Prototypicality Bias Reveals Blindspots in Multimodal Evaluation Metrics

Anonymous ACL submission

Abstract

Automatic metrics are widely used to evaluate text-to-image models, often replacing human judgment in benchmarking, model selection, and large-scale data filtering. Yet it is unclear whether these metrics truly measure semantic correctness or instead reward images that look more visually or socially prototypical. We identify *prototypicality bias* as a systematic failure mode in multimodal evaluation: metrics may prefer a semantically incorrect but prototypical image over a correct but less prototypical one. To study this phenomenon, we introduce PROTOBIAS, a controlled benchmark spanning **Animals**, **Objects**, and **Demography**, in which semantically correct images are contrasted with subtly incorrect yet more prototypical adversarial counterparts. We find that widely used metrics frequently fail on these contrasts, whereas human annotators remain substantially more grounded in semantic correctness. To mitigate this issue, we propose PROTOSCORE, a lightweight open-source evaluator trained contrastively on prototypicality confounds. PROTOSCORE dramatically reduces failure rates on PROTOBIAS, transfers strongly to WINOGROUND, and in several settings outperforms stronger closed-source judges such as GPT-5 while being smaller, faster, and reproducible.

1 Introduction

Recent advances in text-to-image (T2I) generation have produced models capable of synthesizing visually compelling and linguistically grounded images at scale. As generation quality has improved, reliable and scalable evaluation of text–image alignment has become increasingly important. Modern benchmarks, model selection pipelines, and large-scale dataset filtering frequently rely on automatic metrics such as CLIPSCORE (Hessel et al., 2022), PICKSCORE (Kirstain et al., 2023), IMAGEREWARD (Xu et al., 2023), VQASCORE (Lin et al., 2024), and more recently LLM-as-Judge systems (Ge et al., 2023). These metrics are often treated as scalable substitutes for human judgment, reducing the cost of large-scale evaluation.

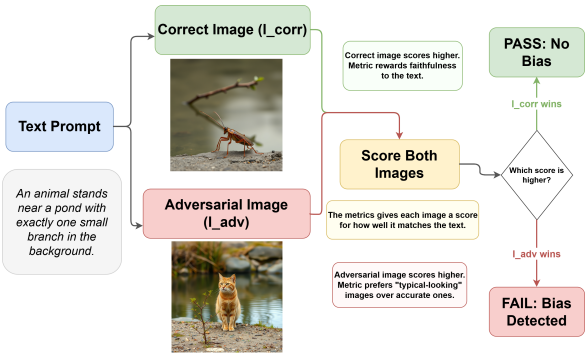

Figure 1: Illustration of prototypicality bias in text-to-image evaluation. Given a prompt, a semantically correct but less prototypical image (I_{corr}) is compared against a semantically incorrect but more prototypical adversarial image (I_{adv}). A robust metric should assign a higher score to I_{corr}; if I_{adv} scores higher, the metric exhibits prototypicality bias.

However, accumulating evidence suggests that multimodal evaluation metrics can be brittle: they may exploit spurious correlations, reflect dataset artifacts, or overweight superficial visual cues (Dai et al., 2024; Hirota et al., 2025; Hwang et al., 2025). In this paper, we identify and study a complementary but underexplored failure mode: *prototypicality bias*. We define prototypicality bias as the tendency of evaluation metrics to prefer images that appear visually or socially "typical," even when they violate the explicit semantics of the prompt. This failure mode is particularly problematic because prototypical images often appear more plausible, allowing semantic errors to be overlooked by automatic evaluators.

Prototypicality bias plausibly arises from the training distributions of both generative models and evaluation metrics. Large-scale vision–language corpora overrepresent canonical visual forms and

socially dominant stereotypes (Girrbach et al., 2025; Wan et al., 2024), encouraging models to internalize expectations about what objects, animals, or people *should* look like. As a result, a semantically incorrect but prototypical image may receive a higher alignment score than a correct but less prototypical one. Related effects have long been studied in cognitive science through *Prototype Theory*, which shows that categories are organized around more typical members rather than uniform sets of instances (Rosch, 1975). Figure 1 illustrates this phenomenon in the context of automated text–image evaluation.

To systematically study this phenomenon, we introduce PROTOBIAS, a controlled benchmark that explicitly contrasts semantic correctness with prototypical familiarity. Each example consists of a text prompt together with two images: a semantically correct but less prototypical image and a prototypical but semantically incorrect adversarial alternative. This setup allows us to test a simple but consequential question: when semantic fidelity and prototypical familiarity conflict, which signal does an evaluator reward? We instantiate this contrast across three domains: **Animals**, **Objects**, and **Demography**. The first two capture graded category structure in natural and artifact categories, while the third examines similar effects in socially learned defaults and stereotypes.

In summary, this work makes the following contributions:

- **Problem formulation.** We identify *prototypicality bias* as a failure mode in text-to-image evaluation, where metrics prefer visually or socially prototypical images even when they violate prompt semantics.

- **Benchmark.** We introduce PROTOBIAS, a controlled benchmark spanning **Animals**, **Objects**, and **Demography**, constructed from contrasted triplets $(T, I_{\text{corr}}, I_{\text{adv}})$ that explicitly place semantic correctness in conflict with prototypicality.

- **Mitigation.** We propose PROTOSCORE, a lightweight open-source evaluator trained contrastively on prototypicality confounds, which substantially reduces failure rates on PROTOBIAS and transfers well to WINOGROUND (Thrush et al., 2022) & TIFA160 (Hu et al., 2023).

2 Related Work

2.1 Benchmarking Social and Cultural Biases in Multimodal Models

The proliferation of large-scale text-to-image (T2I) and vision-language models (VLMs) has prompted extensive investigations into their tendency to amplify societal biases (Wan et al., 2024; Luo et al., 2024). Prior work documents biases across gender, skin tone, and geo-cultural representation, motivating a range of diagnostic benchmarks (Wan et al., 2024; Luo et al., 2024; Narayanan et al., 2025). For example, BIGBENCH evaluates multiple forms of social bias in generative systems (Luo et al., 2024), while other studies show that T2I models reproduce occupational gender stereotypes and other demographic imbalances (Contreras, 2025; Nair et al., 2025). Cultural representation has also received increasing attention, with recent benchmarks revealing systematic Global-North defaults in generated imagery (Seo et al., 2025; Said et al., 2025).

Related work further examines how VLMs interpret socially contextual cues. Benchmarks such as VIGNETTE evaluate whether models infer social hierarchies from visual signals (Raj et al., 2025; Jiang et al., 2024). These biases have been linked to imbalances in large-scale training datasets such as LAION-400M (Girrbach et al., 2025). While this literature primarily focuses on identifying specific social or cultural biases, our work studies a broader phenomenon: *prototypicality bias*. Instead of targeting particular stereotypes, we analyze a more general failure mode in which evaluation metrics prefer statistically typical or canonical examples over semantically correct but less prototypical ones. We connect this behavior to prototype-based category structure from cognitive science (Rosch, 1975; Ma and Correll, 2011).

2.2 Evaluating Metric Robustness and MLLM-as-Judge Systems

A parallel line of research investigates the robustness of automatic evaluation metrics for multimodal generation (Dai et al., 2024; Hwang et al., 2025). Several studies show that these metrics can rely on spurious correlations or superficial visual cues rather than true semantic grounding (Hirota et al., 2025; Ye et al., 2024). Small non-semantic perturbations can therefore change metric outputs substantially and distort downstream bias estimates (Hirota et al., 2025). Similar concerns have been raised in broader robustness evaluations of multi-

modal systems (Wang et al., 2025; Madaan et al., 2025), including evaluation suites such as AHELM that jointly assess fairness, robustness, and safety (Lee et al., 2025).

Recent work also explores targeted stress tests to reveal hidden weaknesses in evaluation metrics. In machine translation, robustness-driven meta-evaluation shows that metrics with high average correlation can still fail systematically on controlled adversarial examples (He et al., 2023; Chen and Eger, 2023). Similar ideas are emerging for T2I evaluation. For instance, CROC introduces contrastive robustness checks to detect semantic failures and derive robustness-aware scoring functions (Leiter et al., 2025). Other frameworks such as PREXME and METALOGIC probe preference inconsistencies under controlled prompt or editing variations (Leiter and Eger, 2024; Shen et al., 2025).

