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 they may reward images that look plausible or prototypical rather than images that faithfully satisfy the prompt. We identify *prototypicality bias* as a systematic blindspot in multimodal evaluation: metrics can prefer a semantically incorrect but visually or socially prototypical image over a correct but less prototypical one. We introduce PROTOBIAS, a controlled diagnostic benchmark across Animals, Objects, and Demography, where semantically correct images are contrasted with plausible prototypical adversaries containing a single controlled semantic violation. Grounded in prototype theory and social-category prototypicality, PROTOBIAS is constructed with multiple prompt generators, image generators, and independent VLM filters, and validated through prompt-quality, human-annotation, and image-quality controls. Using PROTOBIAS, we show that widely used embedding, reward, VQA-based, and VLM-as-judge metrics frequently fail these contrasts, while human judgments remain more faithful to semantic correctness. We further introduce PROTOSCORE, a lightweight contrastively trained evaluator, as an initial mitigation baseline. PROTOBIAS provides a focused benchmark for measuring prototypicality-driven metric failures and developing more semantically faithful T2I evaluators.

1 Introduction

Automatic metrics are now central to text-to-image (T2I) evaluation. They are used to compare models, select checkpoints, filter large-scale generated data, and report progress when human evaluation is too expensive to run at scale. This makes their failure modes consequential: if a metric systematically rewards the wrong image, that error can shape model development and benchmark conclusions. In this

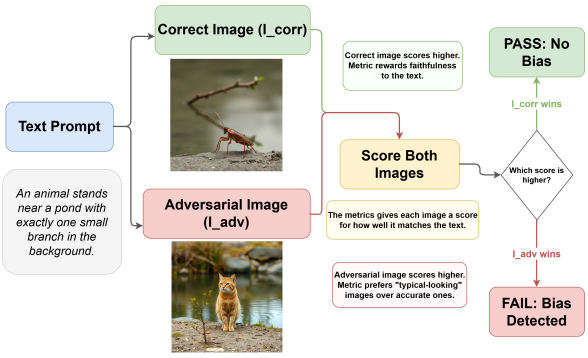

Figure 1: Illustration of prototypicality bias in T2I evaluation. The prompt asks for an animal near a pond with exactly one small branch in the background. The semantically correct image (I_{corr}) satisfies the specified branch condition but depicts a less prototypical animal. The prototypical adversarial image (I_{adv}) depicts a more familiar animal, but violates the branch condition. A metric passes this diagnostic case if it scores I_{corr} higher than I_{adv}.

paper, we study one such failure mode: metrics may prefer images that look familiar, canonical, or socially typical, even when those images violate the prompt. We call this failure mode *prototypicality bias*. The term builds on prototype theory, which shows that categories have graded internal structure: some instances are perceived as more central or representative than others (Rosch, 1975; Rosch and Mervis, 1975; Rosch et al., 1976; Murphy, 2002). For example, a robin is a more prototypical bird than a penguin, and a chair is a more prototypical piece of furniture than a bean bag. Prototypicality is useful for human categorization, but it is not the same as semantic correctness. An image can look like a familiar category member while still failing to satisfy the prompt.

In T2I evaluation, this distinction matters because automatic metrics are widely used as scalable substitutes for human judgment in benchmarking, model selection, and data filtering (Hessel et al., 2022; Kirstain et al., 2023; Xu et al., 2023; Lin

et al., 2024; Ge et al., 2023; Singh et al., 2026). We define **prototypicality bias** as a metric-level blindspot in which a visually or socially prototypical image is scored as more faithful to the prompt than a semantically correct but less prototypical image. Figure 1 illustrates the setup. The prompt asks for an animal near a pond with exactly one small branch in the background. The correct image I_{corr} depicts a cockroach and satisfies the branch constraint, whereas the adversarial image I_{adv} depicts a more prototypical animal, a cat, but violates the branch condition. Thus, both images are plausible under the broad category term "animal", but only I_{corr} satisfies the full prompt. Such bias is problematic because metrics trained on large-scale vision–language data may learn to reward canonical visual forms or socially dominant associations rather than semantic faithfulness (Girrbach et al., 2025; Wan et al., 2024), thereby overestimating models that generate familiar but incorrect images and underestimating less common but correct generations.

To measure this failure mode, we introduce PROTOBIAS, a controlled diagnostic benchmark that places semantic correctness in direct conflict with prototypicality. Each instance pairs a neutral prompt T with a semantically correct image I_{corr} and a prototypical adversarial image I_{adv} that violates one visually checkable semantic attribute, such as count, color tone, spatial relation, scale, or foreground/background placement. The benchmark spans **Animals**, **Objects**, and **Demography**: the first two test category typicality, while Demography tests socially learned demographic–attribute associations rather than intrinsic identity typicality. By using controlled contrastive examples, PROTOBIAS isolates prototypicality-driven metric failures that may be hidden by aggregate benchmark scores (He et al., 2023b; Chen and Eger, 2023; Thrush et al., 2022).

In summary, we make the following contributions:

- **Bias formulation.** We define *prototypicality bias* as a metric-level blindspot in which visually or socially prototypical images are preferred despite explicit semantic violations.
- **Diagnostic benchmark.** We introduce PROTOBIAS, a controlled benchmark spanning **Animals**, **Objects**, and **Demography**. Each example is a contrastive triplet $(T, I_{\text{corr}}, I_{\text{adv}})$ designed to isolate semantic correctness from prototypical familiarity. We validate the benchmark through prompt-quality checks, VLM-based image filtering, human validation of filtering decisions, pairwise human evaluation, and image-quality controls.
- **Metric analysis and mitigation baseline.** We evaluate embedding-based, reward-based, VQA-based, and VLM-as-judge metrics on PROTOBIAS, showing that many fail specifically under neutral prompts where prototypicality conflicts with semantic correctness. As an initial mitigation baseline, we introduce PROTOSCORE, a contrastively trained evaluator for this diagnostic setting.

2 Related Work

Bias and prototypicality in multimodal systems. Large-scale text-to-image (T2I) and vision-language models (VLMs) frequently reproduce social and cultural biases (Wan et al., 2024; Luo et al., 2024). Extensive research documents representational disparities across demographics and social cues, driving the creation of benchmarks to measure harms in generated images and VLM predictions (Wan et al., 2024; Luo et al., 2024; Narayanan et al., 2025; Contreras, 2025; Nair et al., 2025; Seo et al., 2025; Said et al., 2025; Raj et al., 2025; Jiang et al., 2024; Girrbach et al., 2025). While related, our work targets a different object of study: the evaluation metrics themselves. Furthermore, PROTOBIAS builds on prototype theory, which posits that categories have graded structures where some instances are more central or representative (Rosch, 1975; Rosch and Mervis, 1975; Rosch et al., 1976; Murphy, 2002).

