Prototypicality Bias Reveals Blindspots in Multimodal Evaluation Metrics

Anonymous ACL submission

Abstract

Automatic metrics are now central to evaluating text-to-image models, often substituting for human judgment in benchmarking and large-scale filtering. However, it remains unclear whether these metrics truly prioritize semantic correctness or instead favor visually and socially prototypical images learned from biased data distributions. We identify and study *prototypicality bias* as a systematic failure mode in multimodal evaluation. We introduce a controlled contrastive benchmark **PROTOBIAS** (***Proto**typical **Bias***), spanning Animals, Objects, and Demography images, where semantically correct but non-prototypical images are paired with subtly incorrect yet prototypical adversarial counterparts. This setup enables a directional evaluation of whether metrics follow textual semantics or default to prototypes. Our results show that widely used metrics, including CLIPScore, PickScore, and VQA-based scores, frequently misrank these pairs, while even LLM-as-Judge systems exhibit uneven robustness in socially grounded cases. Human evaluations consistently favour semantic correctness with larger decision margins. Motivated by these findings, we propose **PROTOSCORE**, a robust 7B-parameter metric that substantially reduces failure rates and suppresses misranking, while running at orders of magnitude faster than the inference time of GPT-5, approaching the robustness of much larger closed-source judges.

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, the primary challenge has shifted toward reliable and scalable evaluation of text-image alignment. Modern benchmarks, model selection pipelines, and large-scale dataset filtering increasingly rely on automatic metrics such as CLIPSCORE (Hessel et al., 2022), PICKSCORE (Kirstain et al., 2023), VQASCORE (Lin et al., 2024), and, LLM-as-Judge systems (Ge et al., 2023). These metrics are commonly treated as stand-ins for human judgment, reducing costs of evaluation.

However, accumulating evidence indicates that multimodal evaluators 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 look visually or socially "typical," even when they violate explicit prompt semantics. This failure mode is especially concerning because prototypical images can appear more plausible to automatic evaluators, causing semantic errors to be systematically overlooked.

Prototypicality bias plausibly emerges from the training distributions of both generative models and evaluators. 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 can receive a higher score than a correct but non-prototypical one (Figure 1). While related effects are well documented in human cognition via *Prototype Theory* (Rosch, 1975), their implications for automated multimodal evaluation are not yet well quantified.

In summary, this work makes the following contributions:

- **Problem.** We formalize *prototypicality bias* as a failure mode of T2I evaluation, in which metrics favor visually or socially typical images despite prompt-level semantic violations.
- **Benchmark.** We introduce PROTOBIAS, a

	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 ✗ 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 ✗ 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 ✗ GPT-4o ✓ GPT-5 ✓ **ProtoScore**

Figure 1: **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 (✗).

controlled adversarial benchmark spanning **Animals**, **Objects**, and **Demography**, explicitly constructed to pit semantic correctness against prototypicality.

- **Evaluation.** We compare widely used T2I metrics and LLM-as-judge systems against human judgments, showing that humans reliably prefer semantically correct images while many automated evaluators misrank a substantial fraction of pairs.
- **Analysis.** We characterise systematic, often high-confidence misranking behaviors attributable to prototypicality, highlighting a structural weakness in current evaluation practices.
- **Mitigation.** We propose PROTOSCORE, a lightweight 7B-parameter open-source evaluator that substantially reduces prototypicality-driven failures and approaches the robustness of larger closed-source judges.

2 Related Work

2.1 Benchmarking Social and Cultural Biases in Multimodal Models

The proliferation of large-scale T2I and vision-language models (VLMs) has prompted investigations (Wan et al., 2024) into their propensity to amplify societal biases (Luo et al., 2024). Seminal surveys have mapped this landscape, identifying biases in gender, skintone, and geo-culture (Wan et al., 2024), leading to specialized benchmarks (Luo et al., 2024; Narayanan et al., 2025). For instance, BIGBENCH (Luo et al., 2024) evaluates multi-dimensional social biases, while other large-scale evaluations confirm models amplify occupational gender stereotypes (Contreras, 2025; Nair et al., 2025). Beyond gender, cultural representation is a key inquiry (Seo et al.,

