparing the proportion of retrieved images from each group against a predefined target distribution (typically uniform), and measures the largest deviation using the maximum absolute log-ratio between observed and expected frequencies. While *Skew@k* is more suitable for real-world applications where the underlying IRD is unknown, it introduces a different limitation: it is agnostic to the demographic group referenced in the query itself. In other words, it treats all queries uniformly, regardless of whether they refer to specific groups (e.g., "male" or "female"), and thus may fail to capture disparities in retrieval quality that disproportionately affect certain groups.

As in the classical supervised learning setting, biases in T2IR systems can originate from the data, either the training data used to learn the VLP or the IRD itself, or from the VLP models, which may learn and amplify existing biases in the training data [8]. To address this, researchers have proposed mitigation strategies for T2IR that align with the three established families of fairness techniques: pre-processing, in-processing, and post-processing methods [8], as discussed in Section 3.2. However, what remains largely unexplored is the relationship between fairness in T2IR systems and other important metrics, such as robustness (see Section 2).

Beyond metric definitions, (un)fairness in T2IR is typically evaluated through empirical studies on controlled datasets. These datasets are often artificially constructed by combining images labeled with sensitive demographic information and designing queries that explicitly refer to those groups. This process often starts from image collections such as FairFace or UTKFace, where each image is labeled with specific demographic information, such as racial categories (e.g., "East Asian", "White", etc.), followed by the creation of queries designed to target specific subgroups. In this controlled setting, where queries are designed to match entire demographic groups, fairness is evaluated based on whether the retrieved results reflect balanced group representation, since multiple valid images may correspond to the same query. However, this setup does not reflect the complexity of more realistic T2IR benchmarks such as MSCOCO or Flickr30K. In these datasets, each query is typically linked to a single ground-truth image, resulting in a one-to-one correspondence. As a result, fairness analysis becomes significantly more challenging, as it must account not only for the demographic composition of the retrieved images but also for whether the correct image appears among the top-k results. To structure fairness analysis in such settings, queries in these benchmarks can generally be divided into two categories: (i) *Neutral queries*, which do not mention any sensitive attribute (e.g., "CEO of a big company"), and (ii) *Attribute-specific queries*, which explicitly refer to a sensitive attribute (e.g., "female CEO of a big company"). Although real-world benchmarks more faithfully reflect the nature of the T2IR task, artificially constructed datasets are often preferred for (un)fairness analysis because real benchmarks do not offer explicit annotations of sensitive attributes. Indeed, while *attribute-specific queries* can be mapped to demographic groups based on the presence of explicit indicators in the text [35], *neutral queries* provide no such information. However, to overcome this limitation, one can leverage existing demographic annotations of the ground-truth images (e.g., gender or race labels). This enables the definition of a third query category: (iii) *Attribute-labeled queries*, which are indirectly associated with a sensitive group through the label of their ground-truth image. This categorization facilitates fairness evaluation in realistic benchmarks where sensitive attribute labels are not directly available. By associating each query with a demographic group, disparities in retrieval performance across groups can be quantified. This, in turn, enables fairness metrics to be meaningfully applied alongside standard retrieval metrics, capturing both relevance (i.e., whether the correct image appears among the top-k) and demographic parity.

4. (Un)fairness settings and metrics

In this section, we define the evaluation settings for assessing (un)fairness (see Section 4.1) and introduce a set of metrics for measuring demographic disparities (see Section 4.2) in T2IR systems. Both settings and metrics are defined under the assumption that each query belongs to the class of *attribute-labeled queries*, meaning that it is associated with a label for the sensitive attribute under investigation. This allows the evaluation of (un)fairness in VLPs-based T2IR using realistic datasets such as MSCOCO.

4.1. Natural and adversarial setting

In the *natural setting*, the goal is to evaluate how model-intrinsic biases and IRD D biases may lead to unfair outcomes across subgroups.

For this purpose, we construct multiple versions of the IRD D, each characterized by a different degree of demographic imbalance. Specifically, we define a balanced dataset in which images are equally distributed between groups a and b (i.e., $D_{50\%}$, where 50% belong to group a and 50% to group b). This setup isolates the model's influence by removing skew in the data distribution. Conversely, to study the effect of polarization, we construct polarized datasets D_p in which one group is underrepresented (e.g., $D_{30\%}$ where 30% belongs to group a and 70% to group b). We can then analyze the system's behavior across groups by comparing retrieval

accuracy and group representation in the top-k results. By comparing retrieval behavior across these configurations, we can assess whether the observed disparities arise from the model, the data, or their interaction.

In addition, we consider an *adversarial setting* in which an attacker actively manipulates the IRD D by injecting one or more poisoned images with the goal of influencing retrieval outcomes. Specifically, the attacker aims to increase the retrieval frequency of images belonging to a target group, while ensuring that the retrieved content remains semantically aligned with the search query. This allows the attacker to promote biased representations without compromising relevance (see Section 5).

4.2. Accuracy and (Un)fairness metrics

In this section, we introduce a set of accuracy and (un)fairness metrics for VLP-based T2IR systems designed to address the key limitations identified in Section 3. To this end, we employ state-of-the-art accuracy metrics and adapt the generalized fairness formulation proposed by Franco et al. [62], originally introduced for classification tasks, to the context of VLP-based T2IR systems. This formulation is capable of generalizing most fairness definitions in classification, including DP, EO, and Equal Odds.

For binary classification, the generalized fairness $F(f)$ for a function f is defined as:

$$F(f) = \frac{1}{g} \sum_{\breve{\mathcal{Y}} \in \{\breve{\mathcal{Y}}_1, \breve{\mathcal{Y}}_2, \ldots, \breve{\mathcal{Y}}_g\}} \Big| \mathbb{E}_{(X,Y)|S(X)=a, Y \in \breve{\mathcal{Y}}} \{\ell(f(X), \cdot)\} - \mathbb{E}_{(X,Y)|S(X)=b, Y \in \breve{\mathcal{Y}}} \{\ell(f(X), \cdot)\} \Big|, \tag{3}$$

where $\ell : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}$ is a prescribed loss function that quantifies the discrepancy between the model score and the true label Y; $\breve{\mathcal{Y}}$, $\breve{\mathcal{Y}}_1, \breve{\mathcal{Y}}_2, \cdots, \breve{\mathcal{Y}}_g \subseteq \mathcal{Y}$; and $\cdot \in \{y, \pm 1\}$.

By setting $\{\breve{\mathcal{Y}}_1, \breve{\mathcal{Y}}_2, \ldots, \breve{\mathcal{Y}}_g\} = \breve{\mathcal{Y}}_1 = \mathcal{Y}$, $\cdot = 1$, and $\ell(f(X), Y) = [Y \neq \hat{Y}]$ in $F(f)$ we obtain the DP [63]:

$$F(f) = \Big| \mathbb{P}\{\hat{Y} = 1 | S(X) = a\} - \mathbb{P}\{\hat{Y} = 1 | S(X) = b\} \Big|. \tag{4}$$

By setting $\{\breve{\mathcal{Y}}_1, \breve{\mathcal{Y}}_2, \ldots, \breve{\mathcal{Y}}_g\} = \breve{\mathcal{Y}}_1 = \{+1\}$, $\cdot = 1$, and $\ell(f(X), Y) = [Y \neq \hat{Y}]$, we obtain the EO [64]:

$$F(f) = \Big| \mathbb{P}\{\hat{Y} = 1 | Y = 1, S(X) = a\} - \mathbb{P}\{\hat{Y} = 1 | Y = 1, S(X) = b\} \Big|. \tag{5}$$

Finally, by setting $\{\breve{\mathcal{Y}}_1, \breve{\mathcal{Y}}_2, \ldots, \breve{\mathcal{Y}}_g\} = \{\{-1\}, \{+1\}\}$, $\cdot = Y$, and $\ell(f(X), y) = [Y \neq \hat{Y}]$, we obtain the Equal Odds [64]:

$$F(f) = \frac{1}{2} \sum_{\breve{Y} \in \{\pm 1\}} \Big| \mathbb{P}\{\hat{Y} = \breve{y} | Y = \breve{Y}, S(X) = a\} - \mathbb{P}\{\hat{Y} = \breve{y} | Y = \breve{Y}, S(X) = b\} \Big|. \tag{6}$$

It is important to note that T2IR is a ranking task rather than a binary classification task. For this reason, classical classification-based fairness metrics cannot be transferred directly to the T2IR setting. Instead, we adopt the general principle of group disparity that underlies these metrics and instantiate it through retrieval-specific quantities computed over the top-k ranked results. In this sense, the proposed metrics quantify disparities across demographic groups in ranking-dependent outcomes, including the occurrence of the ground-truth image, the attribute consistency of the retrieved images, and, in adversarial settings, the occurrence of poisoned samples among the top-k results.

We therefore adapt $F(f)$ to VLP-based T2IR systems. The main difference between classification and VLP-based T2IR systems lies in the definition of the loss function. In particular, for T2IR systems, the loss depends on whether specific results are present or absent in the retrieval output.

For T2IR, let $q \in \mathcal{Q}$ be a query from a set of possible queries, and let the IRD be $D' := D \cup \mathcal{P}$, where $\mathcal{P}$ is a set of poisoned images (empty in the *natural setting* and non-empty in the *adversarial setting*). Let $S(q)$ and $S(v)$ denote the binary sensitive group, a or b, of a query q and an image $v \in D'$, respectively. The ground-truth image associated with a query q is defined as $v^\star(q)$. Finally, let $\mathrm{Top}_k(q; D') = (v_1, \ldots, v_k)$ denote the ordered list of the top-k images retrieved for q from D'.

Based on these definitions, we first introduce accuracy and fairness metrics applicable to both *natural* and *adversarial* settings, as they do not depend on whether the IRD includes poisoned images. With regard to accuracy, we employ the Top-k Accuracy (ACC), which measures the probability that the ground-truth image is retrieved among the top-k results:

$$\mathrm{ACC@}k = \mathbb{P}\{v^\star(q) \in \mathrm{Top}_k(q; D')\}. \tag{7}$$

Concerning fairness, and inspired by $F(f)$, we first define two quantities analogous to the loss terms in $F(f)$. The first is the ACC@k conditioned on the sensitive group:

$$\mathrm{ACC}_g\mathrm{@}k = \mathbb{P}\left\{v^\star(q) \in \mathrm{Top}_k(q; D') | S(q) = g\right\}. \tag{8}$$

The second is the Attribute-Match Rate (AMR), which measures the likelihood that a retrieved image shares the same sensitive attribute as the query:

$$\mathrm{AMR}_g\mathrm{@}k = \mathbb{E}\left\{ \frac{1}{k} \sum_{v \in \mathrm{Top}_k(q; D')} S(v) = S(q) \Big| S(q) = g \right\}. \tag{9}$$

Then, following the idea behind $F(f)$, we compute the gaps between subgroups, which yield two fairness metrics[2]:

$$\Delta\text{ACC@}k = \left|\text{ACC}_a\text{@}k - \text{ACC}_b\text{@}k\right|, \tag{10}$$

$$\Delta\text{AMR@}k = \left|\text{AMR}_a\text{@}k - \text{AMR}_b\text{@}k\right|. \tag{11}$$

In the *adversarial setting*, we define additional metrics applicable only when the IRD includes poisoned images ($\mathcal{P} \neq \emptyset$). As concerns accuracy, we use the Attack Success Rate (ASR), which measures the probability that at least one poisoned image appears among the top-k retrieved results:

$$\text{ASR@}k = \mathbb{P}\{\text{Top}_k(q; \mathcal{D}') \cap \mathcal{P} \neq \emptyset\}. \tag{12}$$

With respect to fairness, analogously to the *natural setting*, we first define ASR@k conditioned on the sensitive group:

$$\text{ASR}_g\text{@}k = \mathbb{P}\{\text{Top}_k(q; \mathcal{D}') \cap \mathcal{P} \neq \emptyset | S(q) = g\}, \tag{13}$$

and then, we compute the gaps between subgroups, leading to additional fairness metrics[3]:

$$\Delta\text{ASR@}k = \left|\text{ASR}_a\text{@}k - \text{ASR}_b\text{@}k\right|. \tag{14}$$

The proposed metrics do not require knowledge of the IRD distribution, since they are computed solely from attribute-labeled queries and retrieved results. The controlled IRDs used in our experiments serve only to isolate the effect of demographic imbalance under well-defined conditions.

Applying these metrics nevertheless requires demographic annotations with a reasonable degree of reliability for the ground-truth images and, when measuring attribute consistency, for the retrieved images. As in standard group-based fairness evaluation [63,64], annotation noise may affect the estimated disparities, especially when it is asymmetric or group-dependent. A systematic analysis of this effect would require a dedicated study involving explicit noise models or additional verified annotations [66,67]; it is therefore beyond the scope of the present work, while remaining an interesting direction for future investigation.

5. Threat model

In this section, we present the threat model assumed for the *adversarial setting*, detailing the attacker's knowledge and capabilities (see Section 5.1) and the attacker's goal (see Section 5.2).

5.1. Attacker's knowledge and capabilities

We consider a white-box scenario in which the attacker has full knowledge of the retrieval pipeline, including the underlying VLP, the image and text encoders, and the similarity function, as described in Section 3.1. This reflects realistic scenarios where retrieval systems are built on publicly available models such as CLIP [4], making internal components accessible rather than developed in-house. Indeed, VLPs are rarely trained from scratch by end users, which explains why a significant body of prior work on adversarial attacks adopts the white-box setting [10,11].

In line with classical poisoning assumptions [68], we further assume that the attacker can modify a (small) subset of the IRD. This is a plausible condition, as image collections are often sourced from open or semi-trusted repositories [69,70]. In practice, this assumption is realistic in scenarios where IRDs are constructed from open or semi-curated sources, such as web-scale image collections, user-contributed content platforms, or continuously updated repositories [71]. In such settings, the insertion of a limited number of carefully crafted samples represents a plausible attack vector, even in the presence of basic moderation or filtering mechanisms.

Additionally, the attacker has access to a small external dataset, denoted as $\mathcal{E}$, which is annotated with the same binary sensitive attribute under investigation, such as gender or race. This auxiliary dataset can be independently curated by the attacker using publicly available data sources, allowing it to be aligned with the specific semantic concepts that the attacker aims to manipulate. As a result, such a dataset can be effective in practice without requiring access to the target system's private data [71].

To reflect realistic deployment conditions, we do not assume that the attacker has access to the exact user query submitted at test time, denoted as qu. In practice, queries are unpredictable and may vary across semantically equivalent formulations, making perfect knowledge of qu an overly strong assumption. Instead, we adopt a more realistic setting in which the attacker selects a target query qt that captures the semantic concept they intend to manipulate. The actual user query qu is then unknown but semantically equivalent to qt, reflecting natural linguistic variability [72].

In practical deployments, retrieval systems may include proprietary components, such as task-specific similarity functions or additional scoring layers, which can restrict the attacker's knowledge to a gray-box or even black-box setting. Nevertheless, we

[2] The extension to multivalued sensitive attributes, following [65], considers the case where $S(X) \in \mathcal{G}$ with $|\mathcal{G}| > 2$. It can be formalized as follows: $\Delta\text{ACC@}k = \frac{1}{|\mathcal{G}|}\sum_{s_1\in\mathcal{G}}\left|\frac{1}{|\mathcal{G}|}\sum_{s_2\in\mathcal{G}}\text{ACC}_{s_2}\text{@}k - \text{ACC}_{s_1}\text{@}k\right|$, and $\Delta\text{AMR@}k = \frac{1}{|\mathcal{G}|}\sum_{s_1\in\mathcal{G}}\left|\frac{1}{|\mathcal{G}|}\sum_{s_2\in\mathcal{G}}\text{AMR}_{s_2}\text{@}k - \text{AMR}_{s_1}\text{@}k\right|$. Note that, for $|\mathcal{G}| = 2$, corresponding to a binary sensitive attribute, this definition reduces to the one used in our paper.

[3] The extension to multivalued sensitive attributes, following [65], considers the case where $S(X) \in \mathcal{G}$ with $|\mathcal{G}| > 2$. It can be formalized as follows: $\Delta\text{ASR@}k = \frac{1}{|\mathcal{G}|}\sum_{s_1\in\mathcal{G}}\left|\frac{1}{|\mathcal{G}|}\sum_{s_2\in\mathcal{G}}\text{ASR}_{s_2}\text{@}k - \text{ASR}_{s_1}\text{@}k\right|$. Note that, for $|\mathcal{G}| = 2$, corresponding to a binary sensitive attribute, this definition reduces to the one used in our paper.

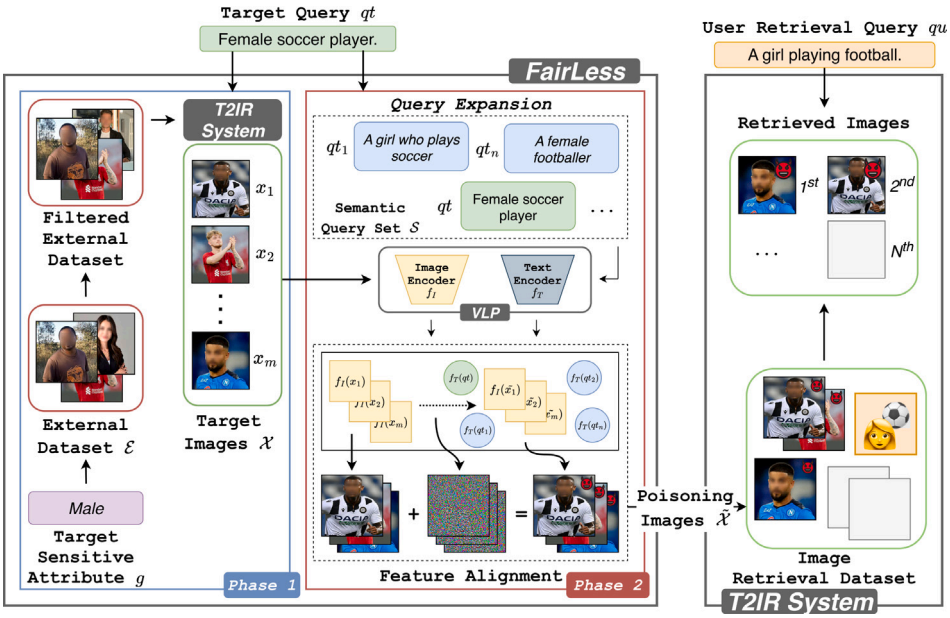

Fig. 1. FairLess workflow.

emphasize that the white-box setting remains important for assessing worst-case vulnerabilities and for providing an upper bound on the effectiveness of the attack. Reducing the attacker's knowledge may make attacks more complex, but it may also create a false sense of security; security by obscurity should therefore be avoided whenever possible.

Nevertheless, to investigate a scenario that is more realistic than the white-box setting, we consider the case in which the attacker has only partial knowledge of the similarity function ϕ used in the system. This is particularly realistic, since while the choice of the underlying VLP is often fixed to a widely available architecture, the similarity function can be easily tuned or replaced in development and deployment. Accordingly, we examine a scenario where the attacker relies on cosine similarity to craft poisoned samples, while the actual retrieval system may employ a different measure, such as the ITM score [5], which remains unknown to the attacker. This setting can be interpreted as a form of gray-box knowledge, in which the attacker has access to the main components of the system but not to all implementation details.