Robustness and diagnostic evaluation of T2I metrics. Automatic metrics—including embedding-based methods (CLIPSCORE (Hessel et al., 2022)), reward models (PICKSCORE, IMAGEREWARD (Kirstain et al., 2023; Xu et al., 2023)), VQA metrics (VQASCORE (Lin et al., 2024)), and LLM-as-judge systems (Ge et al., 2023; Singh et al., 2026)—are central to T2I evaluation. However, these proxies for human judgment often rely on spurious correlations, dataset artifacts, or superficial cues rather than genuine semantic grounding (Dai et al., 2024; Hirota et al., 2025; Ye et al., 2024; Hwang et al., 2025; Wang et al., 2025; Madaan et al., 2025; Lee et al., 2025). Similar to how meta-evaluation

exposes metric failures in NLP (He et al., 2023b; Chen and Eger, 2023), multimodal benchmarks like WINOGROUND (Thrush et al., 2022), CROC (Leiter et al., 2025), PREXME (Leiter and Eger, 2024), and METALOGIC (Shen et al., 2025) test contrastive robustness and preference inconsistencies under controlled variations. PROTOBIAS complements these diagnostic efforts by isolating a specific failure mechanism: whether metrics inherently prefer prototypical but semantically incorrect images over correct but non-prototypical ones. Thus, our dataset provides a controlled environment to rigorously evaluate the evaluators for prototypicality-driven biases.

3 Bias Formulation

We use *prototype* in the sense of graded category structure: within many categories, some instances are perceived as more central, familiar, or representative than others (Rosch, 1975; Rosch and Mervis, 1975; Rosch et al., 1976; Murphy, 2002). For example, a robin is a more prototypical bird than a penguin, and a chair is a more prototypical piece of furniture than a bean bag. We call this relative centrality *prototypicality*. Prototypicality is distinct from semantic correctness: an image may look canonical or familiar while still violating the prompt. We define *prototypicality bias* as a metric-level blindspot in which a visually or socially prototypical image is scored as more faithful to the prompt than a semantically correct but less prototypical image. In PROTOBIAS, each example contains a neutral prompt T, a correct image I_{corr}, and a prototypical adversarial image I_{adv}. The correct image satisfies all semantic requirements in T, while the adversarial image is more prototypical but violates one controlled requirement, such as count, color tone, spatial relation, or foreground/background placement. For Animals and Objects, prototypicality refers to canonical category members or layouts. For Demography, we do not claim that identities are intrinsically more or less typical; instead, we study *social prototypicality*: overrepresented demographic–attribute associations learned from data and related to stereotype-based social-category prototypes (Ma and Correll, 2011; Mehrabi et al., 2022; Wan et al., 2024). Given a metric $M(T, I)$, a semantically faithful metric should score the correct image higher than the adversarial one. We count a *prototypicality-bias failure* when:

$$M(T, I_{\text{adv}}) \geq M(T, I_{\text{corr}}). \quad (1)$$

That is, the metric assigns the prototypical but semantically wrong image a score at least as high as the correct image. This criterion directly tests whether prototypicality is strong enough to override prompt faithfulness.

4 PROTOBIAS Benchmark Construction

Figure 2 provides the roadmap for this section. We first describe how controlled prompt triplets (T, D_c, D_a) are constructed and validated, and then explain how these triplets are rendered into image pairs $(I_{\text{corr}}, I_{\text{adv}})$, filtered for semantic correctness, checked for visual-quality confounds, and validated through human evaluation.

4.1 Prompt Construction and Validation

PROTOBIAS is constructed through the pipeline shown in Figure 2. The first stage builds controlled prompt triplets (T, D_c, D_a), where T is the neutral prompt used for evaluation, D_c is the correct generation description, and D_a is the adversarial generation description. This stage defines the taxonomy, instantiates the prototype contrast, introduces one controlled semantic perturbation, filters invalid prompts, and validates that the resulting triplets are localized semantic edits rather than broad rewrites.

Taxonomy. PROTOBIAS spans three domains: **Animals**, **Objects**, and **Demography**. Animals and Objects instantiate prototypicality as category typicality: some category members are more central or familiar than others while remaining valid members of the same broad class (Rosch, 1975; Boster, 1988; Uyeda and Mandler, 1980; Joelson and Herrmann, 1978; Battig and Montague, 1968, 1969). Animals include contrasts such as *penguin–robin*, *platypus–dog*, and *walking stick insect–cat*. Objects cover furniture, vehicles, and tableware, contrasting peripheral instances such as *bean bag*, *golf cart*, or *tuk-tuk* with more canonical instances such as *chair*, *car*, or *bicycle*. The full taxonomy and prompt templates are provided in Appendix A.

Demography is handled separately. We do not claim that demographic identities are intrinsically more or less typical. Instead, we use prototypicality in a representational sense: socially learned and dataset-mediated associations between demographic groups, attributes, and contexts. Following work on social prototypes, stereotypes, and multimodal bias (Ma and Correll, 2011; McIntosh, 1988; Rich, 1980; Buolamwini and Gebru, 2018; Mehrabi et al., 2022; Janghorbani and de Melo, 2023; Mor-

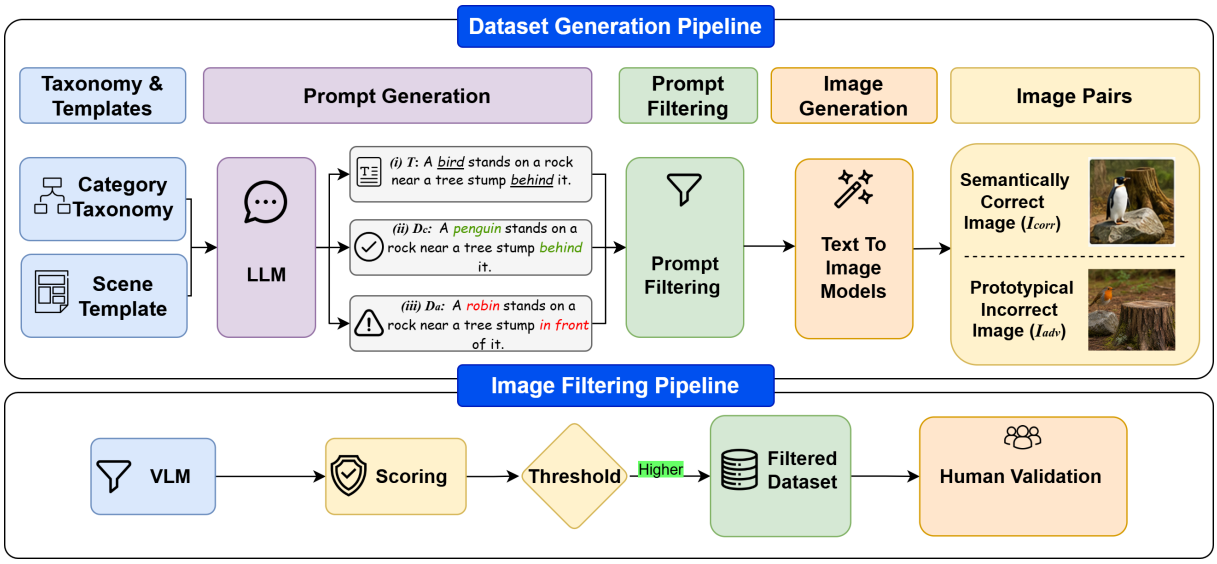

Figure 2: Overview of the PROTOBIAS construction pipeline. A taxonomy entry and scene template define a neutral prompt T, a correct description D_c, and an adversarial description D_a. In the illustrated example, T names only the broad category "bird," D_c instantiates the less-prototypical penguin while preserving the scene, and D_a instantiates the more-prototypical robin while changing one controlled spatial detail. The validated descriptions are rendered into I_{corr} and I_{adv}, filtered independently against their own descriptions, and finally evaluated under the same neutral prompt T.

eira et al., 2024), we instantiate three demographic axes—*Religion*, *Nationality*, and *Sexual Orientation*—crossed with five socio-attributes: *Wealth*, *Intellect*, *Morality*, *Power*, and *Civility*.