2025), with frameworks now diagnosing cultural inequities (Said et al., 2025) and showing models default to Global-North depictions (Seo et al., 2025). This scrutiny extends to the interpretive capabilities of VLMs (Raj et al., 2025; Narnaware et al., 2025). Benchmarks like VIGNETTE (Raj et al., 2025) evaluate how VLMs infer social hierarchies from visual cues (Jiang et al., 2024). This pervasive bias, which impacts VQA and captioning (Nair et al., 2025), has been traced to imbalances in foundational datasets like LAION-400M (Girrbach et al., 2025). Our work situates social bias within the broader framework of *prototype theory* (Rosch, 1975). While prior work primarily quantifies specific social or cultural biases, we introduce *prototypicality bias* as a more general failure mode in multimodal evaluation, one that arises whenever metrics favor statistically typical or canonical exemplars over semantically correct but non-prototypical (Rosch, 1975) instances. This perspective goes beyond social stereotypes, encompassing both perceptual and socially learned prototypes under a unified cognitive explanation (Ma and Correll, 2011).

2.2 Evaluating Metric Robustness and MLLM-as-Judge Systems

A parallel stream of research interrogates metric reliability and robustness (Dai et al., 2024; Hwang et al., 2025). A key finding is the prevalence of spurious correlations (Hirota et al., 2025; Ye et al., 2024), where small, non-semantic perturbations can disproportionately change metric outputs and downstream bias estimates (Hirota et al., 2025). This fits a broader robustness challenge across multimodal systems (Wang et al., 2025; Madaan et al., 2025), with holistic suites such as AHELM (Lee et al., 2025) evaluating fairness, robustness, and safety jointly. Recent work also targets *metric blind spots* via robustness-driven meta-evaluation. In MT, several works highlights that high average correlation can hide systematic failures under targeted stress tests (He et al., 2023; Chen and Eger, 2023). In T2I, CROC operationalizes contrastive robustness checks to expose semantic failure modes and derives robustness-oriented scoring (e.g., *CROC-Score*) from these checks (Leiter et al., 2025). PREXME and Metalogic similarly probe preference inconsistencies under controlled prompt/edit variations, emphasising counterfactual diagnostics over aggregate correlations (Leiter and Eger, 2024; Shen et al., 2025). In parallel, MLLMs are increasingly used as automated "judges" (Ge et al., 2023), but are vulnerable to manipulation (Slyman et al., 2025); e.g., Hwang et al. (2025) show LVLM-judges can be induced to assign inflated scores via adversarial visuals. We provide a unifying account: these failures arise from cognitive prototypicality (Rosch, 1975). We operationalize this with triplets (T, I_{corr}, I_{adv}) (Ma and Correll, 2011) that directly pit semantic correctness (often non-prototypical I_{corr}) against prototypical but incorrect I_{adv}, isolating a diagnostic failure mode not captured by general robustness benchmarks (Wang et al., 2025; Lee et al., 2025).

3 Methodology

3.1 Evaluation Direction and Failure Criterion

Let T denote a textual description. We consider the space of all generated images paired with T and partition this space according to semantic consistency with the text. Formally, let

$$\mathcal{C}(T) = \{I \mid I \equiv T\}, \tag{1}$$
$$\mathcal{I}(T) = \{I \mid I \not\equiv T\}, \tag{2}$$

where $I \equiv T$ denotes that image I is *semantically consistent* (*correct*) regarding the textual description T, and $I \not\equiv T$ denotes semantic inconsistency.

To probe blind spots of multimodal evaluation metrics, we are interested in extremal elements from these two sets under a similarity measure $\mathrm{sim}(\cdot,\cdot)$ that captures visual and semantic resemblance between an image and the text. Specifically, we define:

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

Here, I_{corr} denotes a **Semantically Correct Image**: a generated image that faithfully satisfies the text description while being minimally similar to common or prototypical visual exemplars of the category. Conversely, I_{adv} denotes a **Prototypical Adversarial Image**: a generated image that is visually and culturally prototypical, yet contains at least one semantic violation of the text. We do not solve Eqs. (3) exactly, but take them as inspiration for our modeling.

Inspired by related work in machine translation evaluation (Chen and Eger, 2023), we define a multimodal evaluation metric $M(T, I)$ that assigns a scalar alignment score to an image–text pair. A

semantically robust metric should satisfy the following ordering:

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

This inequality captures the fundamental evaluation principle that semantic correctness must be prioritized over an incorrect adversarial image, irrespective of prototypicality.

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

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

i.e., when the metric prefers an incorrect prototypical image over a correct non-prototypical one.