5.2. Attacker's goal

The attacker aims to inject one or more poisoned images into the IRD D that are semantically aligned with the target query qt, but deliberately associated with a specific sensitive attribute. For example, to bias retrieval results for a query such as "a female soccer player", the attacker may insert visually consistent images depicting male individuals. These poisoned samples increase the likelihood that misrepresentative content is returned in response to qu, thereby inducing biased behavior in the system.

Importantly, the attacker's objective depends on whether a single or multiple poisoned images are injected into the IRD D. In the restricted case where only a single poisoned item is injected, the goal is not to entirely disrupt retrieval but rather to highlight the disproportionate influence of a single injection. Even if the correct result remains within the top-k, the mere presence of an adversarial item can undermine system reliability and safety. For instance, in automated pipelines such as Retrieval-Augmented Generation [73], the incorporation of a poisoned result into the generation process has been shown to directly alter downstream outputs [74]. By contrast, when the attacker can inject multiple poisoned images, the objective shifts to concealing the correct outcome from the top-k results, thereby maximizing the bias induced in the user toward a particular subgroup.

6. FairLess: A fairness attack for T2IR

FairLess is designed to amplify existing biases in VLP-based T2IR systems while maintaining semantically relevant retrieval results, i.e., ensuring that the retrieved outputs remain consistent with the user's intended content. For example, FairLess aims to bias the retrieval results for a qt such as "a female soccer player", and for any qu semantically similar to qt, such as "a girl playing football", by returning one or more poisoned images of male soccer players (see Fig. 1). To achieve this, FairLess operates in two main phases.

In Phase 1 (Section 6.1), referred to as *Semantic Targeted Search* (STS), the adversary selects from $\mathcal{E}$ the top-m images $\mathcal{X} = \{x_1, \ldots, x_m\}$ related to qt using the target T2IR system. These are chosen to match the sensitive attribute that the adversary intends

to bias g. In the running example, the adversary retrieves the top-m images labeled as "male". Note that the richer the pool $\mathcal{E}$, the more semantically consistent the top-m images will be with qt, except for the targeted sensitive attribute.

In Phase 2 (Section 6.2), inspired by the Collisio framework [75], a set of n semantically related queries $\mathcal{S} = \{qt_1, \dots, qt_n\}$ is generated starting from qt. Each selected image $x \in \mathcal{X}$ is then adversarially perturbed to maximize its similarity with the queries in $\mathcal{S}$, producing the poisoned set $\tilde{\mathcal{X}} = \{\tilde{x}_1, \dots, \tilde{x}_m\}$. This procedure ensures that the poisoned images remain aligned not only with qt but also with its broader semantic context. Concretely, each query in $\mathcal{S}$ is encoded by the text encoder f_T, and the perturbation for each x_i is optimized so that its embedding $f_I(\tilde{x})$ is closely aligned with all corresponding text embeddings under the similarity function ϕ. The perturbation is constrained within an ε-radius under a chosen ℓ_p norm (e.g., ℓ_∞ or ℓ_2). Finally, the poisoned images $\tilde{\mathcal{X}}$ are injected into the IRD $\mathcal{D}$, with the goal of making them appear among the top retrieval results for any user query qu semantically related to qt.

Note that, in Section 6.3, we also discuss the problem of mitigating this kind of attack and propose a mitigation strategy based on input preprocessing via JPEG compression [33].

6.1. Phase 1: Semantic Targeted Search (STS)

The STS phase aims to select $\mathcal{X}$ from an external dataset $\mathcal{E}$ accessible to the attacker, using the targeted T2IR system. The selected images are semantically aligned with the intended retrieval concept (i.e., they represent the target query) while deliberately belonging to a specific target class g of the sensitive attribute chosen by the attacker. Formally:

$$\mathcal{X} = \mathrm{Top}_m(qt; \{v \in \mathcal{E} \mid S(v) = g\}), \tag{15}$$

or, in a more explicit iterative form:

$$\begin{aligned} &\mathcal{X}_0 = \varnothing, \\ &\mathcal{X}_i = \mathcal{X}_{i-1} \cup \underset{v \in \mathcal{E} \setminus \mathcal{X}_{i-1},\, S(v)=g}{\arg\max} \phi\big(f_T(qt), f_I(v)\big), \\ &\mathcal{X} = \mathcal{X}_m, \end{aligned} \tag{16}$$

where ϕ denotes the similarity function between the text embedding $f_T(qt)$ and the image embedding $f_I(v)$. This formulation makes explicit the selection mechanism typically employed in T2IR systems.

As also discussed in Section 5.1, $\mathcal{E}$ does not need to match the IRD distribution; rather, it only needs to provide sufficient coverage of the targeted semantic concepts, which are defined by the attacker and are therefore fully known. Consequently, the effectiveness of the attack depends on the availability of semantically aligned samples in $\mathcal{E}$. However, such samples can typically be collected from public repositories, since the relevant semantic criteria are specified by the attacker.

6.2. Phase 2: Feature alignment

Having selected $\mathcal{X}$, the aim of Phase 2 of `FairLess` is to manipulate them in order to craft the poisoned samples $\tilde{x} \in \tilde{\mathcal{X}}$ to inject into $\mathcal{D}$ of the T2IR system.

The first step is to generate $\mathcal{S}$. This can be done with different strategies, e.g., using a set of known variations present in a publicly available dataset or using an LLM (see Section 7).

Subsequently, $\mathcal{S}$ is leveraged to generate each poisoned sample $\tilde{x} \in \tilde{\mathcal{X}}$ from the target image $x \in \mathcal{X}$. This is achieved by perturbing x_i so that the representation $f_I(\tilde{x})$ of the poisoned image is maximally aligned with $f_T(q)$ for all $q \in \mathcal{S}$. Formally, $\tilde{x}$ is obtained by solving the following optimization problem:

$$\tilde{x} = \underset{\hat{x} \in \mathcal{B}^p_\varepsilon(x)}{\arg\max} \sum_{q \in \mathcal{S}} \phi\left(f_T(q), f_I(\hat{x})\right), \tag{17}$$

where $\mathcal{B}^p_\varepsilon(x)$ denotes the set of allowable perturbations of x_i within an ε-bounded ℓ_p ball, $\mathcal{B}^p_\varepsilon(x) = \{\hat{x} : \|x - \hat{x}\|_p \le \varepsilon\}$. In our setting, we consider both the ℓ_∞ and ℓ_2 norms, although more sophisticated perturbation sets can also be employed [76]. Once all the poisoned images $\tilde{\mathcal{X}}$ are crafted, they are injected into the IRD $\mathcal{D}$. As a result, when a retrieval is performed with the user query qu, the attack increases the likelihood that the poisoned samples appear among the top-ranked results.

We note that Problem (17) corresponds to a classical adversarial attack formulation on images [77]. Although generally non-convex, the objective is differentiable and can therefore be effectively addressed using gradient-based optimization methods [11,24]. To this end, Algorithm 1 provides the pseudocode for Phase 2 of `FairLess`, formulated as a projected gradient ascent procedure inspired by the classical PGD framework [24], and designed to solve Problem (17). It requires the semantic query set $\mathcal{S}$, the encoders f_I and f_T, the similarity function ϕ, the target images $\mathcal{X}$, the chosen ℓ_p norm, and the parameters K and ε, which are set (see Section 7). Specifically, for each image x, the method performs K iterations (Line 5) of gradient ascent with step size $\eta = \varepsilon/K$ (Line 4). At each step, the gradient g of the alignment objective is computed (Line 6) and used to update $\tilde{x}$: by the sign of g for $p = \infty$ or by its normalized direction for $p = 2$, where the small constant c prevents division by zero. After each update, $\tilde{x}$ is projected onto the ε-ball around x (Line 12) and clipped to the valid pixel range (Line 13). In the end, the crafted poisoned samples $\tilde{\mathcal{X}}$ are returned for injection into the IRD $\mathcal{D}$.

Finally, we briefly discuss the computational complexity of `FairLess`. The computational cost of `FairLess` is dominated by the iterative optimization procedure used to generate poisoned samples, whereas Phase 1 only requires standard retrieval operations.

Algorithm 1: Phase 2 of `FairLess`: *Feature Alignment*

Input: S, semantic query set; f_I, image encoder; f_T, text encoder; ϕ, similarity function; $\mathcal{X}$, target images; K, number of steps; ε, perturbation bound; and $p \in \{2, \infty\}$, norm type.
Output: $\tilde{\mathcal{X}}$, poisoned images.

$\tilde{\mathcal{X}} = \emptyset$
for $x \in \mathcal{X}$ **do**
  $\tilde{x} = x$
  $\eta = \frac{\varepsilon}{K}$
  for $k \in \{1, \dots, K\}$ **do**
    $g = \nabla_{\tilde{x}} \left[\sum_{q \in S} \phi(f_T(q), f_I(\tilde{x})) \right]$
    if $p = \infty$ **then**
      $\tilde{x} = \tilde{x} + \eta \cdot \text{sign}(g)$
    else if $p = 2$ **then**
      $\tilde{x} = \tilde{x} + \eta \cdot \frac{g}{\|g\|_2 + c}$
    end
    $\tilde{x} = \Pi_{\mathcal{B}_\varepsilon^{(p)}(x)}(\tilde{x})$
    $\tilde{x} = \text{clip}(\tilde{x})$
  end
  $\tilde{\mathcal{X}} = \tilde{\mathcal{X}} \cup \{\tilde{x}\}$
end
return $\tilde{\mathcal{X}}$

In particular, for each selected image, the attack performs a fixed number of gradient-based update steps, each involving standard forward and backward passes through the model. The overall complexity therefore scales linearly with the number of optimization steps, the number of selected images, and the size of the semantic query set S. The contribution of S is limited to a linear factor; in practice, $|S|$ is small and typically fixed to a few values, such as 1, 3, or 10, thus having only a limited impact on the overall cost.

6.3. A possible mitigation

Mitigating poisoning attacks in retrieval systems remains challenging, as existing defenses often require adversarial fine-tuning of the underlying model [78,79], modifications to the retrieval pipeline, such as image–text rematching [80], or purification- and reconstruction-based preprocessing [81]. While these strategies can be effective in specific settings, they typically introduce additional computational or architectural requirements, which limits their suitability as lightweight plug-in defenses.

In this work, we investigate a lightweight input-preprocessing mitigation based on JPEG compression [33]. This choice is motivated by the observation that adversarial perturbations often exploit small, high-frequency image variations, which lossy compression may partially suppress. Let $j_c(\cdot)$ denote JPEG compression with quality factor c, where smaller values of c correspond to stronger compression. Using the T2IR notation introduced in Section 3.1, we replace the standard retrieval score $\phi(f_I(v), f_T(q))$ with

$$\phi\left(f_I(j_c(v)), f_T(q)\right), \tag{18}$$

for each image $v \in D$ and query q. Thus, compression is applied to each image before feature extraction, while the image encoder, text encoder, and similarity function are left unchanged. Our goal is not to propose JPEG compression as a complete defense, but rather to assess whether a simple preprocessing step can reduce the effectiveness of `FairLess`. We further note that an adaptive attacker aware of the compression step could optimize poisoned samples through the same preprocessing pipeline, potentially reducing the effectiveness of this mitigation [82]. We leave a systematic analysis of adaptive defenses to future work.

7. Experimental setting

In this section, we present the architectures underlying the VLP-based T2IR systems, describe the datasets employed in our study, and detail the experimental setups used to evaluate (un)fairness under the two proposed scenarios: the *natural* and the *adversarial* setting. We design a set of controlled experiments to isolate the contributions of the different components of the proposed framework, allowing us to assess their individual and combined effects. We publicly release the source code required to replicate the experiments.[4] All results reported in this work are obtained by the authors through independent experiments and are not taken from previous publications. While we rely on publicly available models and datasets, all evaluations are conducted within our experimental framework.

[4] https://github.com/lazzd/FairLess.

7.1. Architectures of the VLP-based T2IR systems

In this work, we consider T2IR systems that employ CLIP- and BLIP-2-based VLPs.

For the CLIP family, we include both standard and debiased variants. Among the standard models, we adopt two versions introduced in [4]: one using a Vision Transformer backbone (ViT-B/16), denoted CLIP_{ViT}, and another relying on a ResNet-101 convolutional network, denoted CLIP_{CNN}. For the debiased variants of CLIP, specifically designed to mitigate demographic bias, we consider CLIP-DEB [17] and FairCLIP [37]. For CLIP-DEB, we rely on the officially released code,[5] which fine-tunes CLIP with a ViT-B/16 backbone on the FairFace dataset to produce two specialized models: one mitigating gender bias ($\text{CLIP-DEB}_{\text{G}}$) and another mitigating racial bias ($\text{CLIP-DEB}_{\text{R}}$). For FairCLIP, we use the released ViT-B/16 checkpoints,[6] which similarly provide two debiased models targeting gender ($\text{FairCLIP}_{\text{G}}$) and race ($\text{FairCLIP}_{\text{R}}$). All CLIP-based T2IR systems rely on cosine similarity ϕ.

For the BLIP family, we consider BLIP-2 [5], a state-of-the-art VLP evaluated in its text–image retrieval configuration, as described in the original work. The BLIP-based T2IR systems can operate with either cosine similarity (BLIP-2) or the Image–Text Matching (ITM) score ($\text{BLIP-2}_{\text{ITM}}$). The ability to employ different similarity functions ϕ in BLIP-based T2IR is necessary to model a more realistic scenario in which we have only partial knowledge of the systems, i.e., we know the T2IR architecture but not the specific similarity function. In particular, in the *adversarial setting* with $\text{BLIP-2}_{\text{ITM}}$, `FairLess` is optimized using cosine similarity, while the T2IR system is evaluated with the ITM score.

7.2. Datasets

We conduct our experiments on the MSCOCO [34] dataset, a widely used benchmark for image-caption retrieval comprising 123,287 images, each paired with up to 6 captions. To enable fairness analysis, we augment MSCOCO with demographic annotations for both gender and race, provided by prior work, which we study independently. We refer to these as two annotation settings, and all experiments are conducted separately within each setting.

For gender, we adopt the annotations of Zhao et al. [35], labeling MSCOCO images as either "male" or "female". The resulting annotated pool comprises 10,923 images, of which 7771 are labeled as male and 3152 as female. We use this set of 10,923 images as $\mathcal{E}$ in Phase 1 of `FairLess`. From the 10,923, we sample subsets of 3000 images to construct IRDs with different gender polarizations (e.g., $\mathcal{D}_{70\%}$, consisting of 70% male and 30% female images). We choose 3000 samples to ensure IRDs are sufficiently large (thus making the retrieval task challenging) while still allowing us to create strongly polarized distributions given the available annotations. Note that not all 3000 images have captions, since we rely on the Karpathy split [83], which provides captions for only 1317 images for which we have labels (923 male and 394 female). We use these captioned images to derive the queries employed in retrieval.

For race, we focus on images depicting a single prominent individual, following the demographic annotations of Zhao et al. [36]. The resulting annotated pool comprises 6577 images, of which 5894 are labeled as "light-skinned" and 683 as "dark-skinned". We use this set as $\mathcal{E}$. Because the data are highly imbalanced, the maximum IRD size is 975, which ensures we can construct polarized IRDs analogous to the gender case (e.g., 70% dark-skinned and 30% light-skinned). Unlike the gender annotations, all 975 images have up to 6 associated captions.

In all experiments, we randomly sample 500 queries (i.e., captions), equally divided between the two demographic groups (250 per group), regardless of IRD polarization. Queries may include captions corresponding to different images or multiple captions from the same image. The sampling process is repeated three times with different random seeds to ensure robustness of results, and the same sampling procedure is consistently applied across all evaluated methods.

7.3. Natural setting

In this scenario, we evaluate the retrieval performance of T2IR systems on IRDs with varying demographic compositions, using the accuracy and fairness metrics defined for the *natural setting* in Section 4.2. Specifically, we test each system on a balanced IRD, $\mathcal{D}_{50\%}$, where both demographic groups are equally represented, as well as on polarized IRDs, $\mathcal{D}_{30\%}$ and $\mathcal{D}_{70\%}$, which exhibit skewed distributions favoring one group. The balanced distribution $\mathcal{D}_{50\%}$ is used to assess the intrinsic bias of the VLP models.

7.4. Adversarial setting

In this scenario, we evaluate the robustness of T2IR systems against the proposed `FairLess` (see Section 6). Similarly to the *natural setting*, we conduct experiments on balanced IRDs $\mathcal{D}_{50\%}$ as well as on polarized IRDs $\mathcal{D}_{30\%}$ and $\mathcal{D}_{70\%}$.

As a baseline, we report the results obtained when $qt = qu$ (denoted as the "=" scenario), in order to show that the case $qt \neq qu$ constitutes a much more challenging scenario, which has never been investigated in the literature.

To construct the semantic query set $\mathcal{S}$ required by `FairLess`, we consider two augmentation strategies:

- $\mathcal{S}$ is generated directly from MSCOCO. In this case, when randomly selecting a query from the datasets described in Section 7.2 as qt, we keep another query of the same image as qu, while the remaining captions (up to 3) are used as augmentations of qt. This procedure is repeated twice, each time holding out a different caption as qu. We refer to this scenario as "DT";

[5] https://github.com/haoyusimon/VL_Debiasing.

[6] https://github.com/Harvard-Ophthalmology-AI-Lab/FairCLIP.

- S is generated using an LLM. In this case, when randomly selecting a query from the datasets described in Section 7.2 as qt, we keep another query of the same image for qu, and we generate $n_q = 3$ additional queries from qt using *Gemma-3-4b-it* [84], with the following prompt:

LLM Instruction Prompt

Given the following caption, generate exactly n_q variations that preserve the original meaning while rephrasing the wording.
Ensure diversity in structure and style while keeping the core message intact.
Return the output as a Python list format, where each variation is a string inside a list.
Here is the original caption: qt

We also tested the cases $n_q = 10$, instead of 3, to evaluate the sensitivity of the results to n_q, and the case $n_q = 0$ to assess whether augmentation is necessary. In the following, we refer to these three configurations as "LLM_3", "LLM_{10}" and "$\emptyset$". Together with DT, these form the set of semantic query configurations evaluated across all experiments.

The "$\emptyset$" configuration and the "=" scenario correspond to simplified variants of the proposed attack, and can thus be seen as baseline attacks within our setting, providing a reference for evaluating the proposed attack under simplified settings.

For the `FairLess` configuration, we specify the optimization parameters used in all experiments. In particular, we set the number of optimization steps to $K = 100$ and evaluate perturbations under both ℓ_∞ and ℓ_2 constraints:

- under ℓ_∞, we apply a perturbation budget of $\varepsilon = 4/255$ for both CLIP-based models and BLIP-2;
- under ℓ_2, we restrict the evaluation to CLIP-based models and vary ε from a minimal budget of 0.05 up to 1.0 to study the effect of progressively stronger perturbations on attack effectiveness.