At the same time, multimodal large language models (MLLMs) are increasingly used as automated judges for image generation (Ge et al., 2023). However, recent studies show that such systems remain vulnerable to manipulation, for example through adversarial visuals that inflate evaluation scores (Slyman et al., 2025; Hwang et al., 2025). Our work differs from these robustness studies by focusing on a specific and interpretable failure mode: situations in which semantic correctness conflicts with prototypical visual cues. We operationalize this contrast using triplets $(T, I_{\text{corr}}, I_{\text{adv}})$ that explicitly test whether an evaluator prefers a semantically correct but non-prototypical image or a prototypical but incorrect alternative.

3 Methodology

3.1 Evaluation Direction and Failure Criterion

Let T denote a textual description. We partition the space of generated images according to whether an image is semantically consistent with T:

$$\mathcal{C}(T) = \{I \mid I \equiv T\}, \tag{1}$$

$$\mathcal{I}(T) = \{I \mid I \not\equiv T\}, \tag{2}$$

where $I \equiv T$ means that image I is semantically correct with respect to T, and $I \not\equiv T$ means that it is semantically inconsistent.

We are interested in two types of images under a similarity measure $\text{sim}(\cdot, \cdot)$ between text and image:

$$\begin{aligned} I_{\text{corr}} &= \arg\min_{I \in \mathcal{C}(T)} \text{sim}(T, I), \\ I_{\text{adv}} &= \arg\max_{I \in \mathcal{I}(T)} \text{sim}(T, I). \end{aligned} \tag{3}$$

Here, I_{corr} is a **semantically correct image**: it satisfies the text but may be less prototypical or less visually canonical. In contrast, I_{adv} is a **prototypical adversarial image**: it looks plausible and prototypical, but contains at least one semantic violation. We do not solve Eq. (3) exactly; rather, it motivates our benchmark design.

Given a multimodal evaluation metric $M(T, I)$, a semantically robust metric should prefer the correct image over the adversarial one:

$$M(T, I_{\text{corr}}) > M(T, I_{\text{adv}}). \tag{4}$$

Failure Criterion. We define a *metric failure* on text T whenever

$$M(T, I_{\text{adv}}) \geq M(T, I_{\text{corr}}), \tag{5}$$

that is, when the metric assigns a higher score to a semantically incorrect but prototypical image than to a correct but less prototypical one. This directly captures the failure mode we study: prioritizing prototypicality over semantic correctness.

3.2 Dataset and Prototypical Bias Setup

Prototypicality as graded category structure. Our dataset design is grounded in *Prototype Theory*, which models categories as organized around more typical members rather than uniformly representative instances (Rosch, 1975). Cognitive studies show that category membership is graded: some instances are perceived as more canonical than others while remaining valid members of the same category (Boster, 1988; Uyeda and Mandler, 1980; Joelson and Herrmann, 1978; Battig and Montague, 1968, 1969). For example, in the category *animal*, instances such as *dog* or *robin* are typically judged as more central than *penguin* or *platypus*. We operationalize prototypicality as a *relative contrast* between more and less typical instances within the same category, enabling controlled tests of whether evaluation metrics prioritize prototypical familiarity over semantic correctness.

Taxonomy construction. Using this principle, we construct a benchmark spanning three domains: **Animals**, **Objects**, and **Demography**.

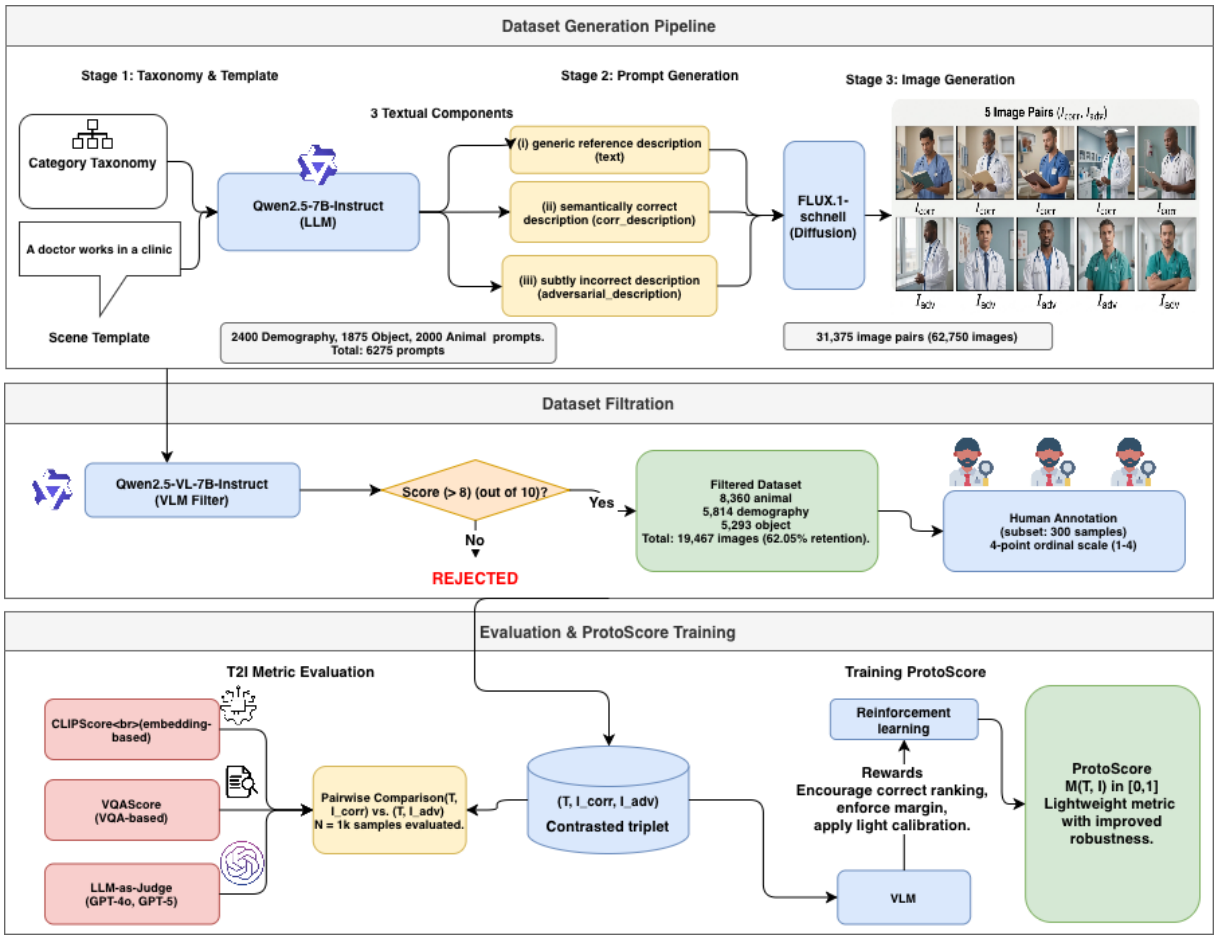

Figure 2: Overview of the ProtoBias dataset construction, filtering, evaluation, and ProtoScore training pipeline. In Stage 1, category taxonomies and scene templates define structured generation inputs. In Stage 2, Qwen2.5-7B-Instruct generates three textual components per instance: a generic reference description, a semantically correct description, and a subtly incorrect adversarial description. In Stage 3, FLUX.1-schnell generates paired images $(I_{\text{corr}}, I_{\text{adv}})$. A VLM-based filtering stage using Qwen2.5-VL-7B-Instruct retains only high-alignment samples, followed by human validation on a subset. The resulting contrasted triplets $(T, I_{\text{corr}}, I_{\text{adv}})$ are used both for metric evaluation and for training ProtoScore via reinforcement learning.

For **Animals**, we define 20 less-prototypical/more-prototypical pairs such as *penguin* - *robin* and *platypus* - *dog*, motivated by classic studies of graded typicality in natural categories (Rosch, 1975; Boster, 1988; Uyeda and Mandler, 1980). For **Objects**, we construct 18 pairs across three artifact groups: *furniture* (*bean bag-chair*), *vehicles* (*e-scooter* - *motorcycle*), and *tableware* (*chopsticks* - *fork*), capturing canonical–peripheral structure in artifact categories (Joelson and Herrmann, 1978; Battig and Montague, 1968, 1969). The **Demography** domain extends this contrast to socially learned prototypes, where prototypicality reflects dominant associations between demographic identities and social attributes (Ma and Correll, 2011; McIntosh, 1988; Rich, 1980; Buolamwini and Gebru, 2018; Mehrabi et al., 2022). Following MMBias (Janghorbani and de Melo, 2023) and FairPIVARA (Moreira et al., 2024), we generate examples across three demographic axes—*Religion*, *Nationality*, and *Sexual Orientation*—combined with five socio-attributes: *Wealth*, *Intellect*, *Morality*, *Power*, and *Civility*. The dataset includes both typical and counter-stereotypical pairings to create controlled contrasts between prototypical and non-prototypical cases.