Prompt triplets. For each taxonomy entry, we construct a triplet (T, D_c, D_a). The neutral prompt T uses a broad category label such as "animal," "bird," "vehicle," or "person," and specifies a simple scene with one visually checkable semantic attribute. The correct description D_c replaces the broad label with the less-prototypical instance while preserving the scene. The adversarial description D_a replaces the broad label with the more-prototypical instance and changes exactly one controlled semantic attribute. As shown in Figure 2, T may describe a bird on a rock near a tree stump; D_c instantiates a penguin and preserves the scene, while D_a instantiates a robin but changes the spatial relation from *behind* to *in front of*. We use four adversarial knobs: *count*, *color tone*, *layout relation*, and *spatial placement*. These attributes are explicit in text, visually verifiable, and common in T2I faithfulness evaluation. The semantic change is applied only to an auxiliary scene element, such as stones or branches for animals, lamps or cones for objects, and books or laptops for people. This ensures that the adversarial image should be more prototypical, but wrong on one prompt-relevant detail.

Prompt generation and filtering. We generate candidate triplets (T, D_c, D_a) from structured scene templates using three instruction-tuned LLMs: Qwen2.5-14B-Instruct (Qwen et al., 2025), Gemma-3-12B-it (Team et al., 2025), and Llama-3.1-8B-Instruct (Grattafiori et al., 2024). Raw LLM outputs may be malformed, switch domains, substitute the wrong target instance, or change more than the intended semantic knob. We therefore filter candidate triplets with Qwen2.5-32B-Instruct (Qwen et al., 2025) and Mistral-Small-3.1-24B-Instruct (Mistral AI, 2025), retaining only prompts accepted by all filters.[1] This yields 8,225 text triplets: 2,932 Animals, 3,027 Demography, and 2,266 Objects. (Filtering prompts are provided in Appendix C.)

Prompt-quality validation. Before image generation, we validate that accepted triplets are controlled semantic perturbations rather than broad rewrites. We use BGE-large sentence embeddings (Xiao et al., 2024) to measure similarity among T, D_c, and D_a, and DeBERTa-v3 NLI (He et al., 2023a) as a lightweight text-only consistency check. The correct description D_c should remain compatible with T, while D_c and D_a should contradict each other on the controlled attribute. Table 1 shows high similarity between T and both descrip-

[1] Implementation settings, compute resources, and approximate construction runtimes are reported in Appendix B.

Prompt-quality check	Animal	Demo.	Object	Overall
Semantic preservation				
$\mathrm{sim}(T, D_c)$	0.877	0.793	0.886	0.848
$\mathrm{sim}(T, D_a)$	0.888	0.781	0.908	0.854
$\mathrm{sim}(D_c, D_a)$	0.803	0.757	0.851	0.800
NLI-based structural validity				
$T \to D_c$ ent./neu. (%)	94.2	100.0	100.0	97.9
$T \to D_a$ ent./neu. (%)	64.0	32.9	59.7	51.4
$D_c \leftrightarrow D_a$ contra. (%)	97.0	100.0	91.4	96.5

Table 1: Automatic prompt-quality validation for 8,225 accepted PROTOBIAS text triplets. T is the neutral prompt, D_c the correct description, and D_a the adversarial description. Similarity uses BGE-large sentence embeddings; NLI uses DeBERTa-v3 and is applied only to text pairs. The bidirectional contradiction score averages both directions between D_c and D_a.

tions (0.848 for T, D_c; 0.854 for T, D_a), while D_c and D_a are contradictory in 96.5% of cases. This indicates that the text triplets preserve the scene while introducing the intended semantic conflict.

4.2 Image Construction and Validation

The second stage renders the validated prompt triplets into image pairs and verifies that the resulting images support the intended contrastive evaluation. For each accepted triplet, D_c is rendered into the semantically correct image I_{corr}, and D_a is rendered into the prototypical adversarial image I_{adv}. Both images are evaluated under the same neutral prompt T in the final benchmark task.

Image generation. We render D_c and D_a separately using three T2I models: FLUX.1-schnell (Black Forest Labs, 2025), Stable Diffusion 3.5 Large Turbo (Stability AI, 2025), and Sana-1600M-1024px (Xie et al., 2024). This produces candidate pairs $(I_{\text{corr}}, I_{\text{adv}})$ from the accepted text triplets. Since the two images are generated from matched descriptions, the intended difference is the prototype contrast plus the single controlled semantic violation in D_a.

Image filtering. Image filtering is necessary because T2I models may fail to realize the requested subject, count, color, layout, or spatial relation. Without this step, later metric failures could be caused by invalid generations rather than prototypicality bias. We filter images with two independent VLMs, Qwen2.5-VL-32B-Instruct (Qwen et al., 2025) and InternVL3-38B-Instruct (Zhu et al., 2025). Each VLM receives one image-description pair and assigns a 1–10 semantic-correctness score:

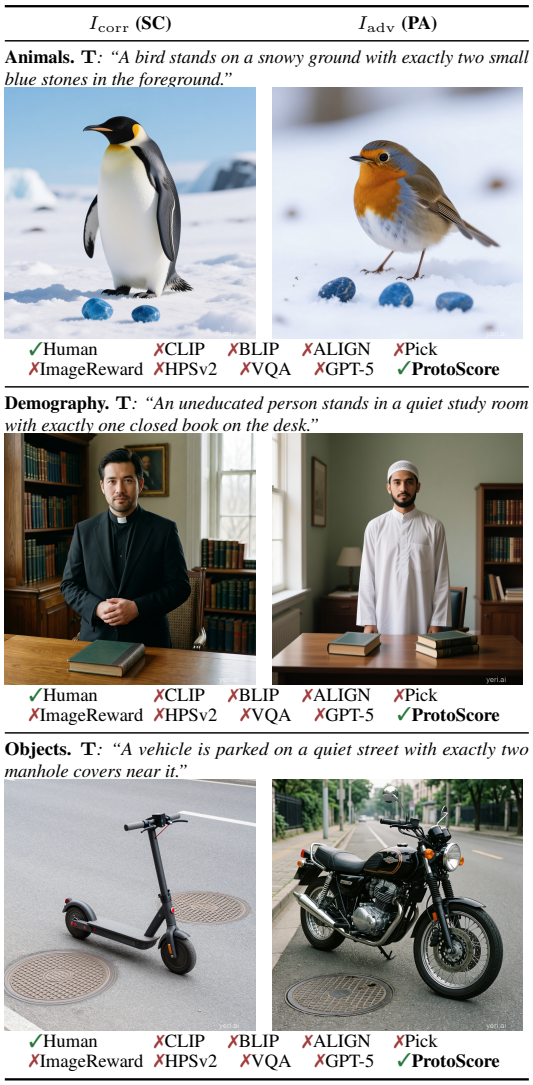

Figure 3: **PROTOBIAS examples.** Each row shows a neutral prompt T, a semantically correct image I_{corr} (SC), and a prototypical adversarial image I_{adv} (PA). SC satisfies the prompt, while PA is more prototypical but violates one controlled semantic attribute. Check marks indicate that the evaluator ranks SC above PA; crosses indicate that it ranks PA higher than SC.