Finally, we analyze different scenarios:

- injection of a single poisoned image into the IRD. This is designed to induce bias in a fully autonomous system that uses T2IR as a component, where such bias can undermine the overall reliability and safety of the system;
- injection of 5 poisoned images, considered only for CLIP-based models due to computational constraints. This serves to bias the entire top-5 retrieval results, thereby inducing stronger bias in the human user.

We also evaluate JPEG compression as an input-preprocessing mitigation applied before retrieval. In this case, we test three quality factors, $\text{QF} \in \{99, 94, 89\}$, where lower values correspond to stronger compression, and compare these configurations against the original adversarial configuration without compression, denoted as "Cl.".

8. Experimental results

In this section, we report the evaluation of VLPs across both the *natural* and *adversarial* settings (see Section 4.1). We begin with the *natural setting* (see Section 8.1), where no poisoned samples are introduced and retrieval performance reflects only intrinsic model biases and the composition of the IRD. We then examine the *adversarial setting* (see Section 8.2), assessing the impact of the `FairLess` attack (see Section 6) under multiple configurations, including single and multiple poisoned insertions (see Sections 8.2.1 and 8.2.4), different semantic augmentation strategies (see Section 8.2.2), and varying target-selection positions (see Section 8.2.3). Within the adversarial setting, we also evaluate a defensive strategy based on JPEG compression (see Section 8.2.5). Finally, we complement the empirical results with a statistical significance assessment of the observed fairness gaps (see Section 8.3). This structure allows us to separate intrinsic biases from attack-induced effects and to systematically analyze how VLPs interact with IRD polarization under increasingly challenging adversarial scenarios.

8.1. (Un)fairness under the natural setting

Table 1 reports the results in the *natural setting* (see Section 7.3), evaluated with the defined metrics (see Section 4.2) on the gender-annotated MSCOCO dataset, across different IRD polarizations (see Section 7.2) and VLP architectures (see Section 7.1). We recall that each image in the gender-annotated subset of MSCOCO has, by construction, at least one caption explicitly mentioning the gender of the depicted subject. As a consequence, the queries used in this setting fall into the class of *Attribute-labeled queries* introduced in Section 3.3, and the sensitive attribute is often made explicit in the text.

Starting from the balanced configuration $D_{50\%}$, which isolates intrinsic model bias, we observe that CLIP-based models achieve modest retrieval performance: ACC@1 is around 30%, and ACC@5 remains below or around 50%. FairCLIP_G performs worst, indicating that fairness-oriented fine-tuning can reduce accuracy. BLIP-2 models instead attain substantially higher accuracy, with ACC@1 above 59% and ACC@5 above 81%, particularly when using the ITM score. In this scenario, CLIP-based models already exhibit non-negligible disparities across gender groups, as indicated by non-zero ΔACC and ΔAMR values that generally favor female queries. CLIP-DEB_G shows the largest gaps, with $\Delta\text{AMR@5}$ exceeding 10 percentage points, suggesting that its debiasing procedure unintentionally over-favors the female group. FairCLIP_G instead displays the opposite trend: while ΔACC still mildly favors female queries, ΔAMR instead shifts in favor of male queries, indicating a higher likelihood of retrieving gender-consistent images for this group. This suggests that FairCLIP_G partially counteracts intrinsic biases in attribute matching. BLIP-2 and BLIP-2_ITM show smaller and less systematic gaps, indicating a weaker intrinsic bias for this architecture.

Moving to polarized IRDs to study dataset-induced bias, we first analyze $D_{30\%}$, where male images are underrepresented (30% male and 70% female). In this setting, the imbalance of the IRD primarily affects attribute matching: all models retrieve

Table 1
Results of the natural setting for gender polarization. For each VLP, results are shown at the top-1 and 5 ranks. POL denotes the IRD polarization: $\mathcal{D}_{50\%}$ (intrinsic model bias) and $\mathcal{D}_{30\%}$, $\mathcal{D}_{70\%}$ (dataset-induced bias), where % refers to male samples. Δs denote absolute gaps between groups, with values favoring female and male groups. Best values for ACC are highlighted in bold within each polarization setting across VLPs.

POL	VLP	RET@1			RET@5		
		ACC ↑	ΔACC ↓	ΔAMR ↓	ACC ↑	ΔACC ↓	ΔAMR ↓
$\mathcal{D}_{50\%}$	CLIP_{ViT} [4]	29.2 ± 1.9	1.6 ± 1.4	2.7 ± 2.2	52.9 ± 1.8	3.3 ± 2.1	3.7 ± 0.6
	CLIP_{CNN} [4]	27.3 ± 1.2	4.9 ± 1.8	3.9 ± 1.7	49.5 ± 1.6	6.3 ± 2.6	3.5 ± 1.9
	$\text{FairCLIP}_{\text{G}}$ [37]	24.1 ± 0.7	2.3 ± 2.0	2.9 ± 1.6	45.1 ± 0.5	3.5 ± 2.8	0.9 ± 1.5
	$\text{CLIP-DEB}_{\text{G}}$ [17]	28.5 ± 1.4	3.8 ± 0.9	8.8 ± 2.1	51.5 ± 0.9	6.1 ± 1.8	10.6 ± 0.7
	BLIP-2 [5]	59.7 ± 2.0	7.0 ± 3.3	0.7 ± 1.3	81.4 ± 0.8	3.7 ± 0.9	2.0 ± 1.7
	$\text{BLIP-2}_{\text{ITM}}$ [5]	**64.7 ± 3.4**	5.2 ± 1.4	0.7 ± 1.3	**84.0 ± 1.8**	4.1 ± 0.5	1.1 ± 0.5
$\mathcal{D}_{30\%}$	CLIP_{ViT} [4]	30.2 ± 0.9	4.4 ± 2.6	15.9 ± 1.8	52.8 ± 0.5	3.5 ± 2.3	23.6 ± 0.5
	CLIP_{CNN} [4]	28.1 ± 1.5	0.1 ± 0.4	17.1 ± 2.4	50.5 ± 0.2	1.7 ± 1.3	22.9 ± 2.4
	$\text{FairCLIP}_{\text{G}}$ [37]	24.4 ± 0.9	3.9 ± 3.2	8.3 ± 0.8	46.4 ± 0.3	3.5 ± 3.4	17.3 ± 2.1
	$\text{CLIP-DEB}_{\text{G}}$ [17]	29.0 ± 0.5	1.5 ± 1.9	21.5 ± 1.8	50.9 ± 0.6	1.1 ± 2.3	31.6 ± 0.6
	BLIP-2 [5]	60.2 ± 1.1	1.0 ± 1.1	3.4 ± 1.4	82.0 ± 0.9	0.7 ± 0.5	17.1 ± 1.0
	$\text{BLIP-2}_{\text{ITM}}$ [5]	**64.6 ± 2.9**	1.2 ± 2.6	3.2 ± 1.9	**84.6 ± 1.4**	0.1 ± 0.3	18.7 ± 0.3
$\mathcal{D}_{70\%}$	CLIP_{ViT} [4]	31.5 ± 1.8	6.2 ± 1.6	8.7 ± 3.8	53.4 ± 1.2	11.3 ± 1.1	15.1 ± 1.1
	CLIP_{CNN} [4]	27.8 ± 0.6	10.1 ± 2.0	10.3 ± 2.5	50.7 ± 1.0	12.3 ± 3.8	16.8 ± 0.7
	$\text{FairCLIP}_{\text{G}}$ [37]	25.0 ± 0.4	8.5 ± 2.9	14.9 ± 0.5	46.4 ± 0.3	11.3 ± 1.1	17.8 ± 2.2
	$\text{CLIP-DEB}_{\text{G}}$ [17]	30.4 ± 1.7	9.1 ± 1.4	4.4 ± 2.8	53.1 ± 0.9	12.8 ± 1.7	11.6 ± 1.5
	BLIP-2 [5]	60.5 ± 1.3	14.5 ± 2.8	4.3 ± 1.1	81.3 ± 1.4	9.9 ± 0.2	11.6 ± 1.5
	$\text{BLIP-2}_{\text{ITM}}$ [5]	**65.1 ± 2.1**	12.1 ± 2.6	4.1 ± 1.3	**84.2 ± 1.4**	7.7 ± 0.3	15.9 ± 2.1

a disproportionate number of female images, resulting in large ΔAMR gaps at both RET@1 and RET@5. This trend is particularly pronounced for $\text{CLIP-DEB}_{\text{G}}$, which exhibits the highest ΔAMR values, suggesting that its debiasing procedure amplifies the reliance on majority-group evidence in this configuration. At the same time, the limited retrieval accuracy of CLIP-based models influences ΔACC differently. Because male images are scarce, once a gender-consistent male candidate is retrieved, it is more likely to be the correct one, which reduces ΔACC or even shifts it in favor of male queries.

Finally, we analyze $\mathcal{D}_{70\%}$, where male images form the majority. As expected, all models show ΔAMR values that favor male queries, indicating higher attribute-match rates for this group. However, these gaps are considerably smaller than those observed in $\mathcal{D}_{30\%}$, where female queries benefited much more substantially from over-representation. This asymmetry suggests that the models are more sensitive to female than to male prevalence in the IRD. A particularly interesting behavior is observed for $\text{CLIP-DEB}_{\text{G}}$: whereas it showed the largest ΔAMR in favor of female queries under $\mathcal{D}_{30\%}$, it now exhibits the smallest ΔAMR among the CLIP-based models. This indicates that the debiasing procedure effectively dampens the influence of IRD imbalance when male images are the majority. BLIP-2 and $\text{BLIP-2}_{\text{ITM}}$ instead display a more moderate and symmetric response, mirroring their behavior in $\mathcal{D}_{30\%}$. In contrast, ΔACC remains consistently in favor of female queries across all architectures and often increases in magnitude, indicating higher retrieval accuracy for female queries despite their numerical minority. This behavior arises from the combination of an intrinsic tendency of the models to favor female-explicit content in the query and the reduced number of female images in the IRD, which, given the generally low accuracy of the models, limits the number of potential confounders and facilitates correct retrieval for female queries.

Table 2, report the analogous results of Table 1, for the MSCOCO race annotated. Unlike the gender annotated one, we recall that the race annotations in this dataset are human-provided. As a consequence, the queries used here do not systematically include explicit references to the sensitive attribute, and retrieval relies more strongly on the visual content of the image.

Starting from the balanced configuration $\mathcal{D}_{50\%}$, we saw that CLIP-based models obtain an higher accuracy than in the gender setting, with the exception of FairCLIPR, whose debiasing substantially lowers performance. However, the models also exhibit clear disparities across groups: ΔACC consistently favors light-skin queries, confirming trends widely reported in the literature. ΔAMR values, in contrast, remain small or close to zero for most models, indicating that attribute matching is not strongly affected by intrinsic bias in this setting. An exception is $\text{FairCLIP}_{\text{R}}$, whose race-oriented debiasing not only reduces accuracy, with RET@1 dropping to 18%, but also produces large ΔAMR values in favor of dark-skin queries. This indicates that the mitigation strategy over-corrects the original bias, prioritizing dark-skin attribute consistency at the cost of overall performance. On the other hand, BLIP-2 and $\text{BLIP-2}_{\text{ITM}}$ behave overall more reliably than the CLIP-based models, achieving high retrieval accuracy and exhibiting comparatively small and stable fairness gaps across both accuracy and attribute matching.

Moving to polarized IRDs, we first consider $\mathcal{D}_{30\%}$, where dark-skin images are the minority. In this configuration, all CLIP-based models show large ΔAMR gaps in favor of light-skin queries, reflecting the larger availability of light-skin images in the IRD. In general, retrieval accuracy also continues to favor light-skin queries across all architectures. However, also in this case, $\text{FairCLIP}_{\text{R}}$ exhibits a distinct pattern: while its ΔAMR still favors light-skin queries, the gap is considerably reduced compared to other models, and ΔACC even shifts slightly in favor of dark-skin queries. This behavior indicates that the model's debiasing continues to push

Table 2
Results of the natural setting for race polarization. For each VLP, results are shown at the top-1 and 5 ranks. POL denotes the IRD polarization: $\mathcal{D}_{50\%}$ (intrinsic model bias) and $\mathcal{D}_{30\%}$, $\mathcal{D}_{70\%}$ (dataset-induced bias), where % refers to dark-skin samples. Δs denote absolute gaps between groups, with values favoring light-skin and dark-skin groups. Best values for ACC are highlighted in bold within each polarization setting across VLPs.

POL	VLP	RET@1			RET@5		
		ACC ↑	ΔACC ↓	ΔAMR ↓	ACC ↑	ΔACC ↓	ΔAMR ↓
$\mathcal{D}_{50\%}$	CLIP_{ViT} [4]	41.4 ± 1.0	10.1 ± 4.7	0.9 ± 1.4	69.1 ± 1.3	7.2 ± 2.9	2.7 ± 2.8
	CLIP_{CNN} [4]	41.0 ± 1.0	5.5 ± 1.7	2.1 ± 2.3	67.1 ± 1.5	5.6 ± 2.6	5.0 ± 1.6
	$\text{FairCLIP}_{\text{R}}$ [37]	18.0 ± 1.8	1.1 ± 1.1	17.8 ± 8.1	41.4 ± 1.4	0.6 ± 1.9	15.6 ± 3.8
	$\text{CLIP-DEB}_{\text{R}}$ [17]	41.0 ± 1.1	8.5 ± 5.0	0.8 ± 2.1	68.4 ± 2.2	7.1 ± 3.0	3.8 ± 3.1
	BLIP-2 [5]	72.1 ± 0.8	5.6 ± 3.5	1.0 ± 0.8	90.3 ± 0.5	3.7 ± 0.6	4.0 ± 1.7
	$\text{BLIP-2}_{\text{ITM}}$ [5]	**74.9 ± 0.2**	3.0 ± 2.1	0.9 ± 1.6	**91.3 ± 0.1**	1.5 ± 0.5	6.2 ± 1.0
$\mathcal{D}_{30\%}$	CLIP_{ViT} [4]	41.2 ± 1.0	5.8 ± 2.6	19.2 ± 1.8	67.4 ± 1.8	3.9 ± 1.9	27.6 ± 3.1
	CLIP_{CNN} [4]	38.7 ± 1.4	5.6 ± 2.6	28.1 ± 1.3	65.9 ± 1.2	6.5 ± 1.4	36.2 ± 2.4
	$\text{FairCLIP}_{\text{R}}$ [37]	18.6 ± 0.5	2.9 ± 2.4	8.3 ± 7.6	41.2 ± 1.7	0.3 ± 3.0	16.3 ± 8.2
	$\text{CLIP-DEB}_{\text{R}}$ [17]	40.2 ± 0.4	5.6 ± 2.1	19.5 ± 2.8	67.0 ± 1.8	5.1 ± 1.9	27.4 ± 3.2
	BLIP-2 [5]	72.2 ± 0.8	3.3 ± 5.0	10.5 ± 3.4	90.5 ± 1.3	2.4 ± 3.2	26.0 ± 2.4
	$\text{BLIP-2}_{\text{ITM}}$ [5]	**74.9 ± 1.2**	2.0 ± 3.2	9.4 ± 1.9	**91.0 ± 1.3**	1.0 ± 2.3	25.0 ± 2.5
$\mathcal{D}_{70\%}$	CLIP_{ViT} [4]	41.1 ± 0.2	9.6 ± 2.7	20.7 ± 3.3	68.4 ± 1.9	7.6 ± 1.9	33.9 ± 2.8
	CLIP_{CNN} [4]	40.9 ± 0.6	8.9 ± 1.9	17.0 ± 1.7	67.0 ± 1.2	11.5 ± 4.7	27.2 ± 0.3
	$\text{FairCLIP}_{\text{R}}$ [37]	17.6 ± 1.3	0.7 ± 1.5	36.3 ± 14.4	40.8 ± 1.5	0.6 ± 1.9	42.2 ± 7.3
	$\text{CLIP-DEB}_{\text{R}}$ [17]	39.6 ± 0.6	8.7 ± 3.1	22.5 ± 3.7	67.0 ± 2.0	7.0 ± 3.2	34.8 ± 3.1
	BLIP-2 [5]	71.0 ± 0.5	4.9 ± 2.2	13.3 ± 1.5	90.7 ± 0.9	4.1 ± 2.2	32.8 ± 2.0
	$\text{BLIP-2}_{\text{ITM}}$ [5]	**74.8 ± 2.3**	3.7 ± 0.8	11.3 ± 3.4	**91.5 ± 0.9**	2.9 ± 1.9	35.1 ± 1.8

results toward the dark-skin group, diminishing the impact of IRD imbalance and partially reversing the accuracy trend observed in other models. BLIP-2 and $\text{BLIP-2}_{\text{ITM}}$ again show more moderate and stable fairness gaps.

Finally, we analyze $\mathcal{D}_{70\%}$, where dark-skin images form the majority. As expected, ΔAMR values now favor dark-skin queries for all models, with gaps that increase with the degree of IRD imbalance. Retrieval accuracy, however, continues to favor light-skin queries for all models except $\text{FairCLIP}_{\text{R}}$, and in many cases the light-skin advantage becomes even larger than in $\mathcal{D}_{30\%}$. This is explained by the combination of: (i) higher intrinsic accuracy for light-skin samples and (ii) the fact that light-skin images are now the minority, which reduces the number of confounders and makes correct retrieval more likely for light-skin queries. $\text{FairCLIP}_{\text{R}}$ again exhibits a markedly different behavior: its ΔAMR shows the strongest preference for dark-skin queries among all models (with large variability across runs), indicating that the debiasing procedure strongly amplifies attribute matching for the minority group in $\mathcal{D}_{30\%}$ and the majority group in $\mathcal{D}_{70\%}$. BLIP-2 and $\text{BLIP-2}_{\text{ITM}}$ maintain the same moderate and symmetric response observed under $\mathcal{D}_{30\%}$.

8.2. (Un)fairness under the adversarial setting

In this section, we consider the *adversarial setting* (see Section 7.4) to assess the impact of the `FairLess` attack (see Section 6) using the defined metrics (see Section 4.2). Our systematic evaluation is designed to analyze how `FairLess` is influenced by the intrinsic behavior of VLPs and by the IRD polarization under different retrieval conditions (see Section 7.2). In this setting, Phase 1 (STS) is performed by setting $\mathcal{E} = D$, meaning that target images are selected from the IRD (see Section 6.1). However, we recall that $\mathcal{E}$ may be set to any external database, and the attack formulation remains valid for arbitrary choices of the target pool.

8.2.1. Impact of a single poisoned insertion

As a first analysis, we study the effect of `FairLess` when a single poisoned image is inserted into the IRD for each concept expressed by a query. To do so, the target is selected following the procedure described in Section 6.1, specifically Eq. (15), by taking the top-ranked image that is most similar to the query concept while belonging to the opposite sensitive-attribute group. In this scenario, we also adopt DT as the semantic augmentation strategy to construct $\mathcal{S}$ in Phase 2 of the attack (see Section 6.2).

Tables 3 and 4 report the results of the single–poisoned insertion scenario for the gender annotated MSCOCO.