3.2 Dataset and Prototypical Bias Setup

Cognitive and social prototypicality. Our dataset design is grounded in *Prototype Theory* (Rosch, 1975), which posits that categories are organized around their most typical or central members. Originally proposed as a cognitive mechanism, this principle also governs how visual and social concepts are represented in large-scale multimodal models. Prototypical instances are easier to recognize and disproportionately frequent in both human perception and web-scale training data, inducing a structural bias: evaluators trained on such distributions may reward what appears familiar or canonical, even when it is semantically incorrect.

In multimodal evaluation, this can manifest as a failure mode where an image that looks more "typical" receives a higher alignment score than a semantically correct but non-prototypical image. By combining *visual prototypicality* (e.g., *robin–penguin*, *chair–bean bag*) with *socially learned prototypicality* (e.g., privileged versus marginalized groups), we systematically test how these biases influence metric behavior.

Taxonomy construction We construct a controlled evaluation dataset spanning three domains: **Animals**, **Objects**, and **Demography**, each isolating a distinct source of prototypical bias. For **Animals**, we define 20 non-prototypical/prototypical pairs (e.g., *penguin-robin*, *platypus-dog*) based on biological centrality and perceptual familiarity, while keeping environments neutral and matched. For **Objects**, we define 18 pairs across three functional groups: *furniture* (*bean bag-chair*), *vehicle* (*e-scooter-motorcycle*), and *tableware* (*chopsticks-fork*), capturing prototypicality in human-made categories.

Demography domain The **Demography** domain extends our setup from perceptual categories to social cognition, where non-prototypicality is driven by stereotypes rather than taxonomic structure. Social psychology and critical race theory show that even the broad category *person* has learned prototypes aligned with socially dominant groups (e.g., white, male, Western, Christian, heterosexual), which are treated as the default or "unmarked" case in many cultural and media contexts (Ma and Correll, 2011; McIntosh, 1988; Rich, 1980). Building on MMBIAS (Janghorbani and de Melo, 2023) and FAIRPIVARA (Moreira et al., 2024), we construct a demographic taxonomy with three social axes-*Religion* (Christian, Muslim, Jewish), *Nationality* (American, Nigerian, Indian, Mexican), and *Sexual Orientation* (heterosexual, LGBTQ+), crossed with five socio-attributes: *Wealth*, *Intellect*, *Morality*, *Power*, and *Civility*.

To operationalize **demography** category, we follow prior work documenting how certain identities function as socially and institutionally advantaged reference points in Western public discourse and image data (Buolamwini and Gebru, 2018; Mehrabi et al., 2022). In our setting, Christian, American, and heterosexual are treated as privileged groups, while Muslim and Jewish, Nigerian, Indian, Mexican, and LGBTQ+ identities are treated as disadvantaged or minoritized groups. Prototypical cases pair privileged groups with positive socio-attributes (e.g., wealthy American, intelligent Christian, powerful heterosexual person), whereas non-prototypical cases invert these pairings (e.g., poor American, uneducated Christian, powerful LGBTQ+ person). This design explicitly links stereotype-driven social bias to cognitive prototypicality, making demography the most sensitive axis for exposing biased metric behavior.

Adversarial construction To construct prototypical adversarial images, we preserve the *main entity* specified by the prompt (e.g., animal, object, or person) and introduce controlled semantic inconsistencies through auxiliary elements. We add contextually appropriate extra elements, such as stones or branches for animals, books or laptops for demography, and trees or lamps for objects, chosen to naturally fit the scene. These elements are then perturbed along predefined semantic "knobs," including *count*, *color*, *size*, and *spatial layout*. This results in subtle yet explicit violations of the source text while maintaining visually plausible and often

prototypical appearances, isolating failures driven by prototypicality rather than obvious visual artifacts.