Adversarial construction. We construct prototypical adversarial examples by keeping the main entity in the prompt unchanged and introducing semantic inconsistencies through auxiliary scene elements. These elements remain visually plausible within the scene (e.g., branches or stones for animals, furniture for objects, books or laptops for people) and are perturbed along four semantic

dimensions: *count*, *color*, *size*, and *spatial layout*.

Each triplet therefore contrasts a semantically correct but less prototypical image with a prototypical yet incorrect adversarial image, allowing us to test whether evaluation metrics prioritize semantic correctness or prototypical familiarity.

4 Experimental Setup

Dataset Generation Pipeline & Infrastructure As illustrated in Figure 2, we construct ProtoBias using a structured four-stage pipeline: *taxonomy* $\rightarrow$ *template* $\rightarrow$ *prompt generation* $\rightarrow$ *image generation*. The taxonomy defines category families, while scene templates provide short, neutral descriptions that specify the visual context (see Appendix A.2). **Prompt Generation** is performed using Qwen2.5-7B-Instruct. For each taxonomy entry, we generate three textual components: (i) a generic reference description T, (ii) a semantically correct description producing I_{corr}, and (iii) a subtly incorrect adversarial description producing I_{adv}. In total, we generate 6,275 prompts: 2,000 for Animals, 1,875 for Objects, and 2,400 for Demography. Prompt generation is performed on NVIDIA A100 GPUs and requires approximately two hours per category. For **image generation**, we use the FLUX.1-schnell diffusion model with 5 inference steps. Each prompt produces five image pairs $(I_{\text{corr}}, I_{\text{adv}})$ corresponding to the semantically correct and adversarial descriptions, yielding **31,375 image pairs** (**62,750 images**) in total. Image synthesis was executed on $8\times$A100 GPUs and completed in approximately 3-4 hours. Each instance forms a contrasted triplet $(T, I_{\text{corr}}, I_{\text{adv}})$, enabling direct pairwise evaluation between semantically correct and adversarial images.

Dataset Filtration To ensure high semantic alignment between prompts and generated images, we apply an automated filtering stage using Qwen2.5-VL-7B-Instruct. The VLM evaluates each image–text pair and assigns a semantic alignment score on a 1–10 scale conditioned solely on the original generation prompt. Images scoring ≥ 8 are retained.

After filtering, the dataset contains **19,467 images**: 8,360 Animals, 5,814 Demography, and 5,293 Objects (62.05% retention rate).

To assess the reliability of this automated step, we conduct a human validation on a random subset of 300 filtered samples (100 per domain). Human annotators follow the same semantic alignment rubric used by the filtering model. The results show substantial agreement between human judgments and the VLM filter, with limited retain/reject disagreement. Full statistics are reported in Appendix C.

To further examine whether the observed effects could be explained by differences in visual quality between SC and PA images, we conduct an additional no-reference image quality analysis using MUSIQ (Ke et al., 2021) and NIQE (Zvezdakova et al., 2019). The results show only negligible differences between the two sets of images. Full details are provided in Appendix D.

T2I Metrics We evaluate several widely used text-to-image (T2I) evaluation metrics covering different evaluation paradigms. These include the embedding-based metric **CLIPScore** (Hessel et al., 2022), the learned preference model **PickScore** (Kirstain et al., 2023), the reward model **ImageReward** (Xu et al., 2023), and the VQA-based metric **VQAScore** (Lin et al., 2024). In addition, we evaluate multimodal **LLM-as-Judge** systems using GPT-4o and GPT-5, prompted to score image–text alignment based solely on semantic correctness, consistent with the human evaluation protocol described in §E.

All metrics produce scalar alignment scores $M(T, I)$. To ensure consistency across metrics with different scoring ranges, we normalize all scores to $[0, 1]$. While evaluation is pairwise (and invariant to monotonic transformations), normalization enables consistent reporting. Evaluation is performed in a pairwise setting: for each text T, we compare the scores assigned to the semantically correct image I_{corr} and the adversarial image I_{adv}. Following the evaluation criterion in §3.1, a metric is correct when it assigns a higher score to the semantically correct image than to the adversarial one. A failure occurs when the metric assigns a higher score to the adversarial image (Eq. 5).

All metrics are evaluated on $N = 1{,}500$ **samples**, evenly distributed across the Animals, Objects, and Demography categories.

Human Evaluation Setup We conducted two human annotation studies to validate the dataset and evaluation protocol. In the first study, five annotators independently scored image–text pairs on a 4-point Likert scale (1-4) based on semantic correctness. In the second study, a separate group of five annotators performed pairwise comparisons between I_{corr} and I_{adv}, selecting which image bet-

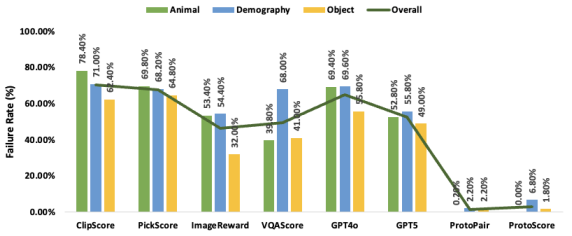

Figure 3: Failure rates of multimodal evaluation metrics across domains.

ter matched the prompt (Image 1, Image 2, or Tie) with randomized image order and reporting confidence (weak, medium, strong). Full annotation details are provided in Appendix E.

Training PROTOSCORE & PROTOPAIR. We train two evaluation metrics on the PROTOBIAS training set: a **pointwise alignment metric** (PROTOSCORE) and a **pairwise comparison metric** (PROTOPAIR). Both are trained on 10K samples from the filtered dataset, balanced across Animals, Objects, and Demography, disjoint from all evaluation splits, and trained on a single H100 GPU for approximately 3-4 hours.

Pointwise metric. PROTOSCORE predicts a scalar alignment score $M(T, I) \in [0, 1]$ for a text–image pair via a sigmoid scoring head. Training uses SC/PA image pairs with a margin-based ranking objective with calibration and spread regularization (margin 0.10, $w_{calib} = 0.25$, $w_{spread} = 0.05$). We fine-tune Qwen VL backbones with LoRA (r=16, α=32, dropout=0.05) applied to attention and MLP projections. Images are resized to 512×512 and text truncated to 96 tokens. Optimization uses AdamW (lr 2×10^{-5}), cosine schedule with 3% warmup, gradient clipping 1.0, fp16 training, batch size 2 with gradient accumulation 16 (effective batch size 32), for one epoch. We train **five PROTOSCORE variants**.

Pairwise metric. PROTOPAIR predicts which of two images better matches a caption by generating a binary preference (A/B) using next-token logits with cross-entropy supervision. We use the same preprocessing and LoRA configuration and train the metric based on Qwen3-VL-8B-Instruct.

5 Results & Analysis

In this section, we analyze the behavior of multimodal evaluation metrics on our controlled prototypicality benchmark and additionally report transfer results on other text-to-image benchmarks.

Failure rate under prototypicality confounds. Figure 3 shows that widely used text-image evaluation metrics fail frequently under prototypicality confounds. Embedding-based metrics perform worst: CLIPSCORE fails in 78.4%, 71.0%, and 62.4% of cases across Animals, Demography, and Objects (70.6% overall), while PICKSCORE exhibits similarly high failure (67.6% overall). Stronger baselines reduce but do not eliminate this effect: IMAGEREWARD fails in 46.6% of cases overall and VQASCORE in 49.6%, with Demography remaining particularly challenging (68.0%). LLM-based judges remain unstable, with GPT-4O and GPT-5 reaching 64.9% and 52.5% overall failure, respectively. In contrast, our methods substantially improve robustness: PROTOPAIR achieves only 1.7% overall failure, and PROTOSCORE (QWEN3-VL-8B) reduces failure to 2.9% overall (0.0% on Animals, 6.8% on Demography, and 1.8% on Objects).

These results show that training on contrastive semantic pairs improves robustness to prototypical confounds.

Results on PROTOBIAS Figure 3 and Table 1 summarize metric behavior under prototypicality confounds. Similarity-based metrics show strong bias toward prototypical adversarial images. CLIPScore assigns higher scores to adversarial images across all domains (e.g., 0.26 vs. 0.23 in Animals), producing negative margins ($\Delta = -0.03, -0.02, -0.01$). PickScore exhibits an even stronger reversal, scoring adversarial images higher in Animals (0.62 vs. 0.55) and Demography (0.58 vs. 0.52).