Table 3 shows the case considering the ℓ_∞ attack with $\varepsilon = \frac{4}{255}$. For RET@1, CLIP-based models exhibit an ASR close to 100% across all IRD polarizations, effectively driving the accuracy of the systems to zero. This shows that the attack is equally effective for both sensitive-attribute groups, as also reflected by the near-zero values of ΔACC and ΔAMR. The attack is less effective on BLIP-2, although ASR remains relatively high when cosine similarity is used as the function ϕ. In contrast, $\text{BLIP-2}_{\text{ITM}}$ is substantially more robust in this setting, with ASR values around 15%, indicating that, in this more challenging scenario where the attacker does not know the similarity function, a larger perturbation budget would be required to achieve high attack success. For RET@5, all models achieve high ASR regardless of IRD polarization. In this case, the attack succeeds whenever the poisoned image appears within the top five retrieved results, making the task easier. It is also noteworthy that the patterns observed in the *natural setting* (Table

Table 3
Results of the adversarial setting for gender polarization under ℓ_∞ ($\varepsilon = \frac{4}{255}$), with the DT augmentation strategy. For each VLP, results are shown at the top-1 and 5 ranks. POL denotes the IRD polarization: $\mathcal{D}_{50\%}$ (intrinsic model bias) and $\mathcal{D}_{30\%}$, $\mathcal{D}_{70\%}$ (dataset-induced bias), where % refers to male samples. Δs denote absolute gaps between groups, with values favoring female and male groups. Best values for ACC and ASR are highlighted in bold within each polarization setting across VLPs.

POL	VLP	RET@1					RET@5				
		ACC ↑	ΔACC ↓	ASR ↑	ΔASR ↓	ΔAMR ↓	ACC ↑	ΔACC ↓	ASR ↑	ΔASR ↓	ΔAMR ↓
$\mathcal{D}_{50\%}$	CLIP_{ViT} [4]	0.0 ± 0.1	0.1 ± 0.1	99.2 ± 0.1	0.1 ± 0.2	0.0 ± 0.1	49.2 ± 2.1	3.6 ± 1.9	99.7 ± 0.1	0.0 ± 0.2	2.7 ± 0.3
	CLIP_{CNN} [4]	0.1 ± 0.2	0.1 ± 0.1	99.4 ± 0.2	0.3 ± 0.3	0.3 ± 0.3	45.8 ± 1.5	5.7 ± 3.1	99.8 ± 0.1	0.2 ± 0.0	3.1 ± 1.0
	$\text{FairCLIP}_{\text{G}}$ [37]	0.0 ± 0.0	0.0 ± 0.0	**100.0 ± 0.0**	0.0 ± 0.0	0.0 ± 0.0	42.1 ± 0.7	3.7 ± 3.2	**100.0 ± 0.0**	0.0 ± 0.0	1.2 ± 1.3
	$\text{CLIP-DEB}_{\text{G}}$ [17]	0.0 ± 0.0	0.0 ± 0.0	99.7 ± 0.1	0.3 ± 0.5	0.3 ± 0.3	47.8 ± 1.6	5.5 ± 1.5	99.8 ± 0.1	0.1 ± 0.1	8.2 ± 0.4
	BLIP-2 [5]	10.6 ± 0.5	1.0 ± 1.3	77.1 ± 1.1	3.2 ± 2.7	3.5 ± 2.6	79.2 ± 1.1	5.2 ± 1.2	92.9 ± 0.7	0.8 ± 0.6	1.1 ± 0.9
	$\text{BLIP-2}_{\text{ITM}}$ [5]	**53.7 ± 2.2**	3.2 ± 2.8	14.7 ± 0.8	1.7 ± 1.4	2.5 ± 2.3	**83.3 ± 1.7**	4.0 ± 0.8	73.0 ± 0.2	3.7 ± 1.3	0.2 ± 0.4
$\mathcal{D}_{30\%}$	CLIP_{ViT} [4]	0.0 ± 0.1	0.1 ± 0.1	99.3 ± 0.1	0.3 ± 0.3	0.3 ± 0.3	49.8 ± 0.8	3.5 ± 1.8	99.5 ± 0.1	0.1 ± 0.2	18.8 ± 0.6
	CLIP_{CNN} [4]	0.1 ± 0.2	0.2 ± 0.3	99.4 ± 0.2	0.7 ± 0.4	0.7 ± 0.5	47.1 ± 0.8	2.2 ± 0.9	99.8 ± 0.2	0.3 ± 0.1	17.6 ± 1.4
	$\text{FairCLIP}_{\text{G}}$ [37]	0.0 ± 0.0	0.0 ± 0.0	**100.0 ± 0.0**	0.0 ± 0.0	0.0 ± 0.0	43.6 ± 0.9	3.3 ± 3.2	**100.0 ± 0.0**	0.0 ± 0.0	12.4 ± 2.0
	$\text{CLIP-DEB}_{\text{G}}$ [17]	0.0 ± 0.0	0.0 ± 0.0	99.6 ± 0.1	0.1 ± 0.3	0.1 ± 0.0	47.9 ± 1.0	0.6 ± 1.7	99.8 ± 0.1	0.2 ± 0.0	24.9 ± 0.4
	BLIP-2 [5]	10.8 ± 0.8	2.2 ± 1.4	77.4 ± 0.4	1.7 ± 1.4	1.5 ± 0.8	80.3 ± 1.0	0.8 ± 0.8	92.9 ± 0.3	0.7 ± 1.0	12.6 ± 0.7
	$\text{BLIP-2}_{\text{ITM}}$ [5]	**53.9 ± 2.2**	2.4 ± 2.9	14.7 ± 1.2	0.9 ± 0.6	1.7 ± 1.1	**83.8 ± 1.4**	0.1 ± 0.1	73.8 ± 0.8	2.1 ± 1.2	14.9 ± 0.5
$\mathcal{D}_{70\%}$	CLIP_{ViT} [4]	0.0 ± 0.1	0.1 ± 0.1	99.2 ± 0.1	0.1 ± 0.3	0.3 ± 0.3	50.6 ± 1.1	11.0 ± 2.1	99.8 ± 0.1	0.1 ± 0.2	11.3 ± 1.4
	CLIP_{CNN} [4]	0.1 ± 0.2	0.1 ± 0.1	99.4 ± 0.2	0.3 ± 0.3	0.0 ± 0.1	47.5 ± 1.6	11.5 ± 4.6	99.7 ± 0.2	0.1 ± 0.1	13.1 ± 0.6
	$\text{FairCLIP}_{\text{G}}$ [37]	0.0 ± 0.0	0.0 ± 0.0	**100.0 ± 0.0**	0.0 ± 0.0	0.0 ± 0.0	43.1 ± 0.5	10.3 ± 2.4	**100.0 ± 0.0**	0.0 ± 0.0	14.0 ± 1.7
	$\text{CLIP-DEB}_{\text{G}}$ [17]	0.0 ± 0.0	0.0 ± 0.0	99.7 ± 0.1	0.3 ± 0.5	0.4 ± 0.3	49.8 ± 0.7	13.5 ± 1.9	99.8 ± 0.1	0.1 ± 0.1	8.1 ± 1.5
	BLIP-2 [5]	10.7 ± 0.4	0.1 ± 1.1	77.0 ± 0.9	4.6 ± 3.9	6.3 ± 4.8	79.3 ± 1.7	11.1 ± 1.1	92.8 ± 0.6	3.0 ± 0.2	9.3 ± 1.4
	$\text{BLIP-2}_{\text{ITM}}$ [5]	**54.0 ± 1.5**	8.1 ± 4.0	14.8 ± 0.9	3.0 ± 1.9	6.6 ± 2.4	**83.7 ± 1.5**	7.9 ± 0.1	73.0 ± 0.9	10.5 ± 0.4	14.3 ± 1.9

Table 4
Results of the adversarial setting for gender polarization under ℓ_2 ($\varepsilon = 0.05$), with the DT augmentation strategy. For each VLP, results are shown at the top-1 and 5 ranks. POL denotes the IRD polarization: $\mathcal{D}_{50\%}$ (intrinsic model bias) and $\mathcal{D}_{30\%}$, $\mathcal{D}_{70\%}$ (dataset-induced bias), where % refers to male samples. Δs denote absolute gaps between groups, with values favoring female and male groups. Best values for ACC and ASR are highlighted in bold within each polarization setting across VLPs.

POL	VLP	RET@1					RET@5				
		ACC ↑	ΔACC ↓	ASR ↑	ΔASR ↓	ΔAMR ↓	ACC ↑	ΔACC ↓	ASR ↑	ΔASR ↓	ΔAMR ↓
$\mathcal{D}_{50\%}$	CLIP_{ViT} [4]	20.1 ± 2.1	0.2 ± 0.9	43.1 ± 0.5	0.7 ± 0.5	0.7 ± 0.7	**50.1 ± 1.9**	3.5 ± 2.1	74.6 ± 0.8	0.2 ± 0.8	2.7 ± 0.4
	CLIP_{CNN} [4]	18.2 ± 1.3	3.8 ± 1.1	43.0 ± 1.2	1.5 ± 1.1	3.0 ± 1.8	46.8 ± 1.4	6.1 ± 2.8	73.0 ± 0.7	1.3 ± 0.7	3.2 ± 1.4
	$\text{FairCLIP}_{\text{G}}$ [37]	**20.3 ± 1.5**	1.1 ± 0.8	28.2 ± 1.7	2.4 ± 1.4	3.5 ± 1.7	43.5 ± 0.9	3.8 ± 3.0	58.1 ± 1.3	2.5 ± 1.8	0.6 ± 1.3
	$\text{CLIP-DEB}_{\text{G}}$ [17]	19.3 ± 1.6	2.7 ± 2.5	**43.6 ± 1.2**	5.2 ± 0.5	8.1 ± 0.6	48.7 ± 1.3	5.7 ± 1.1	**75.1 ± 0.7**	3.9 ± 0.3	9.3 ± 0.6
$\mathcal{D}_{30\%}$	CLIP_{ViT} [4]	20.2 ± 2.0	3.0 ± 2.0	**43.8 ± 0.6**	7.5 ± 1.3	12.0 ± 1.9	**50.5 ± 0.8**	3.7 ± 2.0	**75.3 ± 0.7**	7.5 ± 0.4	20.6 ± 0.4
	CLIP_{CNN} [4]	19.0 ± 1.3	0.1 ± 0.4	43.0 ± 1.5	4.7 ± 0.3	10.4 ± 1.8	48.0 ± 0.4	1.7 ± 0.6	73.3 ± 0.6	7.9 ± 2.3	19.7 ± 1.2
	$\text{FairCLIP}_{\text{G}}$ [37]	**20.4 ± 1.6**	3.8 ± 1.8	28.5 ± 1.7	1.1 ± 2.0	5.2 ± 2.5	44.9 ± 0.4	3.1 ± 3.5	58.9 ± 0.6	12.6 ± 1.2	16.1 ± 2.3
	$\text{CLIP-DEB}_{\text{G}}$ [17]	19.0 ± 1.2	0.2 ± 1.1	43.7 ± 0.9	10.3 ± 1.5	17.9 ± 1.5	48.7 ± 1.0	0.5 ± 1.6	75.2 ± 1.3	10.1 ± 1.6	28.0 ± 0.5
$\mathcal{D}_{70\%}$	CLIP_{ViT} [4]	**21.1 ± 2.0**	2.9 ± 2.0	44.1 ± 0.8	4.8 ± 1.4	8.6 ± 3.0	**51.2 ± 0.9**	10.3 ± 1.8	75.3 ± 1.0	6.8 ± 0.4	13.5 ± 1.2
	CLIP_{CNN} [4]	18.5 ± 1.1	6.0 ± 1.5	43.3 ± 1.3	3.5 ± 1.2	7.4 ± 1.6	48.4 ± 1.7	11.5 ± 3.9	73.5 ± 0.3	4.6 ± 0.2	14.8 ± 0.5
	$\text{FairCLIP}_{\text{G}}$ [37]	20.7 ± 1.4	5.7 ± 2.2	28.2 ± 1.9	7.5 ± 1.6	14.3 ± 1.3	44.8 ± 0.5	11.5 ± 1.9	59.0 ± 1.2	7.6 ± 3.0	16.4 ± 2.3
	$\text{CLIP-DEB}_{\text{G}}$ [17]	20.0 ± 1.7	5.4 ± 2.0	**44.7 ± 0.6**	0.4 ± 0.6	2.0 ± 0.9	50.7 ± 0.0	13.3 ± 1.8	**76.6 ± 0.6**	3.3 ± 0.4	9.3 ± 1.3

1) re-emerge here: since only a single poisoned image shifts into the top positions, the remaining four retrieved items tend to be legitimate images whose distribution reflects the same biases seen under the natural configuration.

Table 4 reports the results obtained under the ℓ_2 attack with a minimal perturbation budget of $\varepsilon = 0.05$ for the CLIP-based models. This small perturbation is intentionally chosen to highlight how group disparities behave when the attack does not fully disrupt the retrieval ranking. Moreover, such a small budget is sufficient in this scenario because the selected targets already exhibit high similarity with the query (they are still the top opposite-attribute matches), meaning that only a limited perturbation is required to push them into the top retrieved positions. In this setting, $\text{FairCLIP}_{\text{G}}$ emerges as the most robust CLIP-based model, although it already exhibited lower accuracy than the other architectures in the *natural setting*. Under $\mathcal{D}_{50\%}$, the attack behaves similarly across gender groups for most models, with small differences in ASR. A notable exception is $\text{CLIP-DEB}_{\text{G}}$, which exhibits a substantially larger ΔASR in favor of attacking male-related concepts, reflecting a stronger consistency of female matches in the retrieved results at RET@1. This pattern may be explained by the debiasing procedure, whose influence on female attribute matching was already visible in the *natural setting* (Table 1), where $\text{CLIP-DEB}_{\text{G}}$ presented slightly higher attribute-match rates for the female group even under balanced conditions. When moving to polarized IRDs ($\mathcal{D}_{30\%}$ and $\mathcal{D}_{70\%}$), the attack generally reduces the disparities in ΔAMR observed in the *natural setting*. As a consequence, the resulting ΔASR values tend to shift toward the minority group in the IRD. This behavior reflects the fact that a single poisoned insertion has a stronger effect when targeting the less represented attribute group, since the IRD contains a larger number of opposite-attribute images, making it more likely to retrieve items that align with the attacked concept.

Table 5
Results of the adversarial setting for race polarization under ℓ_∞ ($\epsilon = \frac{4}{255}$), with the DT augmentation strategy. For each VLP, results are shown at the top-1 and 5 ranks. POL denotes the IRD polarization: $\mathcal{D}_{50\%}$ (intrinsic model bias) and $\mathcal{D}_{30\%}$, $\mathcal{D}_{70\%}$ (dataset-induced bias), where % refers to dark-skin samples. Δs denote absolute gaps between groups, with values favoring light-skin and dark-skin groups. Best values for ACC and ASR are highlighted in bold within each polarization setting across VLPs.

POL	VLP	RET@1					RET@5				
		ACC ↑	ΔACC ↓	ASR ↑	ΔASR ↓	ΔAMR ↓	ACC ↑	ΔACC ↓	ASR ↑	ΔASR ↓	ΔAMR ↓
$\mathcal{D}_{50\%}$	CLIP_{ViT} [4]	0.2 ± 0.2	0.2 ± 0.3	99.4 ± 0.1	0.1 ± 0.4	0.1 ± 0.5	65.6 ± 1.7	7.3 ± 2.5	99.7 ± 0.1	0.1 ± 0.1	1.7 ± 1.8
	CLIP_{CNN} [4]	0.1 ± 0.1	0.1 ± 0.1	99.5 ± 0.3	0.3 ± 0.2	0.1 ± 0.3	63.9 ± 1.5	5.3 ± 2.5	99.8 ± 0.2	0.2 ± 0.2	4.2 ± 0.5
	$\text{FairCLIP}_{\text{R}}$ [37]	0.0 ± 0.0	0.0 ± 0.0	**100.0 ± 0.0**	0.0 ± 0.0	0.0 ± 0.0	37.5 ± 1.2	0.4 ± 2.2	**100.0 ± 0.0**	0.0 ± 0.0	12.9 ± 4.2
	$\text{CLIP-DEB}_{\text{R}}$ [17]	0.1 ± 0.1	0.0 ± 0.2	99.6 ± 0.1	0.1 ± 0.2	0.0 ± 0.2	63.9 ± 1.8	7.1 ± 1.4	99.8 ± 0.1	0.0 ± 0.1	2.6 ± 2.5
	BLIP-2 [5]	11.4 ± 1.1	2.3 ± 0.6	81.6 ± 2.3	1.7 ± 0.2	1.6 ± 0.9	89.1 ± 0.5	3.6 ± 0.9	96.0 ± 0.3	0.9 ± 0.8	3.4 ± 1.8
	$\text{BLIP-2}_{\text{ITM}}$ [5]	**61.3 ± 1.6**	7.4 ± 2.1	17.8 ± 1.9	5.6 ± 2.1	4.5 ± 3.2	**90.3 ± 0.5**	1.7 ± 0.6	85.8 ± 0.8	2.6 ± 1.2	5.0 ± 1.7
$\mathcal{D}_{30\%}$	CLIP_{ViT} [4]	0.1 ± 0.2	0.0 ± 0.1	99.4 ± 0.7	0.1 ± 0.1	0.1 ± 0.2	64.2 ± 1.7	4.7 ± 2.0	99.8 ± 0.3	0.1 ± 0.1	21.5 ± 3.2
	CLIP_{CNN} [4]	0.1 ± 0.0	0.1 ± 0.0	99.4 ± 0.4	0.0 ± 0.0	0.1 ± 0.2	62.8 ± 1.4	5.4 ± 0.7	99.8 ± 0.3	0.1 ± 0.1	28.6 ± 1.6
	$\text{FairCLIP}_{\text{R}}$ [37]	0.0 ± 0.0	0.0 ± 0.0	**100.0 ± 0.0**	0.0 ± 0.0	0.0 ± 0.0	37.6 ± 2.1	1.0 ± 3.1	**100.0 ± 0.0**	0.0 ± 0.0	12.1 ± 6.5
	$\text{CLIP-DEB}_{\text{R}}$ [17]	0.1 ± 0.0	0.0 ± 0.0	99.6 ± 0.4	0.1 ± 0.1	0.0 ± 0.0	62.7 ± 2.1	4.8 ± 2.3	99.8 ± 0.3	0.0 ± 0.0	20.4 ± 3.9
	BLIP-2 [5]	11.5 ± 1.3	1.7 ± 1.6	81.3 ± 1.5	0.4 ± 0.9	3.1 ± 1.2	89.2 ± 1.4	2.9 ± 3.8	96.5 ± 1.2	0.3 ± 0.6	20.2 ± 1.7
	$\text{BLIP-2}_{\text{ITM}}$ [5]	**61.0 ± 1.3**	4.1 ± 3.0	17.7 ± 0.5	2.5 ± 1.9	10.5 ± 5.2	**90.0 ± 1.2**	1.1 ± 2.5	85.2 ± 2.6	2.1 ± 1.9	19.9 ± 1.6
$\mathcal{D}_{70\%}$	CLIP_{ViT} [4]	0.1 ± 0.2	0.0 ± 0.0	99.4 ± 0.3	0.1 ± 0.1	0.2 ± 0.3	64.7 ± 1.9	7.9 ± 2.8	99.7 ± 0.1	0.1 ± 0.1	25.8 ± 2.3
	CLIP_{CNN} [4]	0.0 ± 0.1	0.1 ± 0.1	99.7 ± 0.1	0.1 ± 0.1	0.2 ± 0.2	64.0 ± 1.7	10.7 ± 4.2	99.8 ± 0.1	0.1 ± 0.1	20.5 ± 0.6
	$\text{FairCLIP}_{\text{R}}$ [37]	0.0 ± 0.0	0.0 ± 0.0	**100.0 ± 0.0**	0.0 ± 0.0	0.0 ± 0.0	37.8 ± 1.1	0.5 ± 3.1	**100.0 ± 0.0**	0.0 ± 0.0	33.2 ± 6.8
	$\text{CLIP-DEB}_{\text{R}}$ [17]	0.1 ± 0.2	0.1 ± 0.2	99.7 ± 0.2	0.5 ± 0.3	0.2 ± 0.2	63.2 ± 1.2	7.7 ± 2.5	99.9 ± 0.0	0.2 ± 0.0	26.8 ± 2.7
	BLIP-2 [5]	11.0 ± 1.0	4.5 ± 1.0	82.0 ± 1.3	1.9 ± 1.4	0.0 ± 0.4	89.1 ± 0.8	3.1 ± 1.8	96.8 ± 0.4	1.4 ± 0.9	25.0 ± 1.6
	$\text{BLIP-2}_{\text{ITM}}$ [5]	**61.6 ± 1.4**	6.5 ± 0.5	17.6 ± 0.6	2.9 ± 1.0	6.0 ± 2.4	**90.5 ± 1.0**	3.3 ± 2.2	85.9 ± 1.9	3.2 ± 2.3	28.1 ± 1.6

Table 6
Results of the adversarial setting for race polarization under ℓ_2 ($\epsilon = 0.05$), with the DT augmentation strategy. For each VLP, results are shown at the top-1 and 5 ranks. POL denotes the IRD polarization: $\mathcal{D}_{50\%}$ (intrinsic model bias) and $\mathcal{D}_{30\%}$, $\mathcal{D}_{70\%}$ (dataset-induced bias), where % refers to dark-skin samples. Δs denote absolute gaps between groups, with blue indicating light-skin–favored and green dark-skin–favored values. Best values for ACC and ASR are highlighted in bold within each polarization setting across VLPs.