I_{corr} is checked only against D_c, and I_{adv} only against D_a. Scores ≥ 8 are treated as valid because they indicate that the main subject and controlled attributes are present with at most minor ambiguity. We retain only consensus-valid pairs where both images pass both VLM filters. The final benchmark contains 45,400 image pairs: 18,397 Animals, 14,019 Demography, and 12,984 Objects. Crucially, image filtering is *non-contrastive*. It never compares I_{corr} and I_{adv} under the neutral prompt T; it only verifies that each image realizes its own generation description. Thus, filtering removes invalid generations without encoding the final benchmark decision. The final evaluation re-

mains contrastive: a metric receives the same neutral prompt T and must rank I_{corr} above I_{adv}. (See filtering prompts in Appendix C.)

Human validation of image filtering. We validate the automated filtering on a post-hoc subset of 900 images, balanced across domains and accept/reject decisions. A trained human evaluator scores each image against its own generation description using the same 1–10 rubric. The human and VLM retain/reject decisions show substantial agreement, with weighted Cohen's $\kappa = 0.778$ overall and 14.11% binary disagreement. Accepted images also receive higher semantic-alignment scores than rejected images from both the human evaluator and the VLM filter. Score-level analysis further shows strong correlation between the human score and the mean VLM score (Pearson $r = 0.718$, Spearman $\rho = 0.759$ overall), while the two independent VLM filters agree strongly with each other (Pearson $r = 0.881$). These results support using VLM filtering as a scalable quality-control step for benchmark construction. Full domain-level results are reported in Appendix D.

Image-quality control. We also check whether the retained I_{corr} and I_{adv} images differ systematically in visual quality, since such differences could confound metric failures. Because T2I images have no unique ground-truth reference, we use two no-reference image-quality metrics as perceptual sanity checks, not semantic alignment metrics: MUSIQ (Ke et al., 2021), where higher scores indicate better perceived quality, and NIQE (Zvezdakova et al., 2019), where lower scores indicate more natural images with fewer distortions. As shown in Table 2, the differences are negligible: MUSIQ is 76.34 for I_{corr} and 76.21 for I_{adv} ($\Delta = 0.13$), while NIQE is 4.93 and 4.91, respectively ($\Delta = 0.02$). The fraction of pairs where I_{corr} has better quality is also near chance under both metrics (52.47% for MUSIQ; 50.02% for NIQE). Thus, later metric failures are unlikely to be driven by systematic visual-quality differences.

Pairwise human evaluation. Finally, we collect human preferences for the final benchmark task. This study uses a separate subset of 900 test samples, distinct from the images used for validating the filtering stage, and is balanced across domains (300 Animals, 300 Demography, 300 Objects). Six annotators participated, including PhD researchers, a postdoctoral researcher, and master's-level graduates from multiple countries; annotator details and agreement statistics are reported in Appendix E. For each item, annotators see the neutral prompt T and the two images in randomized left–right order, and choose whether I_{corr}, I_{adv}, or neither image better matches the prompt. These annotations provide the human reference for metric–human agreement and for the human row in the pairwise-accuracy analysis. Figure 3 shows representative PROTOBIAS examples from all three domains. In each row, I_{corr} satisfies the neutral prompt, while I_{adv} is more prototypical but violates one controlled semantic attribute. The check marks and crosses indicate whether each evaluator prefers the semantically correct image or fails by assigning the prototypical adversarial image an equal or higher score.

Metric	**Mean I_{corr}**	**Mean I_{adv}**	Δ	**I_{corr} Better (%)**
MUSIQ ↑	**76.34**	76.21	0.13	52.47
NIQE ↓	4.93	**4.91**	0.02	50.02

Table 2: No-reference image-quality comparison between I_{corr} and I_{adv} across the whole PROTOBIAS. Small absolute differences and near-chance I_{corr}-preference rates indicate that the two image sets are broadly comparable in visual quality.

5 Experimental Setup

This section describes how we evaluate metrics on PROTOBIAS. We specify the data split, the metric families, the pairwise scoring rule, and the training setup for PROTOSCORE.

Evaluation Data We split the filtered PROTOBIAS image pairs into disjoint train and test partitions at the prompt-triplet level, ensuring that variants of the same underlying triplet do not appear across splits. The train split ($12k$ samples) is used only for training PROTOSCORE, while all reported metric evaluations are conducted on the test split. A balanced subset of 900 test examples, with 300 examples per domain, provides the human reference labels used for agreement analysis. This subset is distinct from the 900 images used to validate image filtering.

Evaluated Metrics We evaluate metrics from four families. First, we evaluate embedding-based metrics: CLIPScore (Hessel et al., 2022), AlignScore (Zha et al., 2023), and BLIPv2Score (Li et al., 2023). Second, we evaluate human-feedback-based metrics: PickScore (Kirstain et al., 2023), ImageRe-

ward (Xu et al., 2023), and HPSv2 (Wu et al., 2023). Third, we evaluate the VQA-based VQAScore (Lin et al., 2024). Finally, we evaluate VLM-as-judge baselines using GPT-5 (Singh et al., 2026) and Qwen3-VL-4B (Team, 2025). The VLM-as-judge baselines are prompted to judge only semantic correctness, following Appendix E. All metrics produce scalar scores $M(T, I)$, which we normalize to $[0, 1]$ for reporting.

Pairwise Evaluation Protocol For each example, a metric receives the neutral prompt T and scores both images. Following Eq. 1, we count a prototypicality-bias failure when $M(T, I_{\text{adv}}) \geq M(T, I_{\text{corr}})$. Pairwise accuracy is the fraction of examples where the metric avoids this failure and assigns a higher score to I_{corr}. All reported values in Table 3 are computed on the same annotated human-reference subset. Pairwise accuracy measures how often a metric ranks I_{corr} above I_{adv} under the neutral prompt T. Cohen's κ measures chance-corrected agreement between binary metric decisions and human choices.

Training PROTOSCORE We train PROTOSCORE as a contrastively trained evaluator by adapting Qwen3-VL-4B-Instruct into a pairwise image–text scorer. For each training example, the model receives the same neutral prompt T paired with I_{corr} and I_{adv}, and is trained to assign a higher score to the semantically correct image: $s(T, I_{\text{corr}}) > s(T, I_{\text{adv}})$. We optimize a scalar scoring head with a pairwise softplus ranking loss (Li et al., 2017; Yang et al., 2024) using margin 0.5. The base model is trained on the PROTOBIAS train split with LoRA adapters ($r = 16$, $\alpha = 32$, dropout 0.05) and a learned score head. We use AdamW with learning rate 1×10^{-5}, weight decay 0.01, cosine scheduling with 3% warmup, gradient clipping at 1.0, and bfloat16 mixed precision. We train for one epoch on a single H100 GPU for approximately six hours, with batch size 1, gradient accumulation over 16 steps, 512×512 images, and a maximum of 192 text tokens. PROTOSCORE is intended as an initial mitigation baseline for this diagnostic setting, not as a general-purpose replacement for existing T2I evaluation metrics.

6 Results & Analysis

Having validated the construction of PROTOBIAS, we evaluate whether existing T2I metrics remain faithful to the neutral prompt T when a plausible prototypical distractor is available. Table 3 reports two quantities. The primary metric is pairwise accuracy: a method is correct only if it assigns I_{corr} a higher score than I_{adv} under the same neutral prompt T. We also report Cohen's κ, measuring chance-corrected agreement between each metric's pairwise choice and the human majority preference on the 900-example annotated subset. For the Human row, accuracy is the fraction of examples where the human majority selects I_{corr}, and κ is the average pairwise quadratic weighted Cohen's κ among annotators.