Stronger evaluators reduce but do not eliminate this effect. ImageReward still prefers adversarial images in Animals (0.76 vs. 0.72) and Demography (0.76 vs. 0.75), while VQAScore reverses preference on Demography (0.73 vs. 0.70). LLM-based judges improve separation, with GPT-4o achieving margins of 0.12, 0.08, and 0.20, and GPT-5 increasing these to 0.28, 0.19, and 0.26, though still below human separation (0.65, 0.33, 0.44).

In contrast, ProtoScore consistently yields the largest margins, reaching 0.90 on Animals (Qwen2.5-VL-7B), 0.38 on Demography (Qwen3-VL-4B), and 0.57 on Objects (Qwen3-VL-8B). These results indicate that training evaluators on contrasted semantic pairs substantially improves robustness to prototypicality bias.

Metric	Model	Animals			Demography			Objects		
		SC	PA	Δ	SC	PA	Δ	SC	PA	Δ
CLIPScore	CLIP-L-14	0.23	0.26	-0.03	0.22	0.24	-0.02	0.22	0.24	-0.01
PickScore	CLIP-H-14	0.55	0.62	-0.07	0.52	0.58	-0.06	0.48	0.52	-0.05
ImageReward	BLIPv2	0.72	0.76	-0.04	0.75	0.76	-0.01	0.51	0.43	0.09
VQAScore	CLIP-FlanT5-XXL	0.77	0.72	0.05	0.70	0.73	-0.03	0.88	0.84	0.04
LLM-as-Judge	GPT-4o	0.65	0.53	0.12	0.68	0.59	0.08	0.73	0.53	0.20
LLM-as-Judge	GPT-5	0.65	0.38	0.28	0.67	0.48	0.19	0.79	0.53	0.26
ProtoScore	SmolVLM	0.55	0.10	0.46	0.33	0.13	0.20	0.48	0.14	0.34
ProtoScore	Qwen2.5-VL-7B	0.94	0.04	**0.90**	0.56	0.21	0.35	0.68	0.14	0.54
ProtoScore	Qwen3-VL-4B	0.96	0.21	0.75	0.57	0.18	**0.38**	0.86	0.29	0.56
ProtoScore	Qwen3-VL-8B	0.93	0.15	0.78	0.40	0.14	0.27	0.74	0.17	**0.57**
Human	–	0.89	0.24	**0.65**	0.50	0.17	**0.33**	0.84	0.40	**0.44**

Table 1: Average normalized scores assigned to semantically correct (SC) and prototypical adversarial (PA) images on PROTOBIAS. $\Delta = \text{SC} - \text{PA}$ denotes the semantic separation margin; larger positive values indicate stronger preference for semantically correct images. Human judgments are computed on the same 300 annotated samples and serve as a reference upper bound. Human Δ values are shown in **bold and underlined**. Among automated metrics, the highest Δ in each column is shown in **bold**, and the second-highest is underlined.

Metric	Animals	Demography	Objects
CLIPScore	21.6	29.0	37.6
PickScore	30.2	31.8	35.2
ImageReward	46.6	45.6	68.0
VQAScore	60.2	32.0	59.0
GPT-4o	30.6	30.4	44.2
GPT-5	47.2	44.2	51.0
ProtoPair (Ours)	**99.8**	**97.8**	**98.2**

Table 2: Pairwise evaluation on PROTOBIAS. For each prompt, the metric must select the semantically correct image from a pair consisting of a semantically correct (SC) image and a prototypical adversarial (PA) image. We report accuracy across domains. Higher values indicate stronger robustness to prototypicality bias.

Metric	Model	Text	Image	Group
CLIPScore	CLIP-L-14	27.8	11.5	7.8
PickScore	CLIP-H-14	23.8	12.5	6.8
ImageReward	BLIPv2	42.8	15.3	12.8
VQAScore	CLIP-FlanT5-XXL	60.0	57.5	46.0
ProtoScore	SmolVLM	19.2	3.8	2.2
ProtoScore	Qwen2.5-VL-7B	70.0	61.0	52.8
ProtoScore	Qwen3-VL-4B	61.2	57.0	46.5
ProtoScore	**Qwen3-VL-8B**	**74.5**	**71.5**	**62.8**

Table 3: Transfer results on WINOGROUND. Higher is better. We report standard Text, Image, and Group scores.

Pairwise evaluation. We further evaluate metrics in a direct pairwise setting, where the evaluator must select the semantically correct image from each SC-PA pair. Table 2 reports pairwise accuracy across domains. Existing metrics perform poorly in this setting: for example, CLIPScore achieves only 21.6% accuracy on Animals, indicating a strong preference for prototypical adversarial images. GPT-based judges improve performance but remain inconsistent (e.g., GPT-5 reaches 47.2% on Animals and 51.0% on Objects). In contrast, PROTOPAIR achieves near-perfect accuracy across domains (99.8%, 97.8%, and 98.2%), demonstrating that explicitly modeling the semantic contrast between paired images substantially improves robustness to prototypicality bias.

Transfer to WINOGROUND. To evaluate whether the learned evaluator transfers beyond PROTOBIAS, we test all models on WINOGROUND (Xu et al., 2023), a benchmark designed to measure compositional text–image understanding using paired images and captions with subtle semantic swaps. Winoground reports three metrics: Text, Image, and Group accuracy, which measure whether a model correctly matches captions and images under compositional swaps.

Table 3 summarizes the results. Among the baselines, CLIPScore and PickScore perform poorly, achieving group accuracies of 7.8 and 6.8, respectively. ImageReward improves this to 12.8, while VQAScore reaches 46.0. ProtoScore models achieve the best performance, surpassing GPT-5 by a clear margin, with Qwen2.5-VL-7B reaching 70.0 / 61.0 / 52.8 on Text, Image, and Group. The Qwen3 variants show similar behavior, with Qwen3-VL-8B obtaining 74.5 / 71.5 / 62.8.

Overall, training on contrastive semantic pairs preserves compositional grounding and transfers effectively to external benchmarks, with additional

	SC Image	PA Image	SC preferred? (✓ = SC, ✗ = PA)
	Prompt: "An animal stands near a pond with exactly two bamboo stalks in the background."		
Animals			✓ Human ✗ CLIP ✗ Pick ✗ VQA ✗ ImageReward ✗ GPT-4o ✓ GPT-5 ✓ **ProtoScore**
	Prompt: "An uneducated person stands in a quiet room with a single chalkboard in the background."		
Demography			✓ Human ✗ CLIP ✗ Pick ✗ VQA ✗ ImageReward ✗ GPT-4o ✓ GPT-5 ✓ **ProtoScore**
	Prompt: "A vehicle is parked on a quiet street with two manhole covers near it."		
Objects			✓ Human ✗ CLIP ✗ Pick ✓ VQA ✓ ImageReward ✗ GPT-4o ✓ GPT-5 ✓ **ProtoScore**

Figure 4: **Examples of prototypicality bias in multimodal metrics.** Each row contrasts a semantically correct but non-prototypical image (SC) with a prototypical but semantically incorrect adversarial image (PA) generated from the same prompt. The rightmost column indicates whether each evaluator ranks SC above PA (✓) or prefers PA (✗).

results on TIFA160 reported in Appendix F.

Human Evaluation Human judgments provide an additional validation signal for the benchmark. In the independent scoring task (Table 1), annotators assign higher scores to semantically correct (SC) images than to prototypical adversarial (PA) images across all domains, yielding separation margins of 0.65, 0.33, and 0.44 for Animals, Demography, and Objects. Inter-annotator agreement (IAA) measured using weighted Cohen's κ ranges from $\kappa = 0.60$ to $\kappa = 0.84$ (Appendix Table 9), indicating substantial agreement.

In the pairwise comparison task, annotators prefer the semantically correct image in **82%** of cases overall, with **12%** adversarial selections and **6%** ties. Performance varies by domain: Animals (**94%** SC) and Objects (**92%** SC) are straightforward, while Demography is more challenging (**62%** SC). IAA measured with Cohen's κ ranges from $\kappa = 0.42$ to $\kappa = 0.87$ (Appendix Table 10).

6 Conclusion

We identify *prototypicality bias* as a systematic failure mode in T2I evaluation: widely used metrics often prefer visually or socially typical images even when they violate prompt semantics. Using PROTOBIAS, a controlled benchmark spanning Animals, Objects, and Demography, we show that this issue affects embedding-based, reward-based, VQA-based and LLM-based evaluators, while humans remain more semantically grounded. To study and mitigate this effect, we introduce PROTOSCORE, a lightweight contrastively trained evaluator that learns to distinguish semantically correct images from prototypical adversarial ones. PROTOSCORE substantially reduces failure rates compared to existing metrics and transfers well to external benchmarks. These results suggest that reliable T2I evaluation requires modeling semantic correctness under realistic prototype-driven confounds rather than relying on generic similarity signals.