POL	VLP	RET@1					RET@5				
		ACC ↑	ΔACC ↓	ASR ↑	ΔASR ↓	ΔAMR ↓	ACC ↑	ΔACC ↓	ASR ↑	ΔASR ↓	ΔAMR ↓
$\mathcal{D}_{50\%}$	CLIP_{ViT} [4]	**25.8 ± 0.7**	16.5 ± 3.7	54.4 ± 0.2	16.7 ± 3.4	16.9 ± 3.4	**66.0 ± 1.7**	7.1 ± 2.6	89.4 ± 0.6	5.1 ± 0.3	1.3 ± 1.7
	CLIP_{CNN} [4]	24.6 ± 1.6	13.3 ± 1.9	53.8 ± 2.5	18.3 ± 1.8	19.1 ± 2.2	64.6 ± 1.4	5.4 ± 2.1	87.3 ± 0.9	4.8 ± 1.0	4.9 ± 0.5
	$\text{FairCLIP}_{\text{R}}$ [37]	9.7 ± 0.3	2.9 ± 2.8	**66.0 ± 1.8**	12.4 ± 3.6	5.7 ± 5.0	37.8 ± 1.3	0.3 ± 2.4	**91.9 ± 0.8**	2.5 ± 1.7	13.0 ± 4.3
	$\text{CLIP-DEB}_{\text{R}}$ [17]	24.4 ± 1.5	14.9 ± 2.9	55.5 ± 1.1	16.3 ± 2.6	16.7 ± 3.0	64.5 ± 1.8	6.9 ± 1.3	89.9 ± 0.7	4.9 ± 1.3	2.0 ± 2.7
$\mathcal{D}_{30\%}$	CLIP_{ViT} [4]	**24.1 ± 1.2**	13.6 ± 2.8	56.3 ± 2.6	15.1 ± 4.8	22.3 ± 4.1	**64.5 ± 1.7**	4.6 ± 1.4	88.3 ± 1.0	6.1 ± 1.7	23.0 ± 3.5
	CLIP_{CNN} [4]	21.6 ± 1.5	11.2 ± 2.0	55.3 ± 1.6	20.9 ± 5.4	28.0 ± 3.0	63.3 ± 1.4	6.2 ± 0.7	87.3 ± 1.0	6.5 ± 0.6	30.6 ± 1.9
	$\text{FairCLIP}_{\text{R}}$ [37]	9.1 ± 0.9	1.5 ± 1.2	**68.1 ± 1.5**	15.2 ± 4.5	14.1 ± 4.4	37.8 ± 2.2	1.0 ± 3.2	**92.7 ± 0.7**	3.9 ± 1.7	12.8 ± 6.7
	$\text{CLIP-DEB}_{\text{R}}$ [17]	22.7 ± 3.3	11.9 ± 1.8	58.7 ± 3.3	14.0 ± 3.4	20.5 ± 3.5	63.1 ± 2.1	4.9 ± 2.4	89.6 ± 0.6	4.8 ± 0.6	21.9 ± 3.9
$\mathcal{D}_{70\%}$	CLIP_{ViT} [4]	**25.5 ± 0.2**	16.5 ± 1.5	56.2 ± 0.8	18.3 ± 4.9	11.9 ± 3.1	**65.0 ± 1.8**	8.3 ± 2.8	89.7 ± 0.3	4.6 ± 1.4	26.0 ± 2.3
	CLIP_{CNN} [4]	24.9 ± 0.5	18.5 ± 1.4	55.5 ± 1.7	23.1 ± 4.8	16.3 ± 4.0	64.5 ± 1.4	11.0 ± 4.5	88.2 ± 1.3	5.3 ± 1.9	20.5 ± 0.9
	$\text{FairCLIP}_{\text{R}}$ [37]	9.5 ± 1.0	5.8 ± 1.7	**66.2 ± 2.8**	11.5 ± 6.8	0.0 ± 5.0	38.0 ± 1.3	0.3 ± 3.0	**91.8** ± 0.8	2.0 ± 1.4	33.9 ± 6.9
	$\text{CLIP-DEB}_{\text{R}}$ [17]	24.8 ± 0.9	15.4 ± 2.4	56.6 ± 1.0	17.3 ± 3.4	11.1 ± 3.9	63.8 ± 1.4	7.8 ± 3.1	89.7 ± 0.9	5.1 ± 1.0	27.1 ± 2.5

Tables 5 and 6 report the analogous results of Tables 3 and 4 for the race annotated MSCOCO under the ℓ_∞ ($\epsilon = \frac{4}{255}$) and ℓ_2 ($\epsilon = 0.05$) attack respectively.

Table 5 shows a trend that closely related the gender case. For CLIP-based models, the ℓ_∞ attack is strong enough to disrupt retrieval almost entirely, reaching ASR values close to 100% across all IRD polarizations and driving the accuracy to zero. As a result, the group gaps are essentially eliminated, and the values of ΔACC, ΔASR, and ΔAMR remain extremely small, indicating that the attack affects both sensitive-attribute groups to the same extent. A similar outcome appears for BLIP-2, which also shows high vulnerability in this configuration when cosine similarity is used as the ϕ function. In contrast, $\text{BLIP-2}_{\text{ITM}}$ remains substantially more robust, with significantly lower ASR values. This confirms the behavior already observed in the gender case and indicates that a larger perturbation budget would be required to achieve high attack success under ITM-based retrieval. For RET@5, all models achieve very high ASR across polarizations. Since the attack succeeds whenever the poisoned image enters the top five positions, the task is less demanding, and the poisoned item easily displaces at least one clean candidate regardless of the underlying IRD distribution. As in the gender case, the remaining retrieved items tend to be legitimate images, and their distribution reproduces the same group patterns observed in the *natural setting* (Table 2).

Table 6 reports the results for the CLIP-based models under the ℓ_2 attack with a minimal perturbation budget ($\epsilon = 0.05$). The small perturbation makes it possible to observe an important effect: the adversarial manipulation tends to be substantially more effective on dark-skin concepts, which appear easier to perturb in this setting. In particular, for dark-skin queries the attack more

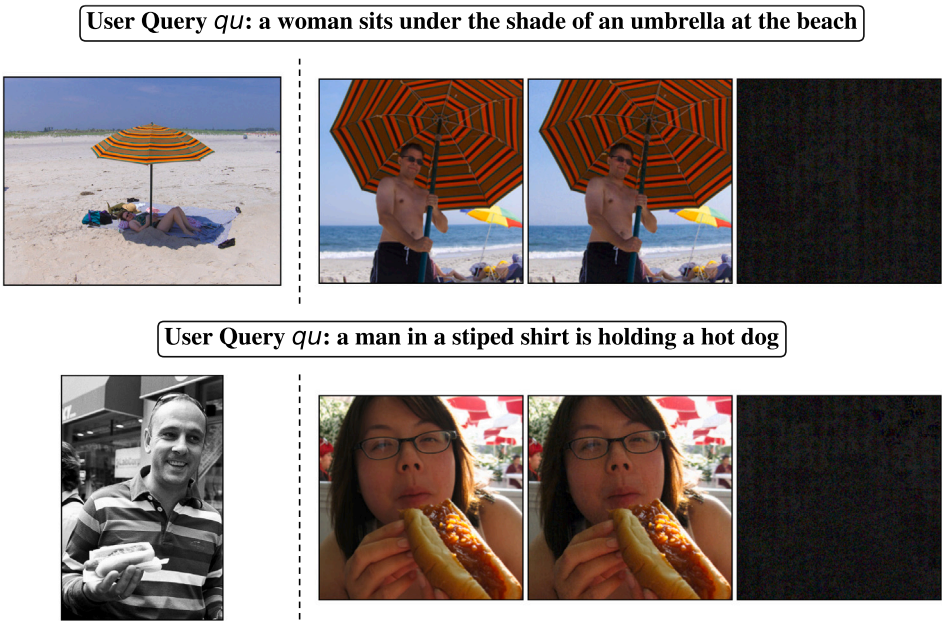

Fig. 2. Poisoned samples for gender polarization under the ℓ_∞ attack ($\epsilon = \frac{4}{255}$) with the DT augmentation strategy. The top row corresponds to CLIP_{ViT} [4] (attacking female concepts), and the bottom row to $\text{FairCLIP}_{\text{G}}$ [37] (attacking male concepts). Each row includes the user query qu, the ground-truth image, the clean and poisoned targets, and the pixel-wise difference (amplified by a factor of 10).

frequently promotes light-skin poisoned images over the corresponding legitimate dark-skin matches, leading in many cases to ΔAMR gaps that favor light-skin samples and thus exacerbating the pre-existing racial bias. Consistently, ΔASR values are large and strongly favor the dark-skin group, which in turn leads to corresponding increases in ΔACC, amplifying the light-skin advantage already seen in the *natural setting* (Table 2). Among the CLIP-based models, $\text{FairCLIP}_{\text{R}}$ exhibits the smallest ASR disparities, but this comes at the cost of the lowest overall accuracy, since its debiasing mechanism systematically favors dark-skin samples, even under attack. Regarding ΔAMR, under $\mathcal{D}_{70\%}$ the bias in favor of light-skin samples decreases, as dark-skin images form the majority of the IRD. In contrast, the ΔACC gap in favor of light-skin queries tends to increase: with fewer light-skin images available in the IRD, light-skin queries compete with a smaller pool of same-attribute candidates and are less affected by the attack, unlike the dark-skin group. For RET@5, ASR values increase substantially for all models and polarizations, as the attack succeeds whenever the poisoned image appears anywhere within the top five retrieved results. Again in this case, the influence of legitimate retrieved samples re-emerges: the remaining four items follow the same distribution and bias patterns observed in the *natural setting*.

Finally, Figs. 2 and 3 present qualitative examples of attacked images using `FairLess` on the CLIP_{ViT} and its mitigated variant FairCLIP under the ℓ_∞ attack with $\epsilon = \frac{4}{255}$, corresponding to the highest perturbation budget considered for gender and race, respectively.

From Figs. 2 and 3, it is possible to observe that the poisoned samples remain semantically aligned with the concept expressed in the user query qu and therefore with the corresponding ground-truth image, while depicting the opposite sensitive attribute as required by the attack objective. The adversarial perturbation is imperceptible to the human eye, as illustrated by the pixel-wise difference map, which is amplified by a factor of 10.

8.2.2. *Effect of semantic augmentation*

In this section, we study the influence of the semantic augmentation strategy used to construct S in Phase 2 of `FairLess` (see Section 6.2). Table 7 summarizes the results for CLIP_{ViT} across both gender and race.

We first consider the baseline scenario $qt = qu$ (denoted as "="), where the attacker has full knowledge of the user query. As expected, this configuration yields the highest ASR across all polarizations and perturbation norms, since the optimization directly targets the exact query used at test time. When moving to the more realistic scenario where $qt \neq qu$, the attack becomes substantially more challenging. In this setting, the adversary has only partial knowledge of the concepts that will be queried at test time, because the target query qt used during optimization does not coincide with the actual user query qu. In this setting, DT emerges as the most effective augmentation strategy, consistently yielding the highest ASR across both gender and race and for all IRD polarizations. This result is expected: although the captions used in S are not identical to the user query, they originate from MSCOCO and therefore supply high-quality, diverse, and naturally aligned semantic variations of the target concept. Regarding the most challenging configuration for the attacker, namely $\emptyset$, where S contains only the target query qt, ASR values drop substantially, particularly under the small-budget ℓ_2 attack. This reflects the inherent difficulty of capturing the semantic variability of unseen user queries when no augmentation is provided. LLM-based augmentation improves over the $\emptyset$ configuration by enriching S with

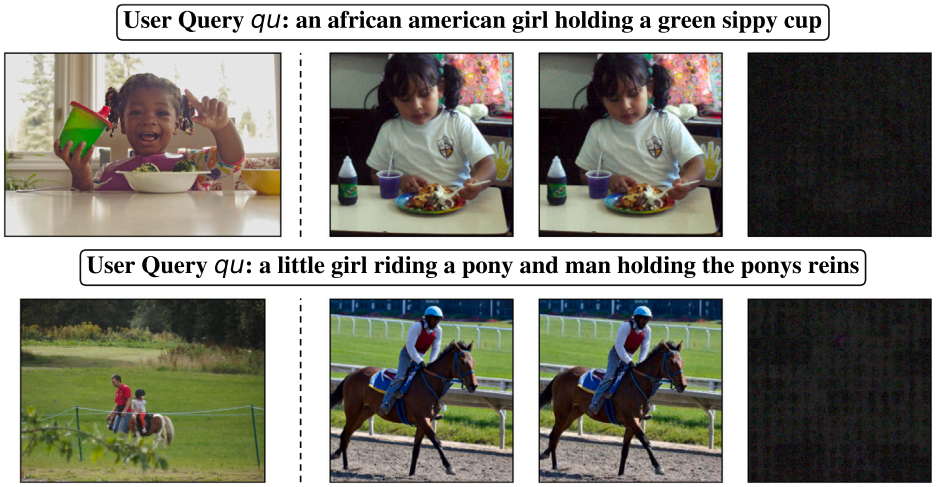

Fig. 3. Poisoned samples for race polarization under the ℓ_∞ attack ($\varepsilon = \frac{4}{255}$) with the DT augmentation strategy. The top row corresponds to CLIP_{ViT} [4] (attacking dark-skin concepts), and the bottom row to $\text{FairCLIP}_{\text{R}}$ [37] (attacking light-skin concepts). Each row includes the user query qu, the ground-truth image, the clean and poisoned targets, and the pixel-wise difference (amplified by a factor of 10).

Table 7
Results of the adversarial setting for both gender and race polarization using CLIP_{ViT} [4], with different augmentation strategies indicated by AUG. ASR at 1 and 5 are reported under ℓ_∞ ($\varepsilon = \frac{4}{255}$) and ℓ_2 ($\varepsilon = 0.05$). POL denotes the IRD polarization: $\mathcal{D}_{50\%}$ (intrinsic model bias) and $\mathcal{D}_{30\%}$, $\mathcal{D}_{70\%}$ (dataset-induced bias), where % refers to male (gender) and dark-skin (race) samples. Best values for ASR are highlighted in bold within each polarization setting.

POL	AUG	GENDER				RACE			
		ℓ_∞		ℓ_2		ℓ_∞		ℓ_2	
		ASR@1 ↑	ASR@5 ↑	ASR@1 ↑	ASR@5 ↑	ASR@1 ↑	ASR@5 ↑	ASR@1 ↑	ASR@5 ↑
$\mathcal{D}_{50\%}$	=	**100.0 ± 0.0**	**100.0 ± 0.0**	**76.8 ± 2.5**	**98.5 ± 0.3**	**100.0 ± 0.0**	**100.0 ± 0.0**	81.4 ± 0.6	99.6 ± 0.3
	DT	99.2 ± 0.1	99.7 ± 0.1	43.1 ± 0.5	74.6 ± 0.8	99.4 ± 0.1	99.7 ± 0.1	54.4 ± 0.2	89.4 ± 0.6
	∅	94.0 ± 1.6	96.0 ± 0.9	29.6 ± 0.9	57.2 ± 2.1	96.0 ± 0.5	98.0 ± 0.3	44.0 ± 0.9	78.1 ± 1.5
	LLM_3	94.6 ± 0.7	96.5 ± 0.7	29.6 ± 0.8	58.0 ± 2.4	96.4 ± 0.6	98.3 ± 0.5	43.9 ± 1.4	78.0 ± 1.2
	LLM_{10}	95.2 ± 0.8	96.7 ± 0.9	29.6 ± 0.7	58.1 ± 2.2	96.9 ± 0.7	98.5 ± 0.4	44.5 ± 1.1	78.0 ± 1.2
$\mathcal{D}_{30\%}$	=	**100.0 ± 0.0**	**100.0 ± 0.0**	**76.1 ± 2.7**	**99.0 ± 0.0**	**100.0 ± 0.0**	**100.0 ± 0.0**	**82.7 ± 1.7**	**99.6 ± 0.3**
	DT	99.3 ± 0.1	99.5 ± 0.1	43.8 ± 0.6	75.3 ± 0.7	99.4 ± 0.7	99.8 ± 0.3	56.3 ± 2.6	88.3 ± 1.0
	∅	93.4 ± 1.4	95.8 ± 0.7	30.0 ± 0.7	57.6 ± 1.4	96.1 ± 0.2	98.2 ± 0.5	45.6 ± 4.3	78.3 ± 0.6
	LLM_3	94.2 ± 1.2	96.4 ± 1.1	30.0 ± 0.9	58.5 ± 1.0	96.7 ± 0.5	98.5 ± 0.7	45.6 ± 3.7	78.6 ± 1.0
	LLM_{10}	94.7 ± 0.9	96.7 ± 0.9	30.0 ± 0.8	58.9 ± 0.9	96.9 ± 0.6	98.8 ± 0.8	46.0 ± 3.9	78.9 ± 1.1
$\mathcal{D}_{70\%}$	=	**100.0 ± 0.0**	**100.0 ± 0.0**	**77.5 ± 2.0**	**98.1 ± 0.5**	**100.0 ± 0.0**	**100.0 ± 0.0**	**83.6 ± 1.6**	**100.0 ± 0.0**
	DT	99.2 ± 0.1	99.8 ± 0.1	44.1 ± 0.8	75.3 ± 1.0	99.4 ± 0.3	99.7 ± 0.1	56.2 ± 0.8	89.7 ± 0.3
	∅	94.0 ± 1.6	96.1 ± 0.9	30.4 ± 1.8	59.0 ± 2.4	95.8 ± 0.9	98.0 ± 0.5	44.5 ± 1.5	78.8 ± 0.9
	LLM_3	94.5 ± 0.8	96.7 ± 0.6	30.5 ± 1.5	59.6 ± 2.5	96.0 ± 0.7	98.2 ± 0.8	44.3 ± 2.2	78.9 ± 1.0
	LLM_{10}	95.3 ± 0.8	96.8 ± 0.9	30.4 ± 1.7	59.9 ± 2.2	96.2 ± 0.6	98.1 ± 0.9	45.4 ± 1.6	79.3 ± 1.1

paraphrased variants of qt. Under the ℓ_∞ attack, both LLM_3 and LLM_{10} yield higher ASR than ∅, with LLM_{10} providing a slightly higher improvement thanks to the greater diversity of generated captions. Under the ℓ_2 attack with a minimal perturbation budget, the improvements introduced by LLM-based augmentation remain limited. This can be attributed to the extremely small perturbation radius used in this setting.