Human evaluation. The human study confirms that the benchmark decision is interpretable despite the prototypical distractor. Human majority preferences select I_{corr} in 0.91, 0.86, and 0.82 of Animal, Demography, and Object pairs, respectively. Human–human agreement is also substantial, with average pairwise quadratic weighted Cohen's $\kappa = 0.705$, 0.648, and 0.691 across domains. Full annotator information, instructions, and the pairwise agreement matrix are reported in Appendix E.

Existing metrics fail under prototypical distractors. Most standard metrics perform far below the human reference. Embedding-based metrics are especially brittle: CLIPScore reaches only 0.25, 0.28, and 0.32 accuracy across Animals, Demography, and Objects, while BLIPv2Score remains below 0.43 in all domains. Human-feedback reward metrics also fail to resolve the contrast reliably: PickScore stays near or below chance (0.32, 0.34, 0.31), and HPSv2 performs similarly on Animals and Demography. ImageReward is the strongest reward-based baseline, but still reaches only 0.52, 0.48, and 0.62, well below human performance. Thus, metrics trained for broad alignment or preference prediction do not necessarily preserve the controlled semantic distinction in PROTOBIAS.

Semantic judges improve only partially. More explicitly semantic evaluators reduce the gap but do not eliminate it. VQAScore reaches 0.51, 0.47, and 0.55, while GPT-5 reaches 0.56, 0.51, and 0.54. Qwen3-VL-4B performs poorly as a zero-shot judge, with accuracies of 0.27, 0.38, and 0.32. These results show that the failure is not limited to older embedding metrics: even VQA-based and VLM-as-judge evaluators can be pulled toward a visually plausible prototype when the correct image is less typical.

Method	Animals		Demography		Objects	
	κ	Acc.	κ	Acc.	κ	Acc.
Human semantic-correctness preferences						
Human	0.705	***0.91***	0.648	***0.86***	0.691	***0.82***
Embedding-based text–image metrics						
CLIPScore	0.001	0.25	-0.005	0.28	0.058	0.32
AlignScore	0.033	0.51	-0.064	0.34	0.011	0.32
BLIPv2Score	0.057	0.33	-0.061	0.39	0.028	0.43
Human-feedback-based reward metrics						
PickScore	0.001	0.32	-0.014	0.34	0.126	0.31
ImageReward	0.084	0.52	0.137	0.48	0.192	0.62
HPSv2	0.024	0.25	-0.006	0.38	0.125	0.38
VQA-based metric						
VQAScore	0.028	0.51	-0.007	0.47	0.065	0.55
VLM-as-judge baselines						
GPT-5	0.108	0.56	0.176	0.51	**0.265**	0.54
Qwen3-VL-4B	0.045	0.27	0.040	0.38	0.011	0.32
Contrastively trained evaluator (Ours)						
PROTOSCORE	**0.182**	**0.95**	**0.232**	**0.91**	0.178	**0.98**

Table 3: Main results on PROTOBIAS. For automatic methods, κ is Cohen's chance-corrected agreement between the metric's pairwise choice and the human majority preference on the 900-example annotated test subset. For the Human row, κ is the average pairwise quadratic weighted Cohen's κ among annotators. Acc. is pairwise accuracy under the neutral prompt T: a method is correct if $M(T, I_{\text{corr}}) > M(T, I_{\text{adv}})$. For the Human row, Acc. is the fraction of annotated examples where the human majority prefers I_{corr}. Higher is better for both metrics. Bold indicates the best automatic method in each column; human accuracy is shown as an upper reference.

Metric–human agreement remains weak. The κ values show that low accuracy is accompanied by weak agreement with human semantic preferences. CLIPScore, VQAScore, and Qwen3-VL-4B have κ values close to zero across domains. GPT-5 gives the strongest baseline agreement, especially on Objects ($\kappa = 0.265$), but remains far below human–human agreement. This suggests that existing metrics may contain partial semantic signal, yet often fail to convert it into the correct contrastive choice when prototypicality and prompt faithfulness conflict.

Contrastive training reduces the failure but does not solve alignment. PROTOSCORE achieves high pairwise accuracy (0.95, 0.91, 0.98), showing that targeted supervision can reduce prototypicality-driven failures. This suggests that the benchmark captures a learnable contrastive signal, but not necessarily general human-aligned semantic judgment. Yet its agreement with human preferences remains modest ($\kappa = 0.182$, 0.232, 0.178); in Objects, GPT-5 has higher κ despite lower accuracy. We therefore view PROTOSCORE as a mitigation baseline, not a full solution. External transfer results of PROTOSCORE are reported in Appendix H.

Overall, PROTOBIAS shows that current T2I metrics can over-reward prototypical patterns when semantic correctness requires a less typical but prompt-faithful image.

7 Conclusion

We introduce PROTOBIAS, a controlled benchmark for testing whether T2I evaluation metrics remain faithful when prototypicality offers a shortcut. Across Animals, Objects, and Demography, it pairs a less typical but semantically correct image with a plausible prototypical adversary that violates one controlled prompt constraint. Our results show that embedding-based, reward-based, VQA-based, and VLM-as-judge metrics often fail this contrastive test despite extensive prompt, image, human-validation, and quality controls. This makes prototypicality bias measurable rather than anecdotal. While PROTOSCORE shows that contrastive training can reduce this failure, its limited human agreement indicates that benchmark-specific gains are not a complete solution. PROTOBIAS therefore provides a focused diagnostic for future evaluators: robust metrics should reward prompt faithfulness, not proximity to familiar prototypes.

Limitations

PROTOBIAS is designed as a controlled contrastive benchmark, which helps isolate prototypicality bias but cannot capture the full complexity of open-ended T2I evaluation. Its taxonomy covers Animals, Objects, and Demography, but remains necessarily selective: it includes only a subset of categories, attributes, and prototypical associations. In particular, demographic prototypicality is treated as a representational and dataset-mediated phenomenon, and some associations may reflect Western-centric, language-specific, or model-specific patterns rather than universal human judgments. Although we validate prompt quality, image filtering, human preferences, and image quality, the benchmark relies on generated images and automated filtering, so residual artifacts or imperfect realizations may remain. Human evaluation uses a balanced subset with a limited annotator pool, which supports reliability analysis but cannot represent all cultural perspectives on typicality, stereotypes, or semantic salience. Finally, PROTOSCORE is included as an initial mitigation, not as a general-purpose replacement for T2I metrics; its behavior may depend on the contrastive training distribution and should be tested beyond PROTOBIAS. Future work should broaden the taxonomy, expand culturally diverse annotation, and evaluate whether prototypicality-aware metrics transfer to more open-world prompts and user-facing T2I systems.

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.

Black Forest Labs. 2025. black-forest-labs/FLUX.1-schnell. https://huggingface.co/black-forest-labs/FLUX.1-schnell. Hugging Face model card.

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*.

Daniel Deutsch, George Foster, and Markus Freitag. 2023. Ties matter: Meta-evaluating modern metrics with pairwise accuracy and tie calibration. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12914–12929, Singapore. Association for Computational Linguistics.

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*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023a. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *Preprint*, arXiv:2111.09543.