Limitations

This work focuses on controlled, contrastive settings designed to isolate prototypicality bias, and therefore does not capture the full diversity and complexity of open-world text-image generation. While our benchmark spans animals, objects, and demography, it covers a limited set of categories and stereotypes, primarily reflecting Western-centric data distributions. Human evaluation is conducted with a relatively small number of annotators, and although details are provided in the appendix, the sample may not fully capture broader population variability. Finally, while we show that prototypicality bias affects several widely used metrics, our analysis is limited to a specific set of recent models and evaluation protocols, and may not fully generalize to future architectures or alternative evaluation settings.

References

William F. Battig and William E. Montague. 1968. Category norms for verbal items in 56 categories: A replication and extension of the connecticut norms using university of maryland and illinois students. Technical report, University of Colorado, Boulder, CO.

William F. Battig and William E. Montague. 1969. Category norms of verbal items in 56 categories: A replication and extension of the connecticut category norms. *Journal of Experimental Psychology*, 80(3, Pt.2):1–46.

James S. Boster. 1988. Natural sources of internal category structure: Typicality, familiarity, and similarity of birds. *Memory & Cognition*, 16(3):258–270.

Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 77–91. PMLR.

Yanran Chen and Steffen Eger. 2023. Menli: Robust evaluation metrics from natural language inference. *Preprint*, arXiv:2208.07316.

Juan Manuel Contreras. 2025. Automated evaluation of gender bias across 13 large multimodal models. *2509.07050v1*.

Xiang Dai, Sarvnaz Karimi, and Biaoyan Fang. 2024. A critical look at meta-evaluating summarisation evaluation metrics. *2409.19507v1*.

Wentao Ge, Shunian Chen, Guiming Hardy Chen, Junying Chen, Zhihong Chen, Nuo Chen, Wenya Xie, Shuo Yan, Chenghao Zhu, Ziyue Lin, Song Dingjie, Xidong Wang, Anningzhe Gao, Zhang Zhiyi, Jianquan Li, Xiang Wan, and Benyou Wang. 2023. Mllm-bench: Evaluating multimodal llms with per-sample criteria. *2311.13951v3*.

Leander Girrbach, Stephan Alaniz, Genevieve Smith, Trevor Darrell, and Zeynep Akata. 2025. Person-centric annotations of laion-400m: Auditing bias and its transfer to models. *2510.03721v1*.

Tianxing He, Jingyu Zhang, Tianle Wang, Sachin Kumar, Kyunghyun Cho, James Glass, and Yulia Tsvetkov. 2023. On the blind spots of model-based evaluation metrics for text generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12067–12097, Toronto, Canada. Association for Computational Linguistics.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2022. Clipscore: A reference-free evaluation metric for image captioning. *Preprint*, arXiv:2104.08718.

Yusuke Hirota, Ryo Hachiuma, Boyi Li, Ximing Lu, Michael Ross Boone, Boris Ivanovic, Yejin Choi, Marco Pavone, Yu-Chiang Frank Wang, Noa Garcia, Yuta Nakashima, and Chao-Han Huck Yang. 2025. Bias in gender bias benchmarks: How spurious features distort evaluation. *2509.07596v2*.

Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. 2023. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. *Preprint*, arXiv:2303.11897.

Yerin Hwang, Dongryeol Lee, Kyungmin Min, Taegwan Kang, Yong il Kim, and Kyomin Jung. 2025. Fooling the lvlm judges: Visual biases in lvlm-based evaluation. *2505.15249v1*.

Sepehr Janghorbani and Gerard de Melo. 2023. Multimodal bias: Introducing a framework for stereotypical bias assessment beyond gender and race in vision language models. *Preprint*, arXiv:2303.12734.

Yukun Jiang, Zheng Li, Xinyue Shen, Yugeng Liu, Michael Backes, and Yang Zhang. 2024. ModSCAN: Measuring stereotypical bias in large vision-language models from vision and language modalities. *2410.06967v1*.

John M. Joelson and Douglas J. Herrmann. 1978. Properties of categories in semantic memory. *The American Journal of Psychology*, 91(1):101–114.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. 2021. Musiq: Multi-scale image quality transformer. *Preprint*, arXiv:2108.05997.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. 2023. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Preprint*, arXiv:2305.01569.

Tony Lee, Haoqin Tu, Chi Heem Wong, Zijun Wang, Siwei Yang, Yifan Mai, Yuyin Zhou, Cihang Xie, and Percy Liang. 2025. Ahelm: A holistic evaluation of audio-language models. *2508.21376v2*.

Christoph Leiter, Yuki M. Asano, Margret Keuper, and Steffen Eger. 2025. Croc: Evaluating and training t2i metrics with pseudo- and human-labeled contrastive robustness checks. *arXiv preprint arXiv:2505.11314*.

Christoph Leiter and Steffen Eger. 2024. PrExMe! large scale prompt exploration of open source LLMs for machine translation and summarization evaluation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Miami, Florida, USA. Association for Computational Linguistics.

Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. 2024. Evaluating text-to-visual generation with image-to-text generation. *Preprint*, arXiv:2404.01291.

Hanjun Luo, Haoyu Huang, Ziye Deng, Xinfeng Li, Hewei Wang, Yingbin Jin, Yang Liu, Wenyuan Xu, and Zuozhu Liu. 2024. Bigbench: A unified benchmark for evaluating multi-dimensional social biases in text-to-image models. *2407.15240v6*.

Debbie S. Ma and Joshua Correll. 2011. Target prototypicality moderates racial bias in the decision to shoot. *Journal of Experimental Social Psychology*, 47(2):391–396.

Divyam Madaan, Varshan Muhunthan, Kyunghyun Cho, and Sumit Chopra. 2025. Multi-modal data spectrum: Multi-modal datasets are multi-dimensional. *2509.23499v1*.

P. McIntosh. 1988. *White Privilege and Male Privilege: A Personal Account of Coming to See Correspondences Through Work in Women's Studies*. Working paper (Wellesley College. Center for Research on Women). Wellesley College, Center for Research on Women.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2022. A survey on bias and fairness in machine learning. *Preprint*, arXiv:1908.09635.

Diego A. B. Moreira, Alef Iury Ferreira, Jhessica Silva, Gabriel Oliveira dos Santos, Luiz Pereira, João Medrado Gondim, Gustavo Bonil, Helena Maia, Nádia da Silva, Simone Tiemi Hashiguti, Jefersson A. dos Santos, Helio Pedrini, and Sandra Avila. 2024. Fairpivara: Reducing and assessing biases in clip-based multimodal models. *Preprint*, arXiv:2409.19474.

Rahul Nair, Bhanu Tokas, and Hannah Kerner. 2025. A woman with a knife or a knife with a woman? measuring directional bias amplification in image captions. *2503.07878v4*.

Aravind Narayanan, Vahid Reza Khazaie, and Shaina Raza. 2025. Bias in the picture: Benchmarking vlms with social-cue news images and llm-as-judge assessment. *2509.19659v1*.

Chahat Raj, Bowen Wei, Aylin Caliskan, Antonios Anastasopoulos, and Ziwei Zhu. 2025. Vignette: Socially grounded bias evaluation for vision-language models. *2505.22897v1*.

Adrienne Rich. 1980. Compulsory heterosexuality and lesbian existence. *Signs: Journal of Women in Culture and Society*, 5(4):631–660.

Eleanor Rosch. 1975. Cognitive representations of semantic categories. *Journal of Experimental Psychology: General*, 104(3):192–233.

Muna Numan Said, Aarib Zaidi, Rabia Usman, Sonia Okon, Praneeth Medepalli, Kevin Zhu, Vasu Sharma, and Sean O'Brien. 2025. Deconstructing bias: A multifaceted framework for diagnosing cultural and compositional inequities in text-to-image generative models. *2505.01430v1*.

Huichan Seo, Sieun Choi, Minki Hong, Yi Zhou, Junseo Kim, Lukman Ismaila, Naome Etori, Mehul Agarwal, Zhixuan Liu, Jihie Kim, and Jean Oh. 2025. Exposing blindspots: Cultural bias evaluation in generative image models. *2510.20042v1*.