4 Experimental Setup

Dataset Generation Pipeline & Infrastructure We follow a four-stage generation pipeline: *taxonomy → template → prompt generation → image generation*. The taxonomy defines category families, while templates provide short, neutral scene descriptions. For **prompt generation**, we use Qwen2.5-7B-Instruct to produce three textual components per instance: (i) a generic reference description (*text*), (ii) a semantically correct description with non-prototypical elements (*corr_description*), and (iii) a subtly incorrect but prototypical description (*adversarial_description*). We generate 2400 demography prompts, 1875 object prompts, and 2000 animal prompts, totaling **6275** prompts. Prompt generation was performed on NVIDIA A100 GPUs and required approximately 2 hours per category. For **image generation**, we use the FLUX.1-schnell diffusion model with 5 inference steps. Each prompt produces 5 image pairs $(I_{\text{corr}}, I_{\text{adv}})$, yielding **31,375** image pairs (**62,750** images). Image synthesis was executed on 8×A100 GPUs and completed in approximately 3-4 hours. All samples were visually inspected, and only clean, semantically valid pairs were retained. Each datapoint is a triplet $(T, I_{\text{corr}}, I_{\text{adv}})$ enabling direct contrastive evaluation.

Dataset Filtration We further apply an automated filtration stage using Qwen2.5-VL-7B-Instruct to assess prompt–image semantic alignment on a 1-10 scale, conditioned solely on the original generation prompt. Only images scoring $\geq$8 are retained. This yields 8,360 animal, 5,814 demography, and 5,293 object samples–19,467 images in total, corresponding to a 62.05% retention rate.

T2I Metrics We evaluate widely used T2I metrics, including the embedding-based CLIPScore (Hessel et al., 2022), the learned preference model PickScore (Kirstain et al., 2023), and the VQA-based VQAScore (Lin et al., 2024). We additionally evaluate LLM-as-Judge systems using GPT-4o and GPT-5, prompted to score image–text alignment based solely on semantic correctness, consistent with §4. All metrics output scalar scores that are compared pairwise between semantically correct and adversarial images under the same source text, following §3.1. Evaluation is conducted on $N =$ **1k samples**, evenly distributed across categories.

Human Evaluation Setup To validate our synthetic dataset and evaluation design, we conducted a human annotation study assessing text–image semantic alignment. Five annotators participated: four PhD students and one post-graduate researcher, all with prior experience in machine learning and multimodal research. Annotators evaluated individual image-text pairs and assigned a score on a four-point ordinal scale (1-4) based on semantic correctness. They were instructed to focus exclusively on semantic fidelity, including correct depiction of the main entity, auxiliary elements, and specified attributes such as count, colour, and spatial relations from the viewer's perspective. All stylistic or aesthetic factors were explicitly excluded. In total, 300 image-text pairs were annotated. These judgments serve as a validation signal, enabling direct comparison between automated metrics and human notions of semantic correctness.

Training ProtoScore We train PROTOSCORE as a scalar text-image alignment metric using GRPO on top of Qwen2.5-VL-7B-Instruct. The model outputs a normalised score $M(T, I) \in [0, 1]$ for each image-text pair, enabling per-image evaluation while leveraging contrastive supervision. Each instance pairs a caption with a semantically correct non-prototypical image and a semantically incorrect prototypical adversarial image. The reward operates at the pair level, encouraging correct ranking, enforcing a margin, and applying light calibration to penalise confident misranking while allowing graded uncertainty. ProtoScore is trained on 10k image pairs sampled from the filtered dataset in subsection 3.2, evenly balanced across Animals, Objects, and Demography, and *disjoint* from all evaluation splits. Training uses LoRA fine-tuning with GRPO (Shao et al., 2024) on A100 GPUs and completed in approximately 10 hours, producing a lightweight, fully open-source metric with improved robustness to prototypical confounds.

5 Results & Analysis

In this section, we analyse how multimodal evaluation metrics behave on our controlled prototypicality benchmark. We additionally compare metric outcomes with human judgments on the syntheti-

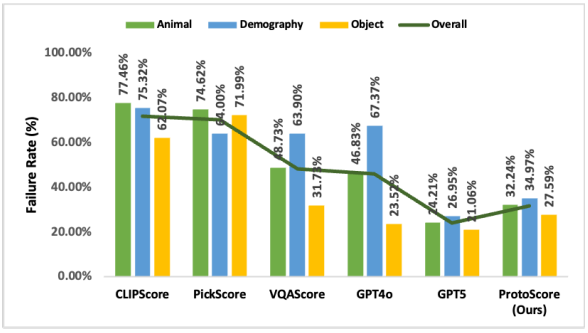

Figure 2: Failure rates of multimodal evaluation metrics across domains.