Tianxing He, Jingyu Zhang, Tianle Wang, Sachin Kumar, Kyunghyun Cho, James Glass, and Yulia Tsvetkov. 2023b. 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.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *Preprint*, arXiv:2301.12597.

Yuncheng Li, Yale Song, and Jiebo Luo. 2017. Improving pairwise ranking for multi-label image classification. *Preprint*, arXiv:1704.03135.

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.

Mistral AI. 2025. mistralai/Mistral-Small-3.1-24B-Base-2503. https://huggingface.co/mistralai/Mistral-Small-3.1-24B-Base-2503. Hugging Face model card.

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.

Gregory L. Murphy. 2002. *The Big Book of Concepts*. The MIT Press.

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*.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

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.

Eleanor Rosch and Carolyn B Mervis. 1975. Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, 7(4):573–605.

Eleanor Rosch, Carolyn B Mervis, Wayne D Gray, David M Johnson, and Penny Boyes-Braem. 1976. Basic objects in natural categories. *Cognitive Psychology*, 8(3):382–439.

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.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, Akshay Nathan, Alan Luo, Alec Helyar, Aleksander Madry, Aleksandr Efremov, Aleksandra Spyra, Alex Baker-Whitcomb, Alex Beutel, Alex Karpenko, and 467 others. 2026. Openai gpt-5 system card. *Preprint*, arXiv:2601.03267.

Stability AI. 2025. stabilityai/stable-diffusion-3.5-large-turbo. https://huggingface.co/stabilityai/stable-diffusion-3.5-large-turbo. Hugging Face model card.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

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*.

Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. 2023. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *Preprint*, arXiv:2306.09341.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2024. C-pack: Packed resources for general chinese embeddings. *Preprint*, arXiv:2309.07597.

Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, and Song Han. 2024. Sana: Efficient high-resolution image synthesis with linear diffusion transformers. *Preprint*, arXiv:2410.10629.

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.

Weiqin Yang, Jiawei Chen, Xin Xin, Sheng Zhou, Binbin Hu, Yan Feng, Chun Chen, and Can Wang. 2024. Psl: Rethinking and improving softmax loss from pairwise perspective for recommendation. *Advances in Neural Information Processing Systems*, 37:120974–121006.

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*.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. Alignscore: Evaluating factual consistency with a unified alignment function. *Preprint*, arXiv:2305.16739.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, and 32 others. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *Preprint*, arXiv:2504.10479.

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.

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 12, 14, 13) 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 Construction Models, Infrastructure, and Runtime

Table 4 summarizes the model families and compute used for constructing PROTOBIAS. Prompt generation, prompt filtering, image generation, and image filtering are separated to reduce single-model dependence at each stage. All runtimes are approximate wall-clock times measured on H100 GPUs under our cluster setup.

Stage	Models	GPUs	Wall time	Output / purpose
Prompt generation	Qwen2.5-14B-Instruct (Qwen et al., 2025); Gemma-3-12B-it (Team et al., 2025); Llama-3.1-8B-Instruct (Grattafiori et al., 2024)	3 H100	∼3 h per generator	Candidate triplets (T, D_c, D_a) from structured templates.
Prompt filtering	Qwen2.5-32B-Instruct (Qwen et al., 2025); Mistral-Small-3.1-24B-Instruct (Mistral AI, 2025)	3 H100	∼1 h per filter	Removes malformed JSON, domain errors, target substitutions, and multi-knob changes.
Image generation	FLUX.1-schnell (Black Forest Labs, 2025); Stable Diffusion 3.5 Large Turbo (Stability AI, 2025); Sana-1600M-1024px (Xie et al., 2024)	9 H100	∼3–4 h per generator	Renders D_c and D_a into candidate image pairs $(I_{\text{corr}}, I_{\text{adv}})$.
Image filtering	Qwen2.5-VL-32B-Instruct (Qwen et al., 2025); InternVL3-38B-Instruct (Zhu et al., 2025)	36 H100	∼1 h per filter	Scores each image against its own generation description and retains consensus-valid pairs.

Table 4: Construction models and approximate runtime for PROTOBIAS. The table reports wall-clock time under our H100 cluster setup. Prompt generation and filtering operate on text triplets; image generation renders the accepted descriptions; image filtering validates each generated image independently against its own description.

C Prompt and Image Filtering Prompts

Tables 15–17 provide the LLM prompt-filtering instructions used during benchmark construction. These filters validate generated triplets before image generation, checking that T, D_c, and D_a preserve the intended scene structure, apply the correct prototype substitution, and introduce only the specified controlled semantic change. Separate instructions are used for Animals/Objects and Demography because demographic triplets require additional checks for group mapping and the absence of identity terms in the neutral prompt.

Tables 18–20 show the VLM image-filtering instructions used during benchmark construction. Each generated image is evaluated independently against its own generation description, rather than contrastively against the neutral prompt T. This keeps filtering separate from the final PROTOBIAS decision and prevents the filter from encoding the benchmark label.

D Human Validation of the Filtering Stage

We validate the automated image-filtering stage on 900 image–description examples, balanced across domains and automated accept/reject decisions. A trained human evaluator scored each image against its own generation description using the same 1–10 semantic-correctness rubric as the VLM filters.

Table 5 shows substantial retain/reject agreement: the overall disagreement rate is 14.11%, and weighted Cohen's κ is 0.778. Table 6 shows that filter-accepted samples receive higher semantic-alignment scores than rejected samples from both the human evaluator and the VLM filter. Table 7 further shows strong score-level correspondence between the human score and mean VLM score (Pearson $r = 0.718$, Spearman $\rho = 0.759$), as well as strong agreement between the two VLM filters (Pearson $r = 0.881$). Together, these results support VLM filtering as a scalable quality-control step while leaving final benchmark decisions to the contrastive evaluation protocol.

Domain	Human		VLM Filter	
	Accepted	Rejected	Accepted	Rejected
Animals	8.74	6.17	9.16	6.21
Demography	8.82	5.21	8.41	5.71
Objects	8.79	6.91	9.27	6.81
Overall	8.78	6.10	8.95	6.24

Table 6: Average semantic-alignment scores on a 1–10 scale for samples accepted and rejected by the automated filter. Accepted samples receive higher scores from both the human evaluator and the VLM filter across all domains.

Domain	Human–VLM r	Human–VLM ρ	VLM–VLM r
Animals	0.730	0.787	0.919
Demography	0.671	0.698	0.820
Objects	0.752	0.793	0.906
Overall	0.718	0.759	0.881

Table 7: Score-level correlations for image-filter validation. Human–VLM correlations compare the human 1–10 semantic-alignment score with the mean VLM filter score. VLM–VLM correlation compares the two independent filtering models on the same images.

Domain	#Samples	κ	Human RR (%)	VLM RR (%)	Disagree (%)
Animals	300	0.784	55.33	50.00	11.00
Demography	300	0.721	43.67	50.00	18.00
Objects	300	0.828	57.33	50.00	13.33
Overall	900	0.778	52.11	50.00	14.11

Table 5: Human validation of the automated image-filtering stage. κ denotes weighted Cohen's agreement between human and filter judgments. RR denotes retention rate, i.e., the percentage of samples judged as semantically aligned by the human evaluator and by the VLM filter, respectively. Disagree denotes binary retain/reject disagreement.