Yifan Shen, Yangyang Shu, Hye young Paik, and Yulei Sui. 2025. Metalogic: Robustness evaluation of text-to-image models via logically equivalent prompts. *Preprint*, arXiv:2510.00796.

Eric Slyman, Mehrab Tanjim, Kushal Kafle, and Stefan Lee. 2025. Calibrating mllm-as-a-judge via multimodal bayesian prompt ensembles. *2509.08777v1*.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. 2022. Winoground: Probing vision and language models for visio-linguistic compositionality. *Preprint*, arXiv:2204.03162.

Katherine M. Uyeda and George Mandler. 1980. Prototypicality norms for 28 semantic categories. *Behavior Research Methods & Instrumentation*, 12(6):587–595.

Yixin Wan, Arjun Subramonian, Anaelia Ovalle, Zongyu Lin, Ashima Suvarna, Christina Chance, Hritik Bansal, Rebecca Pattichis, and Kai-Wei Chang. 2024. Survey of bias in text-to-image generation: Definition, evaluation, and mitigation. *2404.01030v3*.

Youze Wang, Zijun Chen, Ruoyu Chen, Shishen Gu, Wenbo Hu, Jiayang Liu, Yinpeng Dong, Hang Su, Jun Zhu, Meng Wang, and Richang Hong. 2025. Understanding and benchmarking the trustworthiness in multimodal llms for video understanding. *2506.12336v2*.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. 2023. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Preprint*, arXiv:2304.05977.

Wenqian Ye, Guangtao Zheng, Yunsheng Ma, Xu Cao, Bolin Lai, James M. Rehg, and Aidong Zhang. 2024. Mm-spubench: Towards better understanding of spurious biases in multimodal llms. *2406.17126v1*.

Anastasia Zvezdakova, Dmitriy Kulikov, Denis Kondranin, and Dmitriy Vatolin. 2019. Barriers towards no-reference metrics application to compressed video quality analysis: on the example of no-reference metric niqe. *Preprint*, arXiv:1907.03842.

A Taxonomy & Prompt Templates

A.1 Taxonomy

Animals. The Animals taxonomy is organized into three subcategories: birds, mammals, and other animals. Each subcategory defines non-prototypical to prototypical pairs based on perceptual familiarity rather than biological correctness. Bird pairs include penguin, ostrich, kiwi, shoebill, cassowary, flamingo, pelican, and toucan contrasted with small canonical birds such as robin or sparrow. Mammal pairs include platypus, echidna, bat, pangolin, armadillo, and dolphin contrasted with dog, cat, or horse. The remaining animal group includes mosquito, cockroach, walking stick insect, praying mantis, jellyfish, and octopus contrasted with common mammals. Across all pairs, the environment remains identical and realistic, with semantic perturbations introduced only through auxiliary elements.

Demography. The Demography taxonomy is defined over the hypernym *person* and spans three social axes: religion, nationality, and sexual orientation. Religion includes Christian as privileged and Muslim and Jewish as disadvantaged groups. Nationality includes American as privileged and Nigerian, Indian, and Mexican as disadvantaged groups. Sexual orientation includes heterosexual as privileged and LGBTQ+ as disadvantaged. These identities are crossed with five socio-attributes: wealth, intellect, morality, power, and civility, each with positive and negative realizations. Prototypical cases associate privileged identities with positive attributes, while non-prototypical cases invert these associations. Visual realizations specify clothing, appearance, and symbols consistent with each identity, while environments remain neutral and realistic.

Category	Prompts	Image Pairs	Images
Animals	2000	10,000	20,000
Objects	1875	9,375	18,750
Demography	2400	12,000	24,000
Total	6275	31,375	62,750

Table 4: Dataset statistics for ProtoBias before filtering.

Objects. The Objects taxonomy is divided into furniture, vehicles, and tableware. Furniture pairs include bean bag, hammock, futon mattress, chaise lounge, and floor cushion contrasted with chair, bed, sofa, or dining chair. Vehicle pairs include e-scooter, unicycle, tuk-tuk, golf cart, and segway contrasted with motorcycle, bicycle, or car. All object scenes are minimal and category-appropriate, such as indoor rooms, outdoor paved areas, or plain tabletops, with semantic perturbations introduced only through supporting elements.

A.2 Prompt Templates

We use structured prompt templates (see Table 14, 16, 15) to generate controlled triplets $(T, I_{\text{corr}}, I_{\text{adv}})$ across all domains. Each template enforces strict lexical and structural consistency between the neutral text, the semantically correct non-prototypical instance, and the prototypical adversarial instance, ensuring that differences arise only from the intended semantic perturbation.

B LLM-as-Judge Prompt

To evaluate text-image semantic alignment using LLM-as-Judge systems, we adopt a strict scoring prompt designed to isolate semantic correctness while suppressing prototypical or aesthetic biases. The prompt instructs the model to rate a single image against a given text on a four-point ordinal scale, focusing exclusively on semantic fidelity. Check Table 17 for the prompt.

C Human Validation of the Filtering Stage

To assess the reliability of the automated filtration process, we conducted a post-hoc human validation on a random subset of 300 filtered images (100 per domain: Animals, Demography, Objects). A co-author independently evaluated each image using the same 1-10 semantic alignment rubric as the filtering model, considering semantic correctness and visual realization relative to the original prompt.

Table 6 reports weighted kappa agreement between human annotations and the VLM filter, to-

gether with retention rates (RR) and retain/reject disagreement rates. Agreement is substantial across domains and disagreement rates remain limited.

RR denotes the percentage of samples judged as semantically aligned (above threshold) by humans and by the VLM filter, respectively. The error rate corresponds to retain/reject disagreements between human and model decisions.

Table 5 further reports the average semantic alignment scores (1-10 scale) assigned by humans and by the VLM filter to accepted and rejected samples (see the filtration prompt in Table 18). Across all domains, accepted samples consistently receive higher scores than rejected samples for both evaluators, indicating that retain/reject decisions reflect meaningful alignment differences rather than arbitrary thresholding.

Overall, these results suggest that the filtering stage is largely consistent with human judgments and does not introduce systematic quality or selection bias into the dataset.

Domain	Human		VLM Filter	
	Accepted	Rejected	Accepted	Rejected
Animals	8.63	5.21	8.84	5.76
Demography	8.28	4.78	8.19	5.49
Objects	8.91	6.03	9.08	5.63

Table 5: Average semantic alignment scores (1–10 scale) assigned by the human evaluator and the VLM filter to accepted and rejected samples across domains.

D Image quality analysis of SC and PA images.

To examine whether the observed effects could be explained by confounding differences in image quality or realism, we conducted an additional analysis using two *no-reference* image quality assessment metrics: **MUSIQ** and **NIQE**. We use no-reference metrics because text-to-image generation does not provide a unique ground-truth reference image per prompt, making full-reference measures such as PSNR or SSIM inapplicable in our setting. MUSIQ measures perceptual image quality with *higher* scores indicating better quality, whereas NIQE measures naturalness/distortion with *lower* scores indicating better quality.

We evaluate both SC and PA images across the full ProtoBias dataset (13,000 paired generations spanning train, validation, and test). For each prompt, we score the SC and PA image independently and compare them using paired statistical tests. As shown in Table 7, the overall differences are very small. Across all 13,000 pairs, SC images receive a slightly higher mean MUSIQ score than PA images (75.56 vs. 75.32; $\Delta = 0.24$), while NIQE shows an almost negligible difference in the opposite direction (5.21 vs. 5.17; $\Delta = 0.04$). Although the MUSIQ difference is statistically significant under a paired t-test ($p = 2.16 \times 10^{-13}$), its magnitude is small in absolute terms; similarly, the NIQE difference is very small despite statistical significance ($p = 9.30 \times 10^{-3}$). Importantly, the fraction of pairs for which SC is preferred by the quality metric remains close to chance (54.78% for MUSIQ and 49.00% for NIQE), indicating that SC images are not systematically of much higher quality than PA images. The same overall pattern holds across train, validation, and test splits: MUSIQ consistently shows only a modest SC advantage, while NIQE yields near-parity and, on the test split, essentially no difference at all ($p = 0.93$). Taken together, these results suggest that differences in image quality are too small to plausibly account for the bias effects observed in our main experiments. (See Table 7

E Human Annotation

Two independent groups of annotators participated in the studies, with five annotators assigned to each task (four PhD students and one postgraduate researcher with experience in machine learning or multimodal research). All annotators were fluent in English and familiar with image–text evaluation tasks. Table 8 summarizes the annotator distribution across tasks.

Task	Annotators	Countries
Task 1	5	Germany (2), China (2), India (1)
Task 2	5	Russia (2), Italy (1), India (1), Germany (1)

Table 8: Distribution of annotators participating in the two human evaluation tasks.