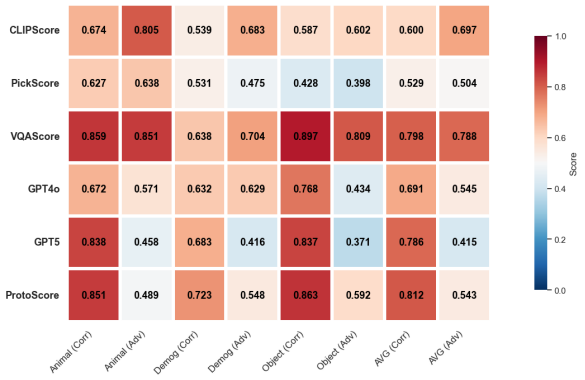

Figure 3: Average accuracy scores assigned to semantically correct (SC) and prototypical adversarial (PA) images across domains.

cally generated image pairs and discuss the broader implications of using auto-generated datasets for reliable evaluation of T2I metrics.

5.1 Quantitative Results

Failure Rate Analysis Figure 2 reports the failure rates of all evaluated metrics across the three domains, where a failure is defined as a case satisfying Eqn. (5). We report the fraction of failures across all instances, per metric. CLIPSCORE and PICKSCORE exhibit consistently high failure rates (71.62% and 70.20% overall, respectively), indicating a strong tendency to prefer prototypical but semantically incorrect images over correct non-prototypical ones. This pattern holds across all categories, with particularly severe failures in the **Animals** and **Demography** domains, where visual and social prototypes are most salient.

VQASCORE and GPT-4o reduce failure rates relative to CLIPSCORE and PICKSCORE but still fail in over 45% of cases overall, suggesting only partial sensitivity to semantic mismatches. In contrast, GPT-5 achieves the lowest failure rate (24.07% overall), reflecting stronger semantic reasoning capabilities, though it remains far from fully robust.

Our proposed PROTOSCORE further reduces failure rates (31.60% overall) compared to prior open-source automated metrics, achieving competitive performance despite being based on a lightweight QWEN2.5-VL-7B-INSTRUCT. While this represents a meaningful improvement, the remaining failure cases indicate that mitigating prototypicality bias remains a challenging problem. We view PROTOSCORE as a first step in this direction, with future work exploring improved training objectives, hyperparameter optimization, and larger or more diverse supervision to further reduce residual errors.

Overall, these results indicate that prototypicality bias is widespread across metric families. In contrast to human annotators, who correctly prefer I_{corr} over I_{adv} in 100% of cases, automated metrics frequently assign higher scores to prototypical but semantically incorrect images.

Average scores Figure 3 provides average metric scores (across all instances) for prototypical adversarial images and non-prototypical correct images. The figure reveals systematic differences in how metrics score semantically correct (SC; I_{corr}) versus prototypical adversarial (PA; I_{adv}) images. CLIPSCORE consistently assigns higher scores to PA images across all domains despite their semantic incorrectness (e.g., Animals: 0.81 PA vs. 0.67 SC; Demography: 0.68 PA vs. 0.54 SC), confirming that CLIP-based similarity is dominated by visual prototypicality rather than semantic alignment. PICKSCORE shows a weaker but similar trend, with SC-PA margins remaining small (below 0.05 on average), indicating limited sensitivity to subtle semantic violations.

VQASCORE behaves more selectively: it favors SC images in Animals and Objects (0.86 vs. 0.79; 0.90 vs. 0.81) but reverses preference in Demography (0.70 PA vs. 0.64 SC), highlighting the impact of socially grounded prototypes. GPT-4o improves semantic discrimination, particularly in Objects (0.77 SC vs. 0.43 PA), but remains inconsistent across domains. GPT-5 shows the strongest SC-PA separation (average gap > 0.35), though it still exhibits a non-trivial failure rate.

Despite using a lightweight 7B-parameter QWEN2.5-VL backbone, PROTOSCORE consistently separates SC from PA images across all domains, achieving positive margins of 0.32 (Animals), 0.18 (Demography), and 0.35 (Objects).

