PROBING THE GENERATIVE WORLDVIEW: UNCOVERING IMPLICIT VISUAL PRIORS IN T2I MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Text-to-image (T2I) models must resolve a fundamental challenge: natural language is radically underspecified compared to the visual world. A prompt like "a man waiting on the side of the road" only conveys a handful of visual attributes, yet the T2I model must commit to many specific choices across different visual dimensions such as weather, lighting, and clothing to depict the scene. To characterize how models fill in what is left unsaid, we introduce *implicit visual priors*, the systematic tendencies T2I models exhibit over unspecified visual attributes, and argue that previously studied biases are a subset of this broader concept. Consequently, we propose ProbeGen, a bottom-up framework that discovers these priors directly from generated images without pre-specified classes or hypotheses. Unlike top-down approaches that anticipate potential biases, our observation-driven method surfaces priors across diverse visual dimensions. Through extensive experiments, we find that implicit visual priors are highly pervasive, with models imposing multiple simultaneous defaults and resolving the same underspecification in measurably distinct ways.

1 INTRODUCTION

What does a text-to-image model like to predict when you leave something unsaid? Consider the prompt "a man waiting on the side of the road". It specifies a subject, an action, and a location, but says nothing about the weather, the man's clothing, or the color of the sky. Yet the model must make concrete choices for every one of these dimensions to create an image. When we generate multiple images from this prompt, a pattern emerges: the outputs are far more homogeneous than the space of plausible images would suggest, defaulting to similar weather, camera angles, and color palettes, despite the prompt specifying none of them (Figure 1). These are not random draws from a broad distribution; they are systematic tendencies that reveal what the model has learned to treat as typical.

We term these systematic tendencies *implicit visual priors*, the learned assumptions that T2I models impose on visual attributes not specified by the input prompt. Modern diffusion models have rapidly evolved from unconditional synthesis (Ho et al., 2020) to highly complex text-conditioned architectures (Rombach et al., 2022; Ramesh et al., 2022; Saharia et al., 2022). However, even with improved conditioning, these models map sparse textual constraints to high-dimensional image manifolds, forcing the model to commit to visual choices wherever the prompt is underspecified. Therefore, every generated image is the product of what the user said and what the model assumed. An implicit visual prior characterizes the latter. While fundamental to model behavior, these priors remain largely understudied, making their characterization essential for comprehensive auditing, informed model selection, and the development of more robust T2I models.

Prior work has studied a narrow slice of this phenomenon, primarily focusing on social biases using closed-set classifiers (Cho et al., 2023; Bianchi et al., 2023; Naik & Nushi, 2023; Vice et al., 2025b) or, more recently, top-down open-set discovery where LLMs hypothesize potential biases (D'Incà et al., 2024). While these approaches have advanced our understanding of harmful model behaviors, they share fundamental limitations that prevent a comprehensive understanding of implicit visual priors. First, because they rely on hypothesis-driven discovery, the set of discoverable priors is bounded by what an LLM thinks to propose. Subtle compositional tendencies are therefore unlikely to surface as they are difficult to anticipate in advance. Second, even when they discover novel biases, existing methods (D'Incà et al., 2024) measure them by forcing attributes into discrete,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

predefined classes (e.g., selecting from a list of laptop brands). This categorical measurement fails to capture the free-form, continuous nature of many visual properties. Third, the focus on social attributes, while important, has left the broader phenomenon uncharacterized.