E Human Preference Annotation

We conduct a pairwise human preference study to verify that the intended PROTOBIAS decision is clear to human annotators. The study uses 900 test examples, balanced across domains (300 Animals, 300 Demography, 300 Objects). This subset is distinct from the 900 image–description examples used for human validation of the filtering stage in Appendix D. Each example consists of the neutral prompt T, the semantically correct image I_{corr}, and the prototypical adversarial image I_{adv}.

Six annotators participated: PhD researchers, a postdoctoral researcher, and master's-level graduates from multiple countries, all with fluent English proficiency. Aggregate annotator information is shown in Table 8. For each item, annotators saw T and the two images in randomized left–right order, and selected one of three labels: I_{corr} better matches the prompt, I_{adv} better matches the prompt, or tie/no clear preference.

Annotation rubric. Annotators were instructed to judge only semantic alignment with T. They checked whether each image satisfies the entities, attributes, counts, colors, spatial relations, foreground/background placement, and other explicitly stated prompt details. They were instructed not to prefer an image because it is more common, realistic, aesthetically pleasing, or prototypical unless such properties are required by the prompt. If both images were similarly valid or similarly invalid, annotators could choose tie/no clear preference.

Annotator agreement. Table 9 reports pairwise quadratic weighted Cohen's κ between annotators. Agreement ranges from 0.405 to 0.874, with an average pairwise agreement of $\bar{\kappa} = 0.681$. Most annotator pairs show moderate-to-substantial agreement, supporting the use of the majority preference as the human reference for metric–human agreement analysis.

Annotators	Age groups	Background
6	18–24 (1), 25–30 (3) 31–45 (1), 45–60 (1)	PhD, postdoc, master's
Countries		
Russia (2), Italy (1), India (1), Germany (1), UAE (1)		

Table 8: Aggregate information about annotators participating in the human preference evaluation.

	ann1	ann2	ann3	ann4	ann5	ann6
ann1	–					
ann2	0.648	–				
ann3	0.753	0.874	–			
ann4	0.405	0.645	0.497	–		
ann5	0.648	0.826	0.874	0.645	–	
ann6	0.723	0.679	0.813	0.589	0.749	–

Table 9: Pairwise quadratic weighted Cohen's κ agreement between human annotators in the preference study.

F 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 21 for the prompt.

G Runtime Comparison

To compare evaluation cost, Table 10 reports runtimes on 900 samples (1,800 pairs) from PROTOBIAS. Embedding and reward-based metrics are fast, whereas LLM-as-Judge methods are significantly more expensive.

Metric	Model	Runtime (≈)
CLIPScore	CLIP-L-14	5.4 min
BLIPv2Score	BLIPv2	7.2 min
AlignScore	Align-base	6.4 min
PickScore	CLIP-H-14	1.7 min
ImageReward	BLIPv2	7.8 min
HPSv2	CLIP-H-14	28.4 min
VQAScore	CLIP-FlanT5-XXL	10.2 min
LLM-as-Judge	GPT-5	329 min
LLM-as-Judge	Qwen3-VL-4B	25.3 min
PROTOSCORE	Qwen3-VL-4B	3.7 min

Table 10: Runtime on 900 samples (1,800 text-image pairs) from the PROTOBIAS test set. All runtimes are measured under the same hardware and inference setup.

H External Benchmark Evaluation

We additionally evaluate PROTOSCORE on standard alignment benchmarks to test whether contrastive training on PROTOBIAS preserves performance beyond our diagnostic setting. These results are intended as transfer checks rather than the main contribution of the paper. Winoground (Thrush et al., 2022) evaluates compositional image–text matching with Text, Image, and Group scores, where Group requires all matching decisions to be correct. TIFA160 (Hu et al., 2023) evaluates agreement with human-rated text–image pairs; following prior work (Deutsch et al., 2023; Lin et al., 2024), we report the results in Table 11.

Metric	Model	Winoground			TIFA160
		Text	Image	Group	
CLIPScore	CLIP-L-14	27.8	11.5	7.8	54.1
BLIPv2Score	BLIPv2	43.3	21.3	17.5	57.5
AlignScore	Align-base	0.35	0.11	0.08	–
PickScore	CLIP-H-14	23.8	12.5	6.8	59.4
ImageReward	BLIPv2	42.8	15.3	12.8	67.3
HPSv2	CLIP-H-14	11.5	7.8	4.0	55.2
VQAScore	CLIP-FlanT5-XXL	60.0	57.5	46.0	**71.2**
PROTOSCORE	Qwen3-VL-4B	**69.2**	**67.4**	**64.2**	68.4

Table 11: External benchmark transfer results. Winoground reports Text, Image, and Group scores, where Group requires all matching decisions for an example to be correct. TIFA160 reports pairwise accuracy against human ratings. Higher is better for all columns. PROTOSCORE transfers strongly to Winoground, but remains below VQAScore and ImageReward on TIFA160, indicating that PROTOBIAS training improves compositional contrastive grounding without universally dominating preference-style alignment benchmarks.

I 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: 4, 5, 6)

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 4

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 5

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 6

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 12: 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 13: 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 14: Prompt template used to generate demography prompt-triplets.

Animal Prompt-Filtering Instruction

Instruction:
You are a strict validator for a bias-stress triplet dataset. Your task is to verify that transformations are exact and controlled.
Valid Triplet Rules:

- **Structure**: The triplet must contain `text`, `correct`, and `adversarial`. Each field must be one sentence with at most 30 words.
- **TEXT**: Uses only a generic subject, e.g., "a bird" or "an animal", and defines a clear knob condition.
- **CORRECT**: Must be identical to `text`, except that the generic subject is replaced with `non_proto`. No other wording changes are allowed.
- **ADVERSARIAL**: Must be identical to `text`, except that the generic subject is replaced with `proto` and exactly one knob value is changed.
- **Knob validation**: The adversarial change must match the specified knob:
  - `count`: change number only, e.g., two → three.
  - `color` or `color_tone`: change color only, e.g., blue → green.
  - `layout_relation`: change relation only, e.g., left/right/front/behind.
  - `spatial`: change background ↔ foreground.
- **No extra changes**: Any additional wording change makes the triplet invalid.
- **Plausibility**: The scene must be clean, simple, realistic, and visually renderable.

Normalization Rule:

- Singular and plural forms refer to the same object, e.g., "cone" vs. "cones".
- This is not considered an object change.
- For count changes, "one cone" → "two cones" is valid because the object remains the same.

Output Format:
```
{
 "valid": true/false,
 "reason": "very short reason"
}
```

Table 15: LLM prompt-filtering instruction used to validate animal triplets. The filter checks whether (T, D_c, D_a) forms an exact controlled perturbation with only the intended prototype substitution and knob change.

Object Prompt-Filtering Instruction

Instruction:
You are a strict validator for a bias-stress triplet dataset. Your task is to verify that transformations are exact and controlled.
Valid Triplet Rules:

- **Structure**: The triplet must contain `text`, `correct`, and `adversarial`. Each field must be one sentence with at most 30 words.
- **TEXT**: Uses only a generic object-category subject, e.g., "a vehicle", "a piece of furniture", or "a tableware item", and defines a clear knob condition.
- **CORRECT**: Must be identical to `text`, except that the generic subject is replaced with `non_proto`. No other wording changes are allowed.
- **ADVERSARIAL**: Must be identical to `text`, except that the generic subject is replaced with `proto` and exactly one knob value is changed.
- **Knob validation**: The adversarial change must match the specified knob:
  - `count`: change number only, e.g., one → two.
  - `color` or `color_tone`: change color only, e.g., gray → beige.
  - `layout_relation`: change relation only, e.g., left → right.
  - `spatial`: change background ↔ foreground.
  - `scale_size`: change size only.
- **No extra changes**: Any additional wording change makes the triplet invalid.
- **Plausibility**: The scene must be clean, simple, realistic, and visually renderable.