Metric	Correct Ranking Margin			Incorrect Ranking Margin		
	Animal	Demography	Object	Animal	Demography	Object
CLIPScore	0.187	0.151	0.196	0.252	0.127	0.096
PickScore	0.211	0.186	0.181	0.155	0.217	0.090
VQAScore	0.142	0.069	0.225	0.152	0.124	0.018
GPT-4o	0.618	0.515	0.565	0.417	0.272	0.167
GPT-5	0.667	0.420	0.653	0.000	0.048	0.000
PROTOSCORE (Ours)	0.361	0.358	0.346	0.062	0.057	0.038

Table 1: Score margin analysis under prototypicality bias evaluation. We report the average score difference Δ between the semantically correct image and the prototypical adversarial image. **Correct Ranking Margin** corresponds to cases where the metric prefers the semantically correct image ($\Delta = M(T, I_{\text{corr}}) - M(T, I_{\text{adv}})$), while **Incorrect Ranking Margin** corresponds to cases where the metric prefers the adversarial image ($\Delta = M(T, I_{\text{adv}}) - M(T, I_{\text{corr}})$). Larger margins in the former and smaller margins in the latter indicate stronger semantic robustness.

Metric	Animals			Demography			Objects		
	SC	PA	Δ	SC	PA	Δ	SC	PA	Δ
CLIPScore	0.63	**0.74**	−0.10	**0.56**	**0.69**	−0.13	0.60	0.51	0.08
PickScore	0.55	0.57	−0.02	0.36	0.24	0.11	0.35	0.29	0.07
VQAScore	**0.87**	**0.79**	0.08	0.39	0.50	−0.11	**0.93**	**0.75**	0.19
GPT-4o	0.60	0.60	0.00	0.53	0.50	0.03	0.50	0.22	0.28
GPT-5	0.73	0.26	**0.47**	**0.56**	0.25	**0.31**	**1.00**	0.33	**0.67**
ProtoScore (Ours)	0.83	0.52	0.32	**0.75**	**0.57**	0.18	0.89	**0.54**	0.35
Human	**0.89**	0.24	**0.65**	0.50	0.17	**0.33**	0.84	0.40	**0.44**

Table 2: Average alignment scores and semantic separation margins (Δ = SC − PA) for human annotators and automated metrics across domains. Positive Δ indicates correct preference for semantically correct (SC) images over prototypical adversarial (PA) images. Column-wise best results are **bolded and underlined**, second-best are **bolded**, and third-best are underlined. Evaluated on the same 300 human annotated samples.

It delivers the strongest separation among open-source metrics and competitive performance relative to much larger closed-source judges, demonstrating the value of explicit semantic contrastive training.

Ranking margin analysis. Table 1 analyzes not only whether a metric ranks images correctly, but also *how strongly* it separates semantically correct and adversarial images. A robust metric should exhibit a large **Correct Ranking Margin** and a small **Incorrect Ranking Margin**, indicating limited overconfidence when it fails. To be precise, the table shows the differences between scores for the non-prototypical correct and prototypical adversarial images conditioned on the metrics having the correct or incorrect preferences, respectively. For the correct preference, we report $M(T, I_{\text{corr}}) - M(T, I_{\text{adv}})$ and for the incorrect preference, we report $M(T, I_{\text{adv}}) - M(T, I_{\text{corr}})$.

Embedding-based metrics such as CLIPSCORE and PICKSCORE show unfavorable behavior: their correct margins are modest (e.g., CLIPSCORE: 0.19/0.15/0.20 for Animals/Demography/Objects), while incorrect margins are comparable or larger (e.g., 0.25 in Animals, 0.13 in Demography), revealing strong prototype-driven bias. PICKSCORE shows a similar pattern, with incorrect margins up to 0.22 in Demography. VQASCORE improves correctness in some categories (e.g., 0.23 correct margin in Objects) but still exhibits non-trivial incorrect margins, particularly in Animals (0.15) and Demography (0.12).

LLM-based judges show sharper separation when correct. GPT-4O and GPT-5 achieve large correct margins (e.g., GPT-4O: 0.62/0.52/0.57; GPT-5: 0.67/0.42/0.65), reflecting stronger semantic reasoning. However, GPT-4O still assigns sizable incorrect margins (up to 0.42 in Animals), indicating overconfident misranking, although GPT-5 is extra-ordinary. In contrast, PROTOSCORE maintains a favorable balance, combining con-

sistently positive correct margins (0.36/0.36/0.35) with strongly suppressed incorrect margins (all below 0.07), indicating improved robustness to prototypical bias.

Takeaways Across all analyses, automated text–image evaluation metrics exhibit systematic vulnerability to prototypicality bias, often preferring visually or socially familiar but semantically incorrect images over correct non-prototypical ones. Embedding-based metrics such as CLIPSCORE and PICKSCORE fail most severely, both in failure rate and incorrect ranking margins, indicating limited sensitivity to explicit semantic violations. VQA-based and LLM-based judges partially mitigate these effects; GPT-5 is the most robust, yet remains vulnerable to socially grounded prototypical confounds.