To address these limitations, we propose ProbeGen, a bottom-up, observation-driven approach to study implicit visual priors. Rather than hypothesizing what priors might exist, we start from the generated images, extracting the visual attributes that are actually present, aggregating them across multiple generations, and analyzing the resulting distributions. This reversal, from images to patterns rather than hypotheses to verification, enables the discovery of priors that no language model would anticipate, captures continuous and compositional visual properties, and evaluates different visual dimensions without prioritizing predefined ones. Our key contributions are:

- We introduce the concept of *implicit visual priors* and propose ProbeGen, a bottom-up framework that uncovers these priors from generated images without pre-specified hypotheses or classes.
- We design complementary distributional analyses, spanning from prompt-level frequency and co-occurrence profiles to global prior prevalence and cross-model comparisons, comprehensively characterizing T2I model behavior.
- We demonstrate the actionability of the uncovered prevalent priors, successfully leveraging them to enhance visual diversity as one of the possible downstream application scenarios.

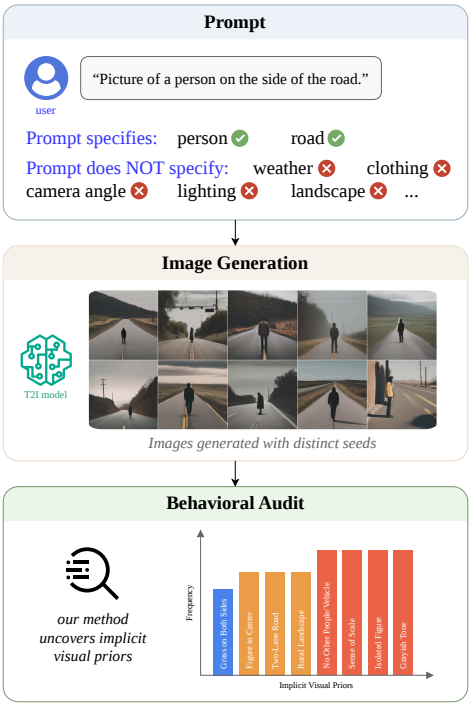

Figure 1: Our method reveals implicit visual priors in T2I models. Despite varying random seeds, the model systematically defaults to the same scene, tone, and composition.

2 Related Work

Evaluating Bias in Text-to-Image Models. Research on bias in T2I models has progressed from closed-set probing of predefined attributes to increasingly open-ended analysis. Early work employed diagnostic prompts and processed generated images through automated classifiers to quantify gender and skin tone distributions (Cho et al., 2023; Bianchi et al., 2023; Naik & Nushi, 2023; Vice et al., 2025a;b). For instance, Bianchi et al. (2023) showed that T2I models amplify demographic stereotypes at scale, while Naik & Nushi (2023) documented that seemingly neutral prompts favor Western norms. Wang et al. (2023) adapted the Implicit Association Test from social psychology to quantify implicit biases using CLIP similarity scores. However, these approaches rely on predefined attribute sets and external classifiers such as FairFace (Karkkainen & Joo, 2021). As Lyu et al. (2025) and Seshadri et al. (2024) demonstrated, such detectors can introduce measurement bias, especially when applied to ambiguous or low-quality generations. The T2ISafety benchmark (Li et al., 2025) measures fairness via KL divergence against uniform reference distributions but remains limited to closed attribute sets, and while Vice et al. (2025b) propose metrics to quantify T2I bias, their analysis also relies on closed-set probing. Wan et al. (2024) provide a comprehensive survey highlighting the diversity and fragmentation of these existing approaches.

Open-Set and Intersectional Bias Discovery. Recognizing that important biases may lie outside predefined sets, recent work has moved toward open-set discovery. OpenBias (D'Incà et al., 2024) uses an LLM to propose potential biases given textual captions, then verifies their presence via a VQA model, discovering novel biases such as laptop brands and train colors. GradBias (D'Inca et al., 2025) extended this by analyzing the influence of individual words on bias manifestation. TIBET (Chinchure et al., 2024) uses an LLM to dynamically generate bias axes and creates counterfactual generations for comparison, while OASIS (Dehdashtian et al., 2025) proposed a sociologically grounded definition of stereotypes, distinguishing them from generic bias by requiring a