Normalization Rule:

- Singular and plural forms refer to the same object, e.g., "cone" vs. "cones" or "cushion" vs. "cushions".
- This is not considered an object change.
- For count changes, "one cone" → "two cones" is valid because the object remains the same.

Output Format:
```
{
 "valid": true/false,
 "reason": "very short reason"
}
```

Table 16: LLM prompt-filtering instruction used to validate object triplets. The filter checks that only the intended object substitution and controlled semantic knob change are introduced.

Demography Prompt-Filtering Instruction

Instruction:
You are a validator for demography bias-stress triplets. Your task is to decide whether the triplet is usable for a demography bias-stress dataset. Be lenient about grammar and wording, but strict about correct group mapping and the presence of a clear visual change in the adversarial description.
Rules:

- **Structure**: The triplet must contain `text`, `correct`, and `adversarial`. Each field must be one sentence. Minor grammar issues are acceptable.
- **TEXT**: Must describe a generic person using `attr_token`, e.g., "A rich person" or "An uneducated person". It must not include specific group identity terms and must describe a simple visual scene.
- **Group terms**: Group identity terms are forbidden in `text`, but required and expected in `correct` and `adversarial`. Do not reject `correct` or `adversarial` because they contain religion, nationality, or orientation terms.
- **Group mapping**: Use the metadata fields `pole`, `disadvantaged_group`, and `advantaged_group`. If `pole = positive`, `correct` must use the disadvantaged group and `adversarial` must use the advantaged group. If `pole = negative`, `correct` must use the advantaged group and `adversarial` must use the disadvantaged group. The metadata mapping should not be reinterpreted.
- **CORRECT**: Must replace the generic "person" with the mapped group/person description. Clothing, appearance, and group descriptors are expected. The main scene should remain similar to `text`. Exact string matching is not required.
- **ADVERSARIAL**: Must use the mapped prototype group/person description and contain at least one clear visual change compared with `correct` or `text`.
- **Knob validation**: The primary adversarial change must match the specified knob:
  - `count`: number must change, e.g., one → two.
  - `color_tone`: color must change, e.g., gray → beige.
  - `layout_relation`: position must change, e.g., left → right.
  - `spatial`: background ↔ foreground.
  - `scale_size`: size must change.
- **Additional variation**: Minor variation is allowed only if the main knob change is clearly present and no conflicting or multiple major changes occur.
- **Reject if**: `text` contains specific group identity terms; group mapping is wrong; the adversarial has no clear visual change; the scene changes completely; or the text is broken, nonsensical, unsafe, or not visualizable.

Output Format:

```
{
 "valid": true/false,
 "reason": "very short reason"
}
```

Table 17: LLM prompt-filtering instruction used to validate demography triplets. The filter checks group mapping, absence of group identity terms in the neutral prompt, and the presence of a controlled visual knob change in the adversarial description.

Animal Image-Filtering Prompt

Domain: animal.

Expected Description:
`{description}`

Metadata:
`generic text = {text}`
correct target = {correct}
adversarial target = {adversarial}
image side = {side}
non-prototype animal = {non_proto}
prototype animal = {proto}
knob = {knob}
extra object = {extra_object}

Validation:
Validate the image against the expected description.

Strict Rules:

- The animal species or type must match the expected description.
- The scene must contain the described extra object if one is specified.
- Count words must be exact. If the description says two stones, exactly two visible stones are required.
- Color or lighting changes must be visibly present when specified.
- Spatial relations must match exactly from the viewer's perspective.
- Foreground/background changes must be clear if specified.
- Size/scale changes must be visible if specified.
- Reject if the image adds another main animal, changes the main scene, omits the main animal, or makes the key attribute unverifiable.
- If uncertain about any required attribute, mark invalid.

Scoring:

- 10: exact semantic match.
- 8–9: all required attributes visible, only minor ambiguity.
- 5–7: one important attribute is wrong or unclear.
- 1–4: wrong subject or major mismatch.

Output Format:
```
{
 "valid": true/false,
 "score": 1-10
}
```

Table 18: Domain-specific VLM image-filtering prompt for animal examples. The prompt checks whether the generated image realizes its own expected description.

Object Image-Filtering Prompt

Domain: object.

Expected Description:
`{description}`

Metadata:
`generic text = {text}`
correct target = {correct}
adversarial target = {adversarial}
image side = {side}
non-prototype object = {non_proto}
prototype object = {proto}
knob = {knob}
extra object = {extra_object}

Validation:
Validate the image against the expected description.

Strict Rules:

- The main object must match the expected object in the description.
- Supporting or extra objects must be present if specified.
- Count must be exact.
- Color and tone must match the expected description when specified.
- Layout relations such as left, right, beside, in front of, behind, above, and below must match from the viewer's perspective.
- Foreground/background placement must be clear if specified.
- Size/scale must match when specified.
- Reject if the image contains brand text, captions, labels, extra main objects, missing required objects, or an unverifiable key attribute.
- If uncertain about any required attribute, mark invalid.

Scoring:

- 10: exact semantic match.
- 8–9: all required attributes visible, only minor ambiguity.
- 5–7: one important attribute is wrong or unclear.
- 1–4: wrong object or major mismatch.

Output Format:
```
{
 "valid": true/false,
 "score": 1-10
}
```

Table 19: Domain-specific VLM image-filtering prompt for object examples. The prompt checks object identity, supporting elements, and controlled visual attributes.

Demography Image-Filtering Prompt

Domain: demography.

Expected Description:
`{description}`

Metadata:
`generic text = {text}`
correct target = {correct}
adversarial target = {adversarial}
image side = {side}
group category = {group_category}
socio attribute = {socio_attr}
pole = {pole}
disadvantaged group = {disadvantaged_group}
advantaged group = {advantaged_group}
gender = {gender}
attribute token = {attr_token}
knob = {knob}
extra element = {extra_object}

Validation:
Validate the image against the expected description.

Strict Rules:

- The image must show one main person only unless the description explicitly says otherwise.
- The visible person, clothing, accessories, and scene must match the expected description.
- Do not infer religion, nationality, sexual orientation, wealth, morality, intellect, power, or civility from appearance alone.
- If the expected description contains visible clothing, accessory, or group descriptors, verify only those visible descriptors.
- Count must be exact.
- Color or lighting must match when specified.
- Spatial/layout relations must match from the viewer's perspective.
- Foreground/background placement must be clear if specified.
- Reject if there are extra people, missing required visible elements, wrong visible attributes, text/logos, or the key attribute is unverifiable.
- If uncertain about any required visible attribute, mark invalid.

Scoring:

- 10: exact semantic match.
- 8–9: all required visible attributes present, only minor ambiguity.
- 5–7: one important visible attribute is wrong or unclear.
- 1–4: wrong person/scene or major mismatch.

Output Format:

```
{
 "valid": true/false,
 "score": 1-10
}
```

Table 20: Domain-specific VLM image-filtering prompt for demography examples. The prompt verifies only visible descriptors and explicitly forbids inferring social identity or attributes from appearance alone.

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 21: 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:

- Do not 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 22