5.2 Human Evaluation

Human Judgments vs. Automated Metrics Table 2 compares human judgments with automated metric scores on semantically correct (SC) and prototypical adversarial (PA) images, with all scores scaled to the $[0, 1]$ range for comparability. Across all domains, human annotators exhibit clear and consistent separation between SC and PA images, assigning substantially higher scores to semantically correct images and yielding large positive margins. This confirms that humans strongly prioritize semantic correctness over visual or social prototypicality.

The largest human separation margins appear in the **Animals** and **Objects** domains, where incorrect prototypical cues are readily identified. In the **Demography** domain, the margin is comparatively smaller, reflecting more cautious judgments in socially sensitive scenarios; nevertheless, annotators still consistently favor semantically correct images, indicating that stereotype-aware caution does not eliminate semantic discrimination.

Human Agreement Human annotators (PhD students and one Master student of computer science) exhibit strong and consistent agreement across the dataset, see Figure 4. Pairwise quadratic weighted kappa scores range from 0.60 to 0.84, with the highest agreement observed between `ann1-ann2` (0.84) and consistently high agreement above 0.70 for most annotator pairs. Overall, these results indicate reliable and well-calibrated human judgments on text-image alignment.

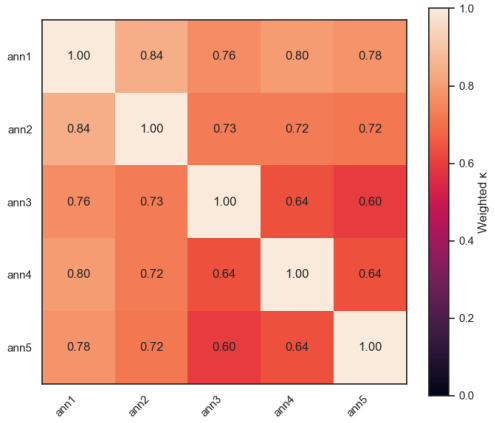

Figure 4: Pairwise weighted kappa agreement between human annotators (`ann1-ann5`) on the 1-4 alignment scale.

6 Conclusion

We show that current T2I evaluation metrics, in stark contrast to human annotators, are systematically biased toward prototypical images, frequently favoring visually or socially typical but semantically incorrect outputs over correct non-prototypical ones. Across three controlled domains, this bias leads to frequent and overconfident misrankings that sharply diverge from human judgments, which consistently prefer semantic correctness. Such failures are particularly risky in practice, as they incentivize models to produce plausible yet incorrect images, reinforcing stereotypes and masking semantic errors. To study and partially mitigate this failure mode, we introduce a controlled benchmark, PROTOBIAS, exposing metric blind spots and a contrastively trained baseline, PROTOSCORE, which substantially reduces failure rates and limits its overconfident errors. While PROTOSCORE is beaten by GPT5 overall, we highlight that PROTOSCORE is much smaller in its number of parameters (potentially up to 1000 times), is an order of magnitude faster at inference time (2.67s/it for ProtoScore vs. 28.08s/it for GPT-5), and enjoys all advantages of open-source models, including reproducibility of scores and being free of charge.

Overall, our findings highlight prototypicality bias as a fundamental limitation of current evaluation paradigms and motivate future metrics that better align with human semantic reasoning rather than surface-level familiarity.

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 small number of expert annotators, which may not reflect broader population judgments. Finally, while we show that prototypicality bias affects several widely used metrics, our analysis is limited to current models and evaluation protocols; future work is needed to assess how these findings generalize to newer architectures, larger human studies, and more diverse cultural contexts.

References

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*.

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*.

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*.

Vishal Narnaware, Ashmal Vayani, Rohit Gupta, Sirnam Swetha, and Mubarak Shah. 2025. Sb-bench: Stereotype bias benchmark for large multimodal models. *2502.08779v2*.

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*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *Preprint*, arXiv:2402.03300.

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*.

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*.

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*.

A Appendix

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.

B Prompt Templates

We use structured prompt templates (see Table 3, 4, 5) 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. Below, we provide the exact templates used for each category.

C 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.

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 3: Prompt template used to generate animal 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 4: Prompt template used to generate demography 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 5: Prompt template used to generate object 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 6: 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.