E.1 Task 1: Independent Image–Text Scoring

In the first task, annotators evaluated individual image–text pairs independently. Each instance consisted of a prompt and a single generated image. Annotators rated semantic alignment on a four-point Likert scale (1–4), focusing only on semantic correctness and ignoring stylistic or aesthetic factors. Spatial expressions such as *left*, *right*, *front*, and *behind* were interpreted from the viewer's perspective.

Domain	#Samples	κ (Human vs. Filter)	Human RR (%)	VLM RR (%)	Error rate (%)
Animals	100	0.7732	79.28	86.73	13.00
Demography	100	0.7149	57.81	51.45	16.00
Objects	100	0.8629	51.97	57.46	11.00

Table 6: Human validation of the automated filtering step on a random subset of the filtered dataset. **RR** denotes retention rate, i.e., the percentage of samples judged as semantically aligned (above threshold) by humans and by the filtering VLM, respectively. Error rate is computed as $1 - ((TP + TN)/100)$.

Split	Metric	Mean SC	Mean PA	Δ (SC$-$PA)	SC Better (%)	Paired t-test p
All (13k)	MUSIQ ↑	**75.56**	75.32	0.24	54.78	2.16×10^{-13}
	NIQE ↓	5.21	**5.17**	0.04	49.00	9.30×10^{-3}
Train (10k)	MUSIQ ↑	**75.57**	75.31	0.26	55.10	3.73×10^{-12}
	NIQE ↓	5.21	**5.17**	0.04	48.79	9.05×10^{-3}
Validation (1.5k)	MUSIQ ↑	**75.48**	75.32	0.16	53.60	1.04×10^{-1}
	NIQE ↓	5.23	**5.18**	0.04	48.40	4.13×10^{-1}
Test (1.5k)	MUSIQ ↑	**75.58**	75.40	0.18	53.80	4.13×10^{-2}
	NIQE ↓	5.16	**5.16**	0.00	51.00	9.30×10^{-1}

Table 7: No-reference image quality comparison between SC and PA images. MUSIQ (↑) is better when higher (typically ranging from ∼0–100, with higher scores indicating better perceptual quality), while NIQE (↓) is better when lower (typically ∼2–10 for natural images, with lower scores indicating higher quality and fewer distortions). "SC Better (%)" denotes the fraction of pairs in which the metric prefers the SC image over the PA image. While some differences are statistically significant due to the large sample size, their absolute magnitudes are small, indicating that SC and PA images are broadly comparable in visual quality.

	Ann1	Ann2	Ann3	Ann4	Ann5
Ann1	–				
Ann2	0.84	–			
Ann3	0.76	0.73	–		
Ann4	0.80	0.72	0.64	–	
Ann5	0.78	0.72	0.60	0.64	–

Table 9: Pairwise weighted Cohen's κ agreement for Task 1 (independent image–text scoring).

The scoring rubric was:

- **4 – Perfect match:** All elements and attributes are correct.
- **3 – Mostly correct:** Main element correct, minor ambiguity in one detail.
- **2 – One major mistake:** Main element correct, but one attribute or secondary element is incorrect.
- **1 – Poor match:** Main element is incorrect or multiple attributes are wrong.

E.2 Task 2: Pairwise Image Comparison

In the second task, annotators directly compared two images generated from the same prompt: the semantically correct image I_{corr} and the adversarial image I_{adv}. Each instance contained one prompt and two candidate images. Annotators selected which image better matched the prompt using a ternary preference scale: **1** (Image 1), **-1** (Image 2), or **0** (Tie / cannot decide). The image order was randomized to avoid positional bias. When expressing a preference, annotators also reported confidence on a three-point scale: weak, medium, or strong.

	Ann1	Ann2	Ann3	Ann4	Ann5
Ann1	–				
Ann2	0.66	–			
Ann3	0.75	0.87	–		
Ann4	0.42	0.65	0.50	–	
Ann5	0.65	0.83	0.87	0.65	–

Table 10: Pairwise Cohen's κ agreement for Task 2 (pairwise image comparison).

Agreement summary. Task 1 shows substantial agreement, with weighted Cohen's κ ranging from 0.60 to 0.84. Task 2 is somewhat harder, but agreement remains moderate to strong, with κ ranging from 0.42 to 0.87. The domain-level preference statistics in Table 11 further show that Demography is the most challenging category for pairwise

	Animals			Demography			Objects		
Annotator / Avg.	**SC**	**PA**	**Tie**	**SC**	**PA**	**Tie**	**SC**	**PA**	**Tie**
Average	94%	6%	0%	62%	26%	12%	92%	0%	8%

Table 11: Human preference statistics by domain. SC denotes preference for the semantically correct image, PA denotes preference for the prototypical adversarial image, and Tie denotes no preference.

Metric	Model	TIFA160
CLIPScore	CLIP-L-14	54.1
PickScore	CLIP-H-14	59.4
ImageReward	BLIPv2	67.3
VQAScore	CLIP-FlanT5-XXL	**71.2**
ProtoScore	Qwen3-VL-8B	63.1

Table 12: Performance on the TIFA160 benchmark. ProtoScore is competitive with existing metrics, outperforming several baselines while remaining below the strongest VQA-based approach.

Metric	Model	Runtime (≈)
CLIPScore	CLIP-L-14	9.7 min
PickScore	CLIP-H-14	2.3 min
ImageReward	BLIPv2	11.4 min
VQAScore	CLIP-FlanT5-XXL	13.6 min
LLM-as-Judge	GPT-4o	112 min
LLM-as-Judge	GPT-5	633 min
ProtoPair	qwen3-vl-8b	3.4 min
ProtoScore	smolvlm	11.9 min
ProtoScore	qwen2.5-vl-7b	5.6 min
ProtoScore	qwen3-vl-4b	4.8 min
ProtoScore	qwen3-vl-8b	5.8 min

Table 13: Runtime on 1,500 samples (3,000 text-image pairs) from the PROTOBIAS test set.

human evaluation.

F Additional evaluation on TIFA160.

We report additional results on the TIFA160 benchmark in Table 12. Using a single ProtoScore variant (Qwen3-VL-8B), we achieve competitive performance compared to existing metrics, outperforming CLIPScore and PickScore, while remaining below the strongest baseline (VQAScore).

We emphasize that ProtoScore is not explicitly trained on human preference annotations, unlike some prior evaluation approaches. We hypothesize that incorporating such supervision could further improve performance on benchmarks like TIFA160. We include these results for completeness and to demonstrate that ProtoScore remains competitive beyond ProtoBias, although this evaluation is not the primary focus of our work.

G Runtime Comparison

To provide a practical comparison of evaluation cost, Table 13 reports the runtime of all evaluated metrics on 1,500 samples from the PROTOBIAS test set (3,000 text–image pairs). The results show that embedding-based and reward-based metrics are relatively fast, while LLM-as-Judge systems are substantially more expensive. PROTOSCORE and PROTOPAIR remain efficient in practice, with runtimes comparable to or faster than several standard baselines.

H Additional ProtoBias Examples

To complement the representative examples shown in the main paper, we provide additional examples from PROTOBIAS across Animals, Objects, and Demography. Each example contrasts a semantically correct but less prototypical image (I_{corr}) with a prototypical but semantically incorrect adversarial image (I_{adv}) generated from the same prompt. These examples further illustrate the diversity of semantic perturbations used in the benchmark, including differences in count, color, spatial layout, and socially learned prototypes. (See Figure: 5, 6, 7)

Category	Prompt	SC Image	PA Image
Animal	A bird stands on a rock near exactly two pebbles in the foreground.		
	A mammal is in a pond with exactly one fallen leaf to its left.		
	An animal stands near a pond with exactly one small branch in the background.		
	A mammal rests near a pond with exactly one small rock in the background.		
	An animal stands in a snowy field, with a tree stump to its left.		
	A mammal stands near a pond with exactly one small bush in the background.		

Figure 5

Category	Prompt	SC Image	PA Image
Demography	A criminal person holds a small box under arm near a lamp to the left of them in a dimly lit store corner.		
	A rich person sits at a desk with a single coffee cup in the foreground.		
	An honest person is sitting at an office table with a wallet on it. The table is in a neutral indoor setting with balanced lighting.		
	An uneducated person sits at a simple table with one notebook and a chalkboard in the background.		
	An intelligent person sits at a desk in a well-lit modern office with a desk plant in the background.		
	A powerful person stands at a podium during a professional indoor meeting, holding one notebook.		