To support this observation, Fig. 4 provides a more fine-grained analysis of how augmentation interacts with the perturbation budget. Fig. 4 reports ASR@1 for both gender and race scenarios across five attack configurations: ℓ_∞ with $\varepsilon = \frac{4}{255}$ and ℓ_2 with $\varepsilon \in \{0.05, 0.3, 0.5, 1.0\}$. In particular, Fig. 4 shows the ASR@1 obtained under the three augmentation strategies (∅, LLM3, LLM10), together with the corresponding improvements of LLM-based augmentation over the ∅ baseline, expressed through $\text{ASR@1}(LLM_3) - \text{ASR@1}(\emptyset)$ and $\text{ASR@1}(LLM_{10}) - \text{ASR@1}(\emptyset)$. For ℓ_2 attacks, as expected, increasing the perturbation budget leads to higher ASR values. More importantly, Fig. 4 shows that larger budgets also make LLM-based augmentation progressively more effective, with stronger perturbations better exploiting the additional semantic variability introduced by the generated captions.

8.2.3. *Influence of target selection position*

In this section, we investigate how the similarity, in terms of raked position, of the target image with respect to the target query affects the effectiveness of `FairLess`, when a single poisoned image is inserted into the IRD. In particular, we compare

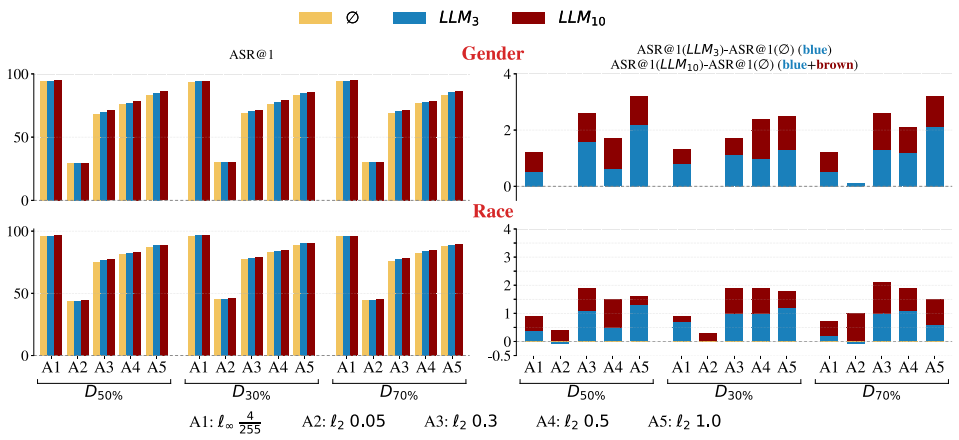

Fig. 4. Adversarial evaluation on CLIP_{ViT} [4] under gender and race scenarios across multiple attack scenarios: ℓ_∞ ($\epsilon = \frac{4}{255}$) and ℓ_2 ($\epsilon \in \{0.05, 0.3, 0.5, 1.0\}$). The left column reports ASR@1 under three augmentation strategies (∅, LLM_3, LLM_{10}), while the right column shows the corresponding ASR@1 differences with respect to ∅, with LLM_3 in blue and LLM_{10} in blue + brown. $\mathcal{D}$ denotes the IRD polarization: $\mathcal{D}_{50\%}$ (intrinsic model bias) and $\mathcal{D}_{30\%}$, $\mathcal{D}_{70\%}$ (dataset-induced bias), where % refers to male (gender) and dark-skin (race) samples.

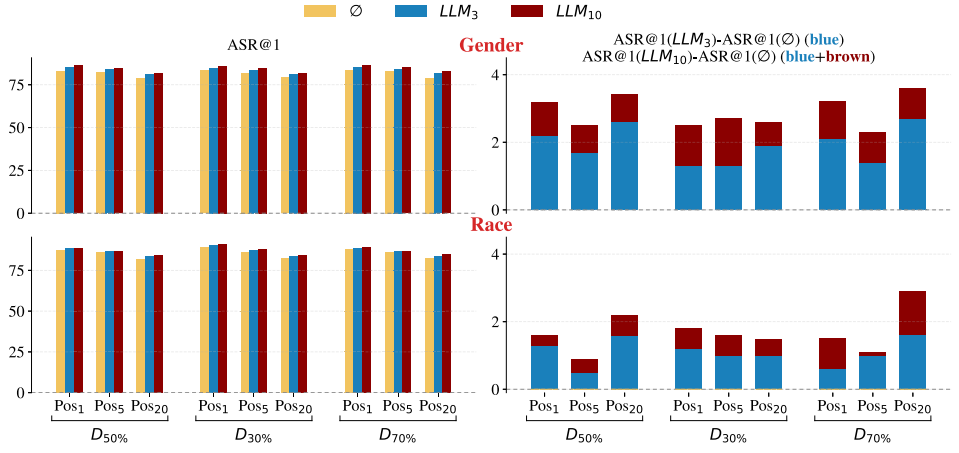

Fig. 5. Adversarial evaluation on CLIP_{ViT} [4] under ℓ_2 attacks with $\epsilon = 1.0$ for gender and race scenarios. The left column reports ASR@1 under three augmentation strategies (∅, LLM_3, LLM_{10}), while the right column shows the corresponding ASR@1 differences with respect to ∅, with LLM_3 in blue and LLM_{10} in blue + brown. Each group corresponds to a target-position setting (Pos), where the 1st, 5th, and 20th most similar opposite-attribute images are selected as poisoning targets. $\mathcal{D}$ denotes the IRD polarization: $\mathcal{D}_{50\%}$ (intrinsic model bias) and $\mathcal{D}_{30\%}$, $\mathcal{D}_{70\%}$ (dataset-induced bias), where % refers to male (gender) and dark-skin (race) samples.

the top-ranked target obtained through Eq. (15) (see Section 6.1) with the 5th and 20th most similar opposite-attribute images, in order to assess whether the attack remains influential even when the chosen target is progressively less aligned with the query concept. For this study, we focus on the more challenging scenario where $qt \neq qu$, so the adversary has only partial knowledge of the concepts that will be queried at test time. Accordingly, we evaluate the baseline ∅ configuration together with the two LLM-based augmentation strategies (LLM_3 and LLM_{10}), which are used to construct S during Phase 2 of the attack (see Section 6.2). Finally, we perform this analysis on CLIP_{ViT} under the ℓ_2 attack with $\epsilon = 1.0$, thus a larger perturbation compared with the minimal one considered ($\epsilon = 0.05$), which allows us to more clearly assess the influence of semantic augmentation.

Fig. 5 shows how the attack outcome varies with the target-selection position. The figure reports the ASR@1 for both gender and race scenarios and shows the results obtained for the three augmentation strategies already introduced. It also includes the corresponding variations relative to the ∅ baseline, expressed as $\text{ASR@1}(LLM_3) - \text{ASR@1}(\emptyset)$ and $\text{ASR@1}(LLM_{10}) - \text{ASR@1}(\emptyset)$. Each group corresponds to one of the target-position settings (Pos), where the 1st, 5th, and 20th most similar opposite-attribute images were selected as poisoning targets.

The results indicate that the attack remains effective even when the poisoning target is not the top-ranked opposite-attribute image. Across all tested positions, `FairLess` still manages to push the poisoned sample into the top retrieved result, although with progressively weaker impact as the target moves down the similarity ranking. This trend is visible in the drop in ASR@1 from

Table 8
Results of the adversarial setting with multiple injections for gender polarization under ℓ_∞ ($\epsilon = \frac{4}{255}$) and ℓ_2 ($\epsilon = 0.05$), using the DT augmentation strategy. For each VLP, results are shown at the top-1 and 5 ranks, where ASR measures the absence of the legitimate image in the top-k. POL denotes the IRD polarization: $\mathcal{D}_{50\%}$ (intrinsic model bias) and $\mathcal{D}_{30\%}$, $\mathcal{D}_{70\%}$ (dataset-induced bias), where % refers to male samples. Δs denote absolute gaps between groups, with values favoring female and male groups. Best values for ACC and ASR are highlighted in bold within each polarization setting across VLPs.

POL	VLP	RET@5 – ℓ_∞ ($\epsilon = \frac{4}{255}$)					RET@5 – ℓ_2 ($\epsilon = 0.05$)				
		ACC ↑	ΔACC ↓	ASR ↑	ΔASR ↓	ΔAMR ↓	ACC ↑	ΔACC ↓	ASR ↑	ΔASR ↓	ΔAMR ↓
$\mathcal{D}_{50\%}$	CLIP_{ViT} [4]	0.1 ± 0.1	0.2 ± 0.2	99.9 ± 0.1	0.2 ± 0.2	0.1 ± 0.0	$\mathbf{40.3 \pm 2.1}$	2.9 ± 3.0	59.7 ± 2.1	2.9 ± 3.0	2.7 ± 1.6
	CLIP_{CNN} [4]	$\mathbf{0.3 \pm 0.2}$	0.3 ± 0.1	99.7 ± 0.2	0.3 ± 0.1	0.3 ± 0.2	37.0 ± 1.1	4.5 ± 2.4	$\mathbf{63.0 \pm 1.1}$	4.5 ± 2.4	1.1 ± 1.1
	$\text{FairCLIP}_{\text{G}}$ [37]	0.0 ± 0.0	0.0 ± 0.0	$\mathbf{100.0 \pm 0.0}$	0.0 ± 0.0	0.0 ± 0.0	38.8 ± 1.4	3.9 ± 2.5	61.2 ± 1.4	3.9 ± 2.5	0.8 ± 1.1
	$\text{CLIP-DEB}_{\text{G}}$ [17]	0.1 ± 0.1	0.2 ± 0.2	99.9 ± 0.1	0.2 ± 0.2	0.0 ± 0.1	38.6 ± 2.2	5.7 ± 1.0	61.4 ± 2.2	5.7 ± 1.0	6.8 ± 1.4
$\mathcal{D}_{30\%}$	CLIP_{ViT} [4]	0.1 ± 0.2	0.3 ± 0.3	99.9 ± 0.2	0.3 ± 0.3	0.3 ± 0.3	$\mathbf{40.8 \pm 0.7}$	0.9 ± 1.4	59.2 ± 0.7	0.9 ± 1.4	15.3 ± 1.6
	CLIP_{CNN} [4]	$\mathbf{0.2 \pm 0.2}$	0.2 ± 0.2	99.8 ± 0.2	0.2 ± 0.2	0.3 ± 0.2	37.8 ± 0.8	0.9 ± 1.4	$\mathbf{62.2 \pm 0.8}$	0.9 ± 1.4	10.6 ± 0.7
	$\text{FairCLIP}_{\text{G}}$ [37]	0.0 ± 0.0	0.0 ± 0.0	$\mathbf{100.0 \pm 0.0}$	0.0 ± 0.0	0.0 ± 0.0	39.9 ± 1.0	3.3 ± 2.5	60.1 ± 1.0	3.3 ± 2.5	14.6 ± 1.8
	$\text{CLIP-DEB}_{\text{G}}$ [17]	0.1 ± 0.1	0.2 ± 0.2	99.9 ± 0.1	0.2 ± 0.2	0.2 ± 0.2	38.7 ± 0.7	2.5 ± 2.1	61.3 ± 0.7	2.5 ± 2.1	19.9 ± 1.2
$\mathcal{D}_{70\%}$	CLIP_{ViT} [4]	0.1 ± 0.1	0.2 ± 0.2	99.9 ± 0.1	0.2 ± 0.2	0.2 ± 0.2	$\mathbf{41.4 \pm 1.4}$	7.7 ± 3.1	58.6 ± 1.4	7.7 ± 3.1	9.9 ± 2.3
	CLIP_{CNN} [4]	$\mathbf{0.3 \pm 0.2}$	0.3 ± 0.2	99.7 ± 0.2	0.3 ± 0.2	0.1 ± 0.1	37.3 ± 1.1	7.7 ± 3.6	$\mathbf{62.7 \pm 1.1}$	7.7 ± 3.6	13.3 ± 0.8
	$\text{FairCLIP}_{\text{G}}$ [37]	0.0 ± 0.0	0.0 ± 0.0	$\mathbf{100.0 \pm 0.0}$	0.0 ± 0.0	0.0 ± 0.0	39.6 ± 1.4	9.1 ± 2.3	60.4 ± 1.4	9.1 ± 2.3	15.4 ± 0.6
	$\text{CLIP-DEB}_{\text{G}}$ [17]	0.1 ± 0.1	0.2 ± 0.2	99.9 ± 0.1	0.0 ± 0.0	0.2 ± 0.2	39.8 ± 1.7	10.9 ± 0.5	60.2 ± 1.7	10.9 ± 0.5	6.1 ± 1.7

Pos_1 to Pos_{20}, which corresponds to the growing semantic mismatch between the target image and the user query. However, we highlight that choosing the top-ranked opposite-attribute match benefits the attacker not only in terms of effectiveness but also in terms of perceived coherence: this choice maximizes semantic alignment with the target concept, making the poisoned result appear more coherent and reinforcing the user's perception of the underlying bias. Regarding augmentation, the effects vary with the target position. At Pos_1, the target image is already strongly aligned with the concept, so LLM-based augmentation mainly serves to refine the semantic information already available in the image. At Pos_5, where the target is less aligned than in Pos_1, the additional paraphrased queries offer only limited benefit, as the semantic gap remains too large for the augmentation to substantively refine the alignment. At Pos_{20}, the situation changes. The target is substantially less aligned with the concept, and here the LLM-generated augmentations become more useful, helping mitigate the semantic mismatch induced by the suboptimal target position.

8.2.4. Impact of multiple poisoned insertions

Finally, we study the effect of `FairLess` when multiple poisoned images are inserted into the IRD for each concept associated with a query. In this setting, the targets are selected according to Eq. (15) (see Section 6.1), by taking the five top-ranked images that are most similar to the query concept while belonging to the opposite sensitive-attribute group. Following the procedure used for the single poisoned insertion analysis (see Section 8.2.1), we adopt DT to generate $\mathcal{S}$ as the semantic augmentation strategy used in Phase 2 of the attack (see Section 6.2). The evaluation is performed on the CLIP-based models under both an ℓ_∞ attack with $\epsilon = \frac{4}{255}$ and an ℓ_2 attack with $\epsilon = 0.05$. We focus on this model architecture because it is widely used in T2IR, offers an effective accuracy–cost balance, and enables comparison with fairness-mitigated variants. Moreover, evaluating multiple poisoned insertions on heavier models such as BLIP would be substantially more computationally demanding. Importantly, unlike the single-insertion analysis, the definition of ASR changes in this setting: the attack is now considered successful when the ground-truth (legitimate) image does not appear among the top-5 retrieved results, due to the presence of multiple poisoned images in the IRD. For this reason, we report only RET@5 results.

Tables 8 and 9 report the results of the multiple–poisoned-insertions scenario for the gender- and race-annotated MSCOCO datasets, respectively.

Table 8 reports the results of the multiple–poisoned insertions scenario for the gender-annotated MSCOCO. Under the ℓ_∞ attack, ASR is essentially 100% for all models and across all IRD polarizations, meaning that none of the legitimate images remain in the top-5 results. Consequently, accuracy collapses to zero and all Δ values also become negligible. This pattern is consistent with the behavior observed in the single-insertion ℓ_∞ case (Table 3), where the attack is strong enough to dominate both accuracy and group disparities, leaving no clearly visible bias pattern. Under the ℓ_2 attack with a minimal perturbation budget ($\epsilon = 0.05$), the success criterion becomes considerably stricter: the attack is not always able to suppress the legitimate image, which allows us to better study how group-related effects re-emerge in this setting. Compared with the single-insertion results (Table 4), ΔAMR preserves the same qualitative trends across all IRD polarizations, while ΔACC gaps are generally attenuated but retain their original direction whenever a clear bias was already present. This indicates that moving from one to five poisoned samples does not fundamentally change the accuracy-based bias patterns, but rather amplifies or reduces them depending on the underlying IRD composition. Regarding ΔASR, the multi-insertion setting reveals a different behavior compared to the single-poison case. For the balanced IRD $\mathcal{D}_{50\%}$, ΔASR values become larger and are mostly male-favored, indicating that male queries are now more likely to lose their ground-truth image from the top-5. Under $\mathcal{D}_{30\%}$, where the single-insertion setting exhibited strong male-favored ΔASR gaps, these disparities become substantially smaller. Finally, for $\mathcal{D}_{70\%}$ we observe an inversion with respect to the single-poison scenario: ΔASR becomes male-favored, meaning that, with five poisoned samples, the attack more frequently suppresses the ground-truth image for male queries. This behavior is consistent with the *natural setting* results in Table 1. Since ASR is defined solely by the presence or absence of the

Table 9
Results of the adversarial setting with multiple injections for race polarization under ℓ_∞ ($\epsilon = \frac{4}{255}$) and ℓ_2 ($\epsilon = 0.05$), using the DT augmentation strategy. For each VLP, results are shown at the top-1 and 5 ranks, where ASR measures the absence of the legitimate image in the top-k. POL denotes the IRD polarization: $\mathcal{D}_{50\%}$ (intrinsic model bias) and $\mathcal{D}_{30\%}$, $\mathcal{D}_{70\%}$ (dataset-induced bias), where % refers to dark-skin samples. Δs denote absolute gaps between groups, with values favoring light-skin and dark-skin groups. Best values for ACC and ASR are highlighted in bold within each polarization setting across VLPs.