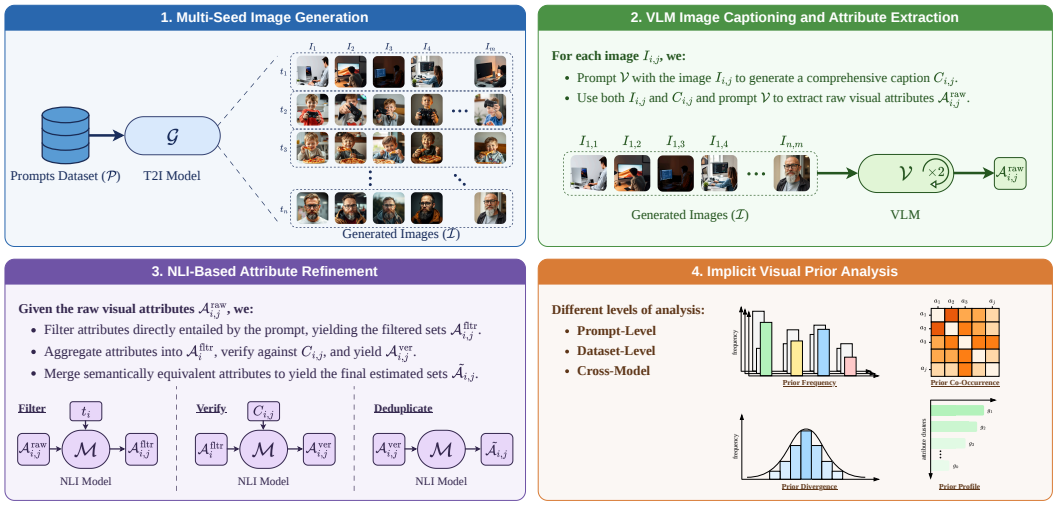

Figure 2: Overview of ProbeGen. A T2I model generates multiple images per prompt, from which a VLM extracts visual attributes. These attributes are refined via a multi-stage NLI process and analyzed to uncover implicit visual priors.

directional violation of real-world distributions. More recently, BiasConnect (Shukla et al., 2025) quantifies how intervening on one bias dimension causally affects another, Porikli & Porikli (2025) document the persistence of stereotypes across model generations, and Vice et al. (2026) use input perturbations to identify bias.

While these methods have expanded the scope of bias detection, they share a common paradigm: top-down, hypothesis-driven discovery. An LLM anticipates what biases might exist, and these hypotheses are then tested in generated images. This is inherently bounded by what the LLM proposes and what can be expressed as categorical class sets. Our work inverts this paradigm. Instead of asking "what biases might exist?" and verifying, we ask "what did the model actually produce?" and find the patterns. This bottom-up approach discovers priors that no LLM would anticipate.

3 Method

We propose ProbeGen, a bottom-up framework for discovering implicit visual priors in T2I models. Given a set of prompts, our pipeline (depicted in Figure 2) extracts visual attributes from generated images, refines them, and analyzes their statistical distributions. In the following subsections, we first define the concept of an implicit visual prior, then detail our extraction and refinement process, and finally introduce our metrics for different types of analysis.

3.1 Implicit Visual Prior Extraction

Image Generation. Every text prompt inherently leaves out the majority of visual dimensions unspecified, i.e., a prompt determines a subject, a setting, perhaps an action, but says nothing about many of the other visual attributes a T2I model must resolve to generate an image. We define an *implicit visual prior* as any visual attribute present in a generated image that is not semantically entailed by the input prompt.

Given a set of n textual prompts $\mathcal{P} = \{t_1, \ldots, t_n\}$, we generate m images per prompt using the target T2I model $\mathcal{G}$ by varying the random seed $s_j \in \{s_1, \ldots, s_m\}$, yielding $I_{i,j} = \mathcal{G}(t_i, s_j)$. Let $\mathcal{A}_{i,j}$ denote the set of all visual attributes present in a given image $I_{i,j}$. We define an attribute $a \in \mathcal{A}