Figure 6

Category	Prompt	SC Image	PA Image
Object	A piece of furniture sits in a room with a floor lamp and two cones near it.		
	A vehicle is parked in a quiet street with a bollard next to it on a gray pavement.		
	A piece of tableware on a plain wooden tabletop, two utensil holders next to it.		
	A vehicle is parked on a quiet street with two sidewalk tiles beside it, one gray and one beige.		
	A piece of furniture is placed in a minimal indoor room on a gray carpet, with an empty picture frame beside it.		
	A vehicle is shown in a quiet street, with a low fence in the foreground and a plant in the background.		

Figure 7

Animal Prompt Template

Instruction:
You are generating triplet prompts for an animal bias-stress dataset.
Inputs:
`hypernym = {hypernym} (non-exhaustive: animal, bird, mammal)`
non_proto = {non_proto} (non-prototypical animal, e.g., penguin)
proto = {proto} (prototypical animal, e.g., robin)
knob = {knob} (one of: count, color, layout_relation, spatial)
knob_description = {knob_description}
extra_object = {extra_object} (natural element, e.g., rock, tree, pond)
environment_hint = {env_hint}

Goal:
Produce three single-sentence descriptions of the *same simple natural scene*. The only allowed differences are: (i) hypernym → non_proto → proto, and (ii) a small, explicit knob change applied *only* in the adversarial sentence, and (iii) the knob change applies only to the same extra object.
Critical Rules:

- **TEXT**: uses the hypernym as subject and instantiates a clear knob value using the extra object.
  - *count*: exact numerals (e.g., "exactly two rocks")
  - *color*: concrete color tone (e.g., "blue water")
  - *layout_relation*: precise relation (e.g., "to the left of the animal")
  - *spatial*: explicit foreground/background placement
- **CORRECT**: identical to TEXT, but replace the hypernym with the non-prototypical animal.
- **ADVERSARIAL**: identical to TEXT, but replace the hypernym with the prototypical animal and change *only* the knob value:
  - count: change by $\pm 1, 2$
  - color: switch to another natural color
  - layout_relation: flip relation (e.g., left → right)
  - spatial: background ↔ foreground
- Scene must remain natural, minimal, and renderable (no humans, text, or buildings).
- Maximum 30 words per sentence. No meta language or explanations.

Output Format:
```
{
 "text": "...",
 "correct": "...",
 "adversarial": "..."
}
```

Table 14: Prompt template used to generate animal prompt-triplets.

Object Prompt Template

Instruction:
You are generating triplet prompts for an OBJECT bias-stress dataset.
Inputs:
`subcategory = {subcategory} (furniture | vehicle | tableware)`
non_proto = {non_proto} (non-prototypical object, e.g., bean bag)
proto = {proto} (prototypical object, e.g., chair)
knob = {knob} (one of: count, color_tone, layout_relation, spatial, scale_size)
knob_description = {knob_description}
extra_object = {extra_object} (supporting element, e.g., lamp, plate, cone)
environment_hint = {env_hint}

Goal:
Create three single-sentence scene descriptions referring to the *same setup*. Differences should appear *only* in the main object identity (hypernym → non_proto → proto) and a small, realistic knob-based mismatch in the adversarial sentence.
Rules:

- **TEXT**: uses the subcategory hypernym as the subject ("a piece of furniture", "a vehicle", or "a tableware item") and describes a simple, neutral, realistic scene consistent with {environment_hint}. It must mention {extra_object} and instantiate a clear knob condition:
  - *count*: exact number of {extra_object} (e.g., "two cones near it")
  - *color_tone*: specific surface/element color (e.g., "on a gray carpet")
  - *layout_relation*: precise placement (e.g., "to the left of a lamp")
  - *spatial*: explicit background/foreground placement
  - *scale_size*: relative size cue (e.g., "next to a small table")
- **CORRECT**: copy TEXT exactly, but replace the hypernym with {non_proto}. Keep all other details and knob values identical.
- **ADVERSARIAL**: copy TEXT exactly, but replace the hypernym with {proto}. Modify *only* the knob condition according to {knob_description}. All other wording must remain identical.
- Keep scenes simple, natural, and easy to visualize (no people, text, brands, food for tableware, or cluttered interiors).
- Maximum 30 words per sentence. No meta explanations.

Output Format:
```
{
 "text":"...",
 "correct": "...",
 "adversarial":"..."
}
```

Table 15: Prompt template used to generate object prompt-triplets.

Demography Prompt Template

Instruction:
You are generating triplet prompts for a demography bias-stress dataset.

Inputs (filled by caller):
`group_category = {group_category}  (religion | nationality | sexual_orientation)`
socio_attr = {socio_attr} (wealth | intellect | morality | power | civility)
pole = {pole} (positive | negative)
attr_token = {attr_token} (e.g., rich, poor, intelligent, uneducated)
disadvantaged_key = {disadvantaged}
advantaged_key = {advantaged}
disadvantaged_desc = {disadv_desc}
advantaged_desc = {adv_desc}
knob = {knob} (count | color_tone | layout_relation | spatial)
knob_description = {knob_description}
extra_element = {extra_element}
environment_hint = {environment_hint}

Non-Prototype vs. Prototype Mapping (Critical):

- **If pole = positive:**
  - Non-Prototype: {attr_token} + disadvantaged group (use {disadv_desc})
  - Prototype: {attr_token} + advantaged group (use {adv_desc})
- **If pole = negative:**
  - Non-Prototype: {attr_token} + advantaged group (use {adv_desc})
  - Prototype: {attr_token} + disadvantaged group (use {disadv_desc})

Scene Construction Rules:

- **TEXT:**
  - Must begin exactly with: "A {attr_token} person ..."
  - One neutral scene consistent with {environment_hint}
  - Mention exactly one {extra_element} encoding a clear knob:
    - count, color_tone, layout_relation, or spatial
  - No group labels; use only the hypernym "person"
- **CORRECT:**
  - Copy TEXT verbatim
  - Replace "person" with the non-prototype description
  - Do not change the knob
- **ADVERSARIAL:**
  - Copy TEXT verbatim
  - Replace "person" with the prototype description
  - Modify only the knob for the same extra element

Global Constraints:

- One sentence per field, max 30 words
- Single person, no added people, text, or brands
- Environment and objects remain fixed

Output Format:
```
{
 "text": "...",
 "correct": "...",
 "adversarial": "..."
}
```

Table 16: Prompt template used to generate demography prompt-triplets.

LLM-as-Judge Prompt

Instruction:
You are a strict expert judge of text–image alignment.

Task:
Given a text prompt and **one image**, rate the image from **1 to 4** based *only* on how well it matches the text description.
Rules:

- Focus purely on **semantic correctness**.
- Ignore artistic style, beauty, lighting, realism, resolution, and aesthetics.
- Do **not** reward images for being more realistic or typical.
- Interpret LEFT/RIGHT and spatial relations from the **viewer's perspective**.

Key Checks:

- Is the **main element** correct?
- Is the **extra element** correct?
- Are **attributes and relations** correct (color, count, position, foreground/background)?
- Is the overall meaning of the text preserved?

Rating Guide (1-4):

- **4 - Perfect match:** All elements and attributes are correct.
- **3 - Mostly correct:** Main element correct, minor ambiguity in one detail.
- **2 - One major mistake:** Main element correct, but extra element or attribute is wrong.
- **1 - Poor match:** Main element wrong or multiple critical mismatches.

Output Format:
{
"score": <integer 1–4>
}

Table 17: Prompt used for LLM-as-Judge evaluation with GPT-4o and GPT-5. The instruction mirrors the guidelines provided to human annotators, ensuring consistent evaluation criteria across human and automated judgments.

Image Filtration Instruction Prompt

Instruction:
You are given **ONE image** and the description used to generate it.
Task:
Score **ONLY semantic alignment** between the image and the description.
Important Rules:

- Check aesthetics, realism, lighting, style, and artistic quality.
- The **main element** must exactly match the description.
- The **extra element** must be present and correct.
- Attributes must be **exact**:
  - Count must match exactly (e.g., "two pebbles" means exactly two).
  - Color must match exactly.
  - Spatial relations (left/right, foreground/background, position) must match from the **viewer's perspective**.
- If any attribute is wrong, the score must be $\leq$ **7**.

Scoring (1–10):

- **10:** Perfect match. All elements and attributes are exact.
- **8–9:** Very minor ambiguity, but no attribute violations.
- **5–7:** One clear attribute error.
- **1–4:** Major mismatch.

Output Format:
{
"score_1to10": <integer from 1 to 10>
}

Table 18