POL	VLP	RET@5 – ℓ_∞ ($\epsilon = \frac{4}{255}$)					RET@5 – ℓ_2 ($\epsilon = 0.05$)				
		ACC ↑	ΔACC ↓	ASR ↑	ΔASR ↓	ΔAMR ↓	ACC ↑	ΔACC ↓	ASR ↑	ΔASR ↓	ΔAMR ↓
$\mathcal{D}_{50\%}$	CLIP_{ViT} [4]	**0.3 ± 0.2**	0.1 ± 0.3	99.7 ± 0.2	0.1 ± 0.3	0.1 ± 0.3	**48.0 ± 1.8**	18.5 ± 4.4	52.0 ± 1.8	18.5 ± 4.4	11.5 ± 1.0
	CLIP_{CNN} [4]	0.2 ± 0.3	0.2 ± 0.3	99.8 ± 0.3	0.2 ± 0.3	0.1 ± 0.0	46.5 ± 1.1	17.5 ± 4.9	53.5 ± 1.1	17.5 ± 4.9	13.5 ± 1.8
	$\text{FairCLIP}_{\text{R}}$ [37]	0.0 ± 0.0	0.0 ± 0.0	**100.0 ± 0.0**	0.0 ± 0.0	0.0 ± 0.0	25.5 ± 1.1	4.0 ± 2.9	**74.5 ± 1.1**	4.0 ± 2.9	2.8 ± 1.3
	$\text{CLIP-DEB}_{\text{R}}$ [17]	**0.3 ± 0.2**	0.2 ± 0.2	99.7 ± 0.2	0.2 ± 0.2	0.1 ± 0.1	46.9 ± 1.4	17.3 ± 3.4	53.1 ± 1.4	17.3 ± 3.4	11.1 ± 1.5
$\mathcal{D}_{30\%}$	CLIP_{ViT} [4]	0.2 ± 0.2	0.1 ± 0.2	99.8 ± 0.2	0.1 ± 0.2	0.2 ± 0.3	**46.1 ± 2.4**	17.1 ± 2.6	53.9 ± 2.4	17.1 ± 2.6	19.7 ± 0.1
	CLIP_{CNN} [4]	**0.3 ± 0.1**	0.0 ± 0.2	99.7 ± 0.1	0.0 ± 0.2	0.1 ± 0.2	45.2 ± 2.2	18.7 ± 1.4	54.8 ± 2.2	18.7 ± 1.4	24.0 ± 2.1
	$\text{FairCLIP}_{\text{R}}$ [37]	0.0 ± 0.0	0.0 ± 0.0	**100.0 ± 0.0**	0.0 ± 0.0	0.0 ± 0.0	24.9 ± 0.7	4.0 ± 3.5	**75.1 ± 0.7**	4.0 ± 3.5	13.7 ± 5.9
	$\text{CLIP-DEB}_{\text{R}}$ [17]	0.2 ± 0.1	0.1 ± 0.1	99.9 ± 0.1	0.1 ± 0.1	0.2 ± 0.2	45.1 ± 1.7	16.2 ± 4.0	54.9 ± 1.7	16.2 ± 4.0	19.3 ± 0.4
$\mathcal{D}_{70\%}$	CLIP_{ViT} [4]	**0.3 ± 0.3**	0.1 ± 0.1	99.7 ± 0.3	0.1 ± 0.1	0.0 ± 0.0	**46.5 ± 0.8**	20.4 ± 2.4	53.5 ± 0.8	20.4 ± 2.4	3.3 ± 2.5
	CLIP_{CNN} [4]	0.2 ± 0.3	0.0 ± 0.1	99.8 ± 0.3	0.0 ± 0.1	0.0 ± 0.1	46.1 ± 1.2	23.3 ± 2.4	53.9 ± 1.2	23.3 ± 2.4	5.5 ± 1.6
	$\text{FairCLIP}_{\text{R}}$ [37]	0.0 ± 0.0	0.0 ± 0.0	**100.0 ± 0.0**	0.0 ± 0.0	0.0 ± 0.0	24.8 ± 1.4	6.6 ± 5.3	**75.2 ± 1.4**	6.6 ± 5.3	4.4 ± 4.0
	$\text{CLIP-DEB}_{\text{R}}$ [17]	**0.3 ± 0.4**	0.1 ± 0.1	99.7 ± 0.4	0.1 ± 0.1	0.0 ± 0.0	46.2 ± 1.1	19.7 ± 1.8	53.8 ± 1.1	19.7 ± 1.8	2.4 ± 1.8

ground-truth image among the top-5 retrieved items, it does not require that the replacement items belong to the opposite sensitive group. As a consequence, groups that already exhibit lower accuracy in the *natural setting* are inherently more prone to having their ground-truth image pushed out of the top-5 under attack. In this setting, for $\mathcal{D}_{50\%}$ and $\mathcal{D}_{70\%}$, ΔACC@5 consistently favors female queries, so male ground-truth images are intrinsically harder for the model to retrieve correctly and thus more often removed from the top-5, leading to male-favored ΔASR. Conversely, under $\mathcal{D}_{30\%}$, where ΔACC@5 favors the male group, this natural robustness advantage counterbalances the attack asymmetry, driving ΔASR values back toward neutrality in this context.

Table 9 reports the results of the multiple poisoned insertions scenario for the race-annotated MSCOCO. Under the ℓ_∞ attack, the behavior is analogous to the gender case: ASR is essentially 100% for all models and IRD polarizations, indicating that the ground-truth image almost never remains in the top-5. As a consequence, accuracy drops to nearly zero and all Δ metrics become negligible. Under the ℓ_2 attack with a minimal perturbation budget ($\epsilon = 0.05$), the attack is not always able to suppress the legitimate image. In this case, the behavior of ΔAMR is broadly consistent across polarizations: for most models and IRD configurations, the gaps are light-skin–favored, increasing the proportion of same-attribute matches for this group. The only exception is $\text{FairCLIP}_{\text{R}}$ under $\mathcal{D}_{70\%}$, which still exhibits a dark-skin–favored gap. This behavior is consistent with the single poisoned insertion setting (Table 6), where its debiasing mechanism reduces racial disparities at the cost of lower overall accuracy. At the same time, ΔACC gaps become larger but remain consistently light-skin–favored across all polarizations, reinforcing the accuracy advantage already present in the *natural setting* (Table 2). The most notable effect of multi-insertion appears in ΔASR. For all CLIP-based models and IRD polarizations, ΔASR is strongly dark-skin–favored, with magnitudes that are considerably higher than in the single-poison case. This indicates that dark-skin queries are systematically more likely to lose their ground-truth image from the top-5 when multiple poisoned images are available in the IRD. Importantly, this behavior aligns with the *natural setting* results in Table 2, because ASR only depends on whether the ground-truth image is absent from the top-5 and does not require that the replacement items belong to the opposite race group. As a consequence, groups that are already less accurate in the *natural setting* are structurally more likely to have their ground-truth image displaced under attack.

8.2.5. Defensive strategy on `FairLess`

We evaluate the effect of JPEG compression [33] as a defense against `FairLess`. For this analysis, we adopt the same single-injection setting introduced in Section 8.2.1 and report results at RET@1. In particular, we consider the ℓ_∞ attack with $\epsilon = \frac{4}{255}$, focusing on the balanced configuration $\mathcal{D}_{50\%}$ and adopting the DT augmentation strategy, which provides a representative and stable attack configuration. JPEG compression is applied at inference time with different quality factors (QFs), where lower values correspond to stronger compression. The row denoted as "Cl." represents the original adversarial configuration without compression.

Table 10 reports the results for both gender and race. For the gender setting, CLIP-DEB refers to $\text{CLIP-DEB}_{\text{G}}$, while for the race setting it refers to $\text{CLIP-DEB}_{\text{R}}$.

We first consider the gender case. For CLIP_{ViT}, the "Cl." configuration exhibits ASR values close to 100%, effectively driving ACC to zero, consistently with the behavior already observed in Table 3. As JPEG compression is applied, ASR progressively decreases, reaching 44.3% at QF = 89, while ACC correspondingly increases up to 17.5%. This indicates that compression weakens the adversarial perturbation, reducing its ability to consistently place the poisoned sample at the top rank. In this configuration, the attack remains almost symmetric across groups, as reflected by the near-zero values of ΔACC and ΔASR in the "Cl." setting. As compression becomes stronger, ΔACC slightly shifts in favor of female queries, in line with the tendency observed in the *natural* setting reported in Table 1.

A similar overall trend is observed for CLIP-DEB. In the "Cl." configuration, ASR is again close to 100%, with ACC near zero. Under compression, ASR decreases to 51.6% at QF = 89, while ACC increases up to 15.0%. However, compared with CLIP_{ViT}, group

Table 10
Effect of JPEG compression as a defense under the ℓ_∞ attack ($\varepsilon = \frac{4}{255}$), using the DT augmentation strategy. For each VLP model, results are reported at top-1 rank. "Cl." denotes the original adversarial configuration, while QF controls the compression strength, with lower values corresponding to stronger compression. Δs denote absolute gaps between groups, with values favoring female and male for gender, and light-skin and dark-skin for race. The best ACC and ASR values are highlighted in bold for each VLP model and each setting.

VLP	QF	GENDER				RACE			
		ACC ↑	ΔACC ↓	ASR ↑	ΔASR ↓	ACC ↑	ΔACC ↓	ASR ↑	ΔASR ↓
$CLIP_{ViT}$ [4]	Cl.	0.0 ± 0.1	0.1 ± 0.1	$\mathbf{99.2 \pm 0.1}$	0.1 ± 0.2	0.2 ± 0.2	0.2 ± 0.3	$\mathbf{99.4 \pm 0.1}$	0.1 ± 0.4
	99	3.4 ± 0.6	0.1 ± 0.1	88.9 ± 1.5	0.4 ± 0.6	3.7 ± 0.5	2.8 ± 1.0	92.4 ± 0.8	1.9 ± 1.6
	94	13.2 ± 0.7	0.8 ± 1.5	61.9 ± 1.2	3.0 ± 2.4	14.3 ± 0.3	10.1 ± 1.6	71.5 ± 1.3	8.5 ± 2.8
	89	$\mathbf{17.5 \pm 0.9}$	1.2 ± 0.8	44.3 ± 0.9	0.2 ± 0.8	$\mathbf{21.0 \pm 1.6}$	12.7 ± 1.4	57.8 ± 1.3	10.5 ± 0.6
CLIP-DEB [17]	Cl.	0.0 ± 0.0	0.0 ± 0.0	$\mathbf{99.7 \pm 0.1}$	0.3 ± 0.5	0.1 ± 0.1	0.0 ± 0.2	$\mathbf{99.6 \pm 0.1}$	0.1 ± 0.2
	99	2.6 ± 0.8	0.2 ± 0.7	91.0 ± 1.4	0.2 ± 1.2	2.9 ± 0.7	1.9 ± 0.6	94.5 ± 0.9	1.5 ± 0.5
	94	11.8 ± 0.6	0.5 ± 1.2	65.9 ± 1.0	0.9 ± 0.5	13.5 ± 0.9	7.5 ± 0.8	75.1 ± 0.8	7.0 ± 2.6
	89	$\mathbf{15.0 \pm 0.6}$	2.5 ± 2.1	51.6 ± 1.9	6.5 ± 2.5	$\mathbf{19.2 \pm 1.5}$	10.7 ± 1.5	63.6 ± 0.7	6.8 ± 1.1

disparities become more visible under stronger compression: ΔACC shifts more clearly in favor of female queries, while, at the strongest compression level, ΔASR shifts in favor of male queries. This indicates that the attack remains relatively more effective on male queries even after partial perturbation removal.

We next consider the race case. For $CLIP_{ViT}$, the "Cl." configuration again shows ASR close to 100% and near-zero ACC, as already observed in Table 5. As JPEG compression is applied, ASR decreases to 57.8% at QF = 89, while ACC increases up to 21.0%, indicating a partial recovery of retrieval performance. In contrast to the gender case, group disparities emerge more clearly as compression becomes stronger: ΔACC progressively shifts in favor of light-skin queries, while ΔASR increases in favor of dark-skin queries. This indicates that the attack remains relatively more effective on dark-skin queries even under compression.

The same overall behavior is observed for CLIP-DEB. Although ASR decreases and ACC increases under compression, ΔACC remains large and continues to favor light-skin queries. At the same time, ΔASR remains consistently in favor of dark-skin queries across QF values, confirming that this group remains more vulnerable to the attack even when the perturbation is partially weakened. Compared with $CLIP_{ViT}$, the debiased model shows slightly smaller, but still substantial, race disparities under compression.

Overall, JPEG compression partially mitigates `FairLess` in our experiments by reducing ASR and restoring part of the retrieval accuracy. However, the attack remains effective under compression, and non-negligible group disparities persist, especially in the race setting.

8.3. Statistical significance assessment

To complement the empirical comparisons reported in the previous subsections, we perform an additional statistical significance assessment of the observed fairness gaps in both the *natural* setting and the *adversarial* setting. For the latter, we consider the single-poisoned-insertion scenario as a representative adversarial configuration. The purpose of this analysis is to determine whether the group disparities discussed above are robust to query resampling, or whether they may instead be partly attributable to sampling variability.

To this end, we adopt a query-level bootstrap procedure, in which each textual query is treated as the statistical unit. When the same query occurs multiple times, its observations are first aggregated across seeds and, when applicable, across repeated retrieval instances; the corresponding metric values are then averaged before resampling. Although multiple textual queries may refer to the same underlying image, we treat queries as independent units, following standard T2IR evaluation protocols in which each query defines a distinct retrieval instance.

Bootstrap samples are generated by resampling queries with replacement within each sensitive group, thus implementing a stratified bootstrap procedure. For each resample, the corresponding inter-group difference is recomputed. For every configuration, we draw 10,000 bootstrap samples to estimate a 95% confidence interval for the group gap, and we report whether this interval excludes zero. This non-parametric approach avoids distributional assumptions on the metrics, which is particularly appropriate in our setting given the complex and non-Gaussian behavior of retrieval outcomes. In practice, configurations for which the 95% bootstrap confidence interval excludes zero are considered statistically significant at the 5% level under a two-sided test.

The analysis is deliberately conducted at the query level, rather than only across repeated runs, in order to directly assess the sensitivity of the observed gaps to the sampled queries, which constitute a primary source of variability in retrieval-based evaluation. Indeed, retrieval performance naturally depends on the specific queries considered, as some queries are intrinsically easier whereas others are more ambiguous. In the *adversarial* setting, attack effectiveness may further vary with the selected target images and with the extent to which individual query concepts can be manipulated. Consequently, metrics such as ACC and ASR are expected to be more sensitive to query-level variability. By contrast, AMR captures the sensitive-attribute composition of the retrieved set more directly, aggregating information over the top-k results rather than relying on a single success event, and empirically leads to more stable significance patterns.

Tables 11 and 12 summarize the results for gender and race in the natural and adversarial settings. For each model, metric, and retrieval cutoff, we report the number of evaluated configurations (*EXPS*) together with the percentage of cases in which

Table 11
Statistical significance assessment in the natural setting for both gender and race polarization. Values report the percentage of configurations for which the 95% query-level bootstrap confidence interval for the group gap excludes zero. EXPS denotes the number of configurations aggregated for each model.

VLP	EXPS	GENDER				RACE			
		RET@1		RET@5		RET@1		RET@5	
		ACC	AMR	ACC	AMR	ACC	AMR	ACC	AMR
$CLIP_{ViT}$ [4]	9	66.7	66.7	44.4	100.0	100.0	66.7	66.7	88.9
$CLIP_{CNN}$ [4]	9	66.7	88.9	66.7	88.9	100.0	66.7	100.0	100.0
FairCLIP [37]	9	55.6	88.9	55.6	66.7	22.2	100.0	0.0	100.0
CLIP-DEB [17]	9	55.6	100.0	66.7	100.0	100.0	66.7	100.0	100.0
BLIP-2 [5]	3	66.7	66.7	66.7	100.0	100.0	66.7	100.0	100.0
$BLIP\text{-}2_{ITM}$ [5]	3	66.7	66.7	66.7	66.7	33.3	66.7	33.3	100.0

Table 12
Statistical significance assessment in the natural setting for both gender and race polarization. Values report the percentage of configurations for which the 95% query-level bootstrap confidence interval for the group gap excludes zero. EXPS denotes the number of configurations aggregated for each model.

VLP	EXPS	GENDER						RACE					
		RET@1			RET@5			RET@1			RET@5		
		ACC	AMR	ASR	ACC	AMR	ASR	ACC	AMR	ASR	ACC	AMR	ASR
$CLIP_{ViT}$ [4]	21	23.8	66.7	47.6	66.7	100.0	61.9	76.2	71.4	71.4	100.0	66.7	66.7
$CLIP_{CNN}$ [4]	21	38.1	61.9	33.3	66.7	100.0	52.4	71.4	71.4	71.4	100.0	100.0	66.7
FairCLIP [37]	21	42.9	61.9	38.1	66.7	71.4	52.4	38.1	52.4	71.4	0.0	100.0	66.7
CLIP-DEB [17]	21	42.9	57.1	42.9	66.7	100.0	42.9	71.4	71.4	76.2	66.7	100.0	66.7
BLIP-2 [5]	3	0.0	66.7	66.7	66.7	66.7	33.3	66.7	33.3	0.0	100.0	100.0	33.3
$BLIP\text{-}2_{ITM}$ [5]	3	33.3	33.3	33.3	66.7	66.7	66.7	100.0	100.0	33.3	33.3	100.0	0.0

the corresponding bootstrap confidence interval excludes zero. For models specialized for a specific sensitive attribute, we use a unified notation in the tables for readability. Specifically, CLIP-DEB and FairCLIP refer to their gender-specific variants in the gender columns, namely $FairCLIP_G$ and $CLIP\text{-}DEB_G$, and to their race-specific variants in the race columns, namely $FairCLIP_R$ and $CLIP\text{-}DEB_R$.

Importantly, a value of 0.0 should not be interpreted as evidence that group disparities are absent. Rather, it indicates that, across the configurations aggregated in that entry, none of the estimated gaps is distinguishable from zero under query resampling. This may occur when empirical gaps are small, when their direction varies across configurations, or when the metric is strongly affected by query-level variability. We also note that, for BLIP-2 and $BLIP\text{-}2_{ITM}$, the number of available configurations is smaller than for the CLIP-based models, especially in the adversarial setting, due to the higher computational cost of running attacks on these architectures. Consequently, the percentages reported for BLIP-based models should be interpreted with additional caution.

Overall, the results are consistent with the conclusions drawn from the main experimental analysis. The most stable and consistently significant disparities are generally observed for AMR, particularly at RET@5, since this metric directly reflects the sensitive-attribute composition of the retrieved set. In contrast, ACC and ASR typically exhibit lower significance rates, especially in the *adversarial* setting and at RET@1. ACC depends on exact retrieval success, which is inherently query-dependent, while ASR further depends on the effectiveness of the attack for each individual query-target pair. Consequently, both metrics are more sensitive to variations in query difficulty, semantic ambiguity, and attack strength.

A further distinction emerges between gender and race. In the race setting, significance rates are generally higher, particularly for CLIP-based models in the *adversarial* setting, consistently with the larger and more systematic disparities observed in the main results. By contrast, in the gender setting, the observed gaps are often smaller and more context-dependent.

Taken together, these findings suggest that the AMR-based fairness patterns discussed in the previous subsections are unlikely to be fully explained by query-sampling variability, providing additional empirical evidence of their robustness. At the same time, they show that the strength and stability of these patterns depend on the metric considered. In particular, the lower significance rates observed for ACC and ASR reflect their greater sensitivity to query-level variability, rather than necessarily implying the absence of meaningful empirical disparities.

9. Conclusions

In this work, we provided the first comprehensive analysis of fairness in Vision–Language Pretrained (VLP) Text-to-Image Retrieval (T2IR) systems under both *natural* and *adversarial* conditions. Our study shows that current T2IR pipelines are influenced not only by the intrinsic biases of the underlying VLPs but also by the demographic composition of the Image Retrieval Database

(IRD). By introducing *attribute-labeled queries* and proposing accuracy- and fairness-oriented retrieval metrics, we enabled a more precise evaluation of demographic disparities in realistic benchmarks such as MSCOCO.

Under the *natural setting*, we demonstrated that CLIP-based models exhibit clear demographic disparities, while BLIP-2 is generally more robust but still susceptible to IRD polarization. Importantly, we observed that fairness-mitigated variants of VLPs do not consistently reduce disparities and may even introduce unintended effects, such as over-correcting for certain demographic groups or amplifying biases when the IRD distribution is skewed.

To investigate fairness vulnerabilities in *adversarial setting*, we introduced `FairLess`, the first data poisoning attack specifically designed to manipulate demographic bias in T2IR. We formalized a white-box threat model and proposed a poisoning strategy that preserves semantic consistency while directing retrieval results toward a targeted demographic subgroup. Our experiments show that `FairLess` is highly effective across models, perturbation budgets, semantic augmentation strategies, and IRD polarizations. Even a *single* poisoned sample can be sufficient to introduce systematic bias, while multiple poisoned insertions can entirely suppress legitimate retrieval results. Furthermore, the attack remains effective even when the attacker does not know the exact user query, highlighting a critical vulnerability in real-world deployments. From a deployment perspective, these findings also suggest that enforcing stricter moderation or maintaining a highly curated image retrieval database could reduce the attack surface, thereby lowering the probability that poisoned samples are successfully indexed and retrieved.

A natural direction for future work is the development of dedicated mitigation strategies against fairness-oriented poisoning attacks in T2IR systems. While this work provides an initial assessment of an input-preprocessing defense, namely JPEG compression, our results indicate that such approaches only partially reduce the effectiveness of `FairLess`. More advanced defenses therefore remain a promising direction and warrant systematic investigation in the context of multimodal retrieval. Extending the analysis to black-box settings represents an important complementary direction. In addition, evaluating the proposed framework on additional datasets, including newly annotated benchmarks, or suitable large-scale, web-collected multimodal corpora, would help assess its generality across different retrieval scenarios. A further avenue for future work is the investigation of annotation noise, with particular attention to how imperfect demographic labels affect the robustness of fairness evaluations. Finally, developing a deeper mechanistic understanding of the interactions among debiasing procedures, dataset polarization, and adversarial perturbations may provide valuable insight into why fairness-oriented poisoning attacks emerge and how they can be mitigated.

Overall, our findings indicate that demographic fairness in T2IR is influenced by a combination of model biases, dataset imbalance, and susceptibility to adversarial manipulation. Mitigating these issues requires not only debiasing VLPs but also developing defenses resilient to poisoning and other adversarial strategies. Future work should focus on designing robust, fairness-aware retrieval mechanisms, developing certified defenses for multimodal retrieval, and expanding fairness evaluations to additional sensitive attributes and emerging VLP architectures.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: This work is partially supported by (i) project ELSA - European Lighthouse on Secure and Safe AI funded by the European Union's Horizon Europe under the grant agreement No. 101070617 (Luca Oneto), (ii) projects SERICS (PE00000014) and FAIR (PE00000013) under the NRRP MUR program funded by the EU - NGEU (Luca Oneto, Fabio Roli, and Davide Anguita), and (iii) project FISA-202300128 funded by the MUR program "Fondo italiano per le scienze applicate" (Fabio Roli).

Lastly, this work was carried out while Dario Lazzaro was enrolled in the Italian National Doctorate on Artificial Intelligence run by the Sapienza University of Rome in collaboration with the University of Genoa, funded by "Unione europea-Next Generation EU, Missione 4 Componente 1 CUP B53C23003580004".

Acknowledgments

This work is partially supported by (i) project ELSA - European Lighthouse on Secure and Safe AI funded by the European Union's Horizon Europe under the grant agreement No. 101070617, (ii) projects SERICS (PE00000014) and FAIR (PE00000013) under the NRRP MUR program funded by the EU - NGEU, and (iii) project FISA-2023-00128 funded by the MUR program "Fondo italiano per le scienze applicate". Lastly, this work was carried out while Dario Lazzaro was enrolled in the Italian National Doctorate on Artificial Intelligence run by the Sapienza University of Rome in collaboration with the University of Genoa, funded by "Unione europea-Next Generation EU, Missione 4 Componente 1 CUP B53C23003580004".

Data availability

All data and code are available at https://github.com/lazzd/FairLess.

References

[1] Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I, et al. Language models are unsupervised multitask learners. OpenAI Blog 2019.
[2] Sangkloy P, Jitkrittum W, Yang D, Hays J. A sketch is worth a thousand words: Image retrieval with text and sketch. In: European conference on computer vision. 2022.
[3] Liu Z, Rodriguez-Opazo C, Teney D, Gould S. Image retrieval on real-life images with pre-trained vision-and-language models. In: IEEE/CVF international conference on computer vision. 2021.
[4] Radford A, Kim JW, Hallacy C, Ramesh A, et al. Learning transferable visual models from natural language supervision. In: International conference on machine learning. 2021.
[5] Li J, Li D, Savarese S, Hoi S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International conference on machine learning. 2023.
[6] Liu Z, Li A, Xu J, Shi D. DCL-net: Dual-level correlation learning network for image–text retrieval. Comput Electr Eng 2025;122:110000.
[7] Wang J, Liu Y, Wang X. Are gender-neutral queries really gender-neutral? Mitigating gender bias in image search. In: Conference on empirical methods in natural language processing. 2021.
[8] Ali J, Kleindessner M, Wenzel F, Budhathoki K, Cevher V, Russell C. Evaluating the fairness of discriminative foundation models in computer vision. In: AAAI/ACM conference on AI, ethics, and society. 2023.
[9] Wang J, Zhang Y, Sang J. Fairclip: Social bias elimination based on attribute prototype learning and representation neutralization. 2022, arXiv preprint arXiv:2210.14562.
[10] Zhang J, Yi Q, Sang J. Towards adversarial attack on vision-language pre-training models. In: ACM international conference on multimedia. 2022.
[11] Lu D, Wang Z, Wang T, Guan W, Gao H, Zheng F. Set-level guidance attack: Boosting adversarial transferability of vision-language pre-training models. In: IEEE/CVF international conference on computer vision. 2023.
[12] Yang Z, He X, Li Z, Backes M, Humbert M, Berrang P, Zhang Y. Data poisoning attacks against multimodal encoders. In: International conference on machine learning. 2023.
[13] Xu Y, Yao J, Shu M, Sun Y, Wu Z, Yu N, Goldstein T, Huang F. Shadowcast: Stealthy data poisoning attacks against vision-language models. In: Neural information processing systems. 2024.
[14] Hu F, Chen A, Li X. Towards making a trojan-horse attack on text-to-image retrieval. In: IEEE international conference on acoustics, speech and signal processing. 2023.
[15] Kleindessner M, Donini M, Russell C, Zafar MB. Efficient fair PCA for fair representation learning. In: International conference on artificial intelligence and statistics. 2023.
[16] Berg H, Hall S, Bhalgat Y, Kirk H, Shtedritski A, Bain M. A prompt array keeps the bias away: Debiasing vision-language models with adversarial learning. In: Conference of the Asia–Pacific chapter of the association for computational linguistics and international joint conference on natural language processing (volume 1: long papers). 2022.
[17] Zhang H, Guo Y, Kankanhalli M. Joint vision-language social bias removal for CLIP. In: Computer vision and pattern recognition conference. 2025.
[18] Biggio B, Nelson B, Laskov P. Poisoning attacks against support vector machines. In: International conference on machine learning. 2012.
[19] Feng J, Cai Q-Z, Zhou ZH. Learning to confuse: Generating training time adversarial data with auto-encoder. In: Neural information processing systems. 2019.
[20] Koh PW, Liang P. Understanding black-box predictions via influence functions. In: International conference on machine learning. 2017.
[21] Shafahi A, Huang WR, Najibi M, Suciu O, Studer C, Dumitras T, Goldstein T. Poison frogs! targeted clean-label poisoning attacks on neural networks. In: International conference on neural information processing systems. 2018.
[22] Gu T, Dolan-Gavitt B, Garg S. Badnets: Identifying vulnerabilities in the machine learning model supply chain. 2017, arXiv preprint arXiv:1708.06733.
[23] Doan K, Lao Y, Zhao W, Li P. LIRA: Learnable, imperceptible and robust backdoor attacks. In: IEEE/CVF international conference on computer vision. 2021.
[24] Madry A, Makelov A, Schmidt L, Tsipras D, Vladu A. Towards deep learning models resistant to adversarial attacks. In: International conference on learning representations. 2018.
[25] Ebrahimi J, Rao A, Lowd D, Dou D. Hotflip: White-box adversarial examples for text classification. In: Annual meeting of the association for computational linguistics (volume 2: short papers). 2018.
[26] Papernot N, McDaniel P, Goodfellow I, Jha S, Celik ZB, Swami A. Practical black-box attacks against machine learning. In: Asia conference on computer and communications security. 2017.
[27] Demontis A, Melis M, Pintor M, Jagielski M, Biggio B, Oprea A, Nita-Rotaru C, Roli F. Why do adversarial attacks transfer? Explaining transferability of evasion and poisoning attacks. In: USENIX security symposium. 2019.
[28] Dehouche N. Implicit stereotypes in pre-trained classifiers. IEEE Access 2021;9:167936–47.
[29] Buolamwini J, Gebru T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In: Conference on fairness, accountability and transparency. 2018.
[30] De-Arteaga M, Romanov A, Wallach H, Chayes J, Borgs C, Chouldechova A, Geyik S, Kenthapadi K, Kalai AT. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In: Conference on fairness, accountability, and transparency. 2019.
[31] Solans D, Biggio B, Castillo C. Poisoning attacks on algorithmic fairness. In: Joint European conference on machine learning and knowledge discovery in databases. 2020.
[32] Ghosh A, Jagielski M, Wilson C. Subverting fair image search with generative adversarial perturbations. In: ACM conference on fairness, accountability, and transparency. 2022.
[33] Dziugaite GK, Ghahramani Z, Roy DM. A study of the effect of jpg compression on adversarial images. 2016, arXiv preprint arXiv:1608.00853.
[34] Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL. Microsoft coco: Common objects in context. In: European conference on computer vision. 2014.
[35] Zhao J, Wang T, Yatskar M, Ordonez V, Chang K-W. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In: Conference on empirical methods in natural language processing. 2017.
[36] Zhao D, Wang A, Russakovsky O. Understanding and evaluating racial biases in image captioning. In: International conference on computer vision. 2021.
[37] Luo Y, Shi M, Khan MO, Afzal MM, Huang H, Yuan S, Tian Y, Song L, Kouhana A, Elze T, et al. Fairclip: Harnessing fairness in vision-language learning. In: IEEE/CVF conference on computer vision and pattern recognition. 2024.
[38] Lee K-H, Chen X, Hua G, Hu H, He X. Stacked cross attention for image-text matching. In: European conference on computer vision. 2018.
[39] Plummer BA, Wang L, Cervantes CM, Caicedo JC, Hockenmaier J, Lazebnik S. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In: IEEE international conference on computer vision. 2015.
[40] Gao W, Kannan S, Oh S, Viswanath P. Estimating mutual information for discrete-continuous mixtures. In: Neural information processing systems. 2017.
[41] Zhang Y, Sang J. Towards accuracy-fairness paradox: Adversarial example-based data augmentation for visual debiasing. In: ACM international conference on multimedia. 2020.

[42] Angwin J, Larson J, Mattu S, Kirchner L. Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks. ProPublica 2016. URL https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.
[43] Poursaeed O, Katsman I, Gao B, Belongie S. Generative adversarial perturbations. In: IEEE conference on computer vision and pattern recognition. 2018.
[44] Karako C, Manggala P. Using image fairness representations in diversity-based re-ranking for recommendations. In: Conference on user modeling, adaptation and personalization. 2018.
[45] Taigman Y, Yang M, Ranzato M, Wolf L. Deepface: Closing the gap to human-level performance in face verification. In: IEEE conference on computer vision and pattern recognition. 2014.
[46] Karkkainen K, Joo J. FairFace: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In: IEEE/CVF winter conference on applications of computer vision. 2021.
[47] Cao M, Li S, Li J, Nie L, Zhang M. Image-text retrieval: A survey on recent research and development. In: International joint conference on artificial intelligence. 2022.
[48] Du Y, Liu Z, Li J, Zhao WX. A survey of vision-language pre-trained models. In: International joint conference on artificial intelligence. 2022.
[49] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, Houlsby N. An image is worth 16x16 words: Transformers for image recognition at scale. In: International conference on learning representations. 2021.
[50] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: IEEE conference on computer vision and pattern recognition. 2016.
[51] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. Neural Inf Process Syst 2017.
[52] Manning CD, Raghavan P, Schutze H. Introduction to Information Retrieval. Cambridge University Press; 2008.
[53] Kelly D, Azzopardi L. How many results per page? A study of SERP size, search behavior and user experience. In: ACM SIGIR conference on research and development in information retrieval. 2015.
[54] Buselli I, López AP, Jiménez EM, Anguita D, Roli F, Oneto L. Mitigating unfair regression in machine learning model updates. In: International conference on machine learning and applications. 2024.
[55] Dastin J. Amazon scraps secret AI recruiting tool that showed bias against women. In: Ethics of data and analytics. 2022.
[56] Oneto L, Navarin N, Biggio B, Errica F, et al. Towards learning trustworthily, automatically, and with guarantees on graphs: An overview. Neurocomputing 2022;493:217–43.
[57] Pessach D, Shmueli E. A review on fairness in machine learning. ACM Comput Surv 2022;55(3):1–44.
[58] Mitchell S, Potash E, Barocas S, D'Amour A, Lum K. Algorithmic fairness: Choices, assumptions, and definitions. Annu Rev Stat Appl 2021;8(1):141–63.
[59] Nanda V, Dooley S, Singla S, Feizi S, Dickerson JP. Fairness through robustness: Investigating robustness disparity in deep learning. In: ACM conference on fairness, accountability, and transparency. 2021.
[60] Wachter S, Mittelstadt B, Russell C. Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI. Comput Law Secur Rev 2021.
[61] Geyik SC, Ambler S, Kenthapadi K. Fairness-aware ranking in search & recommendation systems with application to linkedin talent search. In: ACM SIGKDD international conference on knowledge discovery & data mining. 2019.
[62] Franco D, Oneto L, Anguita D. Fair empirical risk minimization revised. In: International work-conference on artificial and natural neural networks. 2023.
[63] Calders T, Kamiran F, Pechenizkiy M. Building classifiers with independency constraints. In: IEEE international conference on data mining. 2009.
[64] Hardt M, Price E, Srebro N. Equality of opportunity in supervised learning. In: Neural information processing systems. 2016.
[65] Oneto L, Donini M, Pontil M, Shawe-Taylor J. Randomized learning and generalization of fair and private classifiers: From PAC-Bayes to stability and differential privacy. Neurocomputing 2020;416:231–43.
[66] Lamy A, Zhong Z, Menon AK, Verma N. Noise-tolerant fair classification. 2019,
[67] Wang J, Liu Y, Levy C. Fair classification with group-dependent label noise. In: ACM conference on fairness, accountability, and transparency. 2021.
[68] Cinà AE, Grosse K, Demontis A, Vascon S, Zellinger W, Moser BA, Oprea A, Biggio B, Pelillo M, Roli F. Wild patterns reloaded: A survey of machine learning security against training data poisoning. ACM Comput Surv 2023;55(13s).
[69] Feng Y, Ma F, Lin W, Yao C, Chen J, Yang Y. FedPAM: Federated personalized augmentation model for text-to-image retrieval. In: International conference on multimedia retrieval. 2024.
[70] Palazzo M, Dekker FW, Brighente A, Conti M, Erkin Z. Privacy-Preserving data aggregation with public verifiability against internal adversaries. In: USENIX security symposium. 2024.
[71] Carlini N, Jagielski M, Choquette-Choo CA, Paleka D, Pearce W, Anderson H, Terzis A, Thomas K, Tramèr F. Poisoning web-scale training datasets is practical. In: IEEE symposium on security and privacy. SP, 2024.
[72] Rashidi L, Zobel J, Moffat A. Query variability and experimental consistency: A concerning case study. In: ACM SIGIR international conference on theory of information retrieval. 2024.
[73] Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, Küttler H, Lewis M, Yih W-t, Rocktäschel T, Riedel S, Kiela D. Retrieval-augmented generation for knowledge-intensive NLP tasks. In: Neural information processing systems. 2020.
[74] Shereen E, Ristea D, Hasircioglu B, McFadden S, Mavroudis V, Hicks C. One pic is all it takes: Poisoning visual document retrieval augmented generation with a single image. 2025, arXiv preprint arXiv:2504.02132.
[75] Lazzaro D, Mura R, Cinà AE, Laurita G, Vercelli G, Oneto L, Biggio B, Roli F. Poison once, fool many: Practical poisoning attacks against text-to-image retrieval systems. Knowl-Based Syst 2026;334:115090.
[76] Carlini N, Wagner D. Towards evaluating the robustness of neural networks. In: IEEE symposium on security and privacy. 2017.
[77] Biggio B, Roli F. Wild patterns: Ten years after the rise of adversarial machine learning. In: ACM SIGSAC conference on computer and communications security. 2018.
[78] Schlarmann C, Singh ND, Croce F, Hein M. Robust CLIP: unsupervised adversarial fine-tuning of vision embeddings for robust large vision-language models. In: International conference on machine learning. 2024.
[79] Abad Rocamora E, Schlarmann C, Singh ND, Wu Y, Hein M, Cevher V. Robustness in both domains: CLIP needs a robust text encoder. In: Neural information processing systems. 2025.
[80] Yang W, Gao J, Mirzasoleiman B. Robust contrastive language-image pretraining against data poisoning and backdoor attacks. Neural Inf Process Syst 2023.
[81] Shi Y, Du M, Wu X, Guan Z, Sun J, Liu N. Black-box backdoor defense via zero-shot image purification. In: Neural information processing systems. 2023.
[82] Sitawarin C, Tramèr F, Carlini N. Preprocessors matter! realistic decision-based attacks on machine learning systems. In: International conference on machine learning. 2022.
[83] Karpathy A, Fei-Fei L. Deep visual-semantic alignments for generating image descriptions. In: IEEE conference on computer vision and pattern recognition. 2015.
[84] Team G, Kamath A, Ferret J, Pathak S, Vieillard N, Merhej R, Perrin S, Matejovicova T, Ramé A, Rivière M, et al. Gemma 3 technical report. 2025, arXiv preprint arXiv:2503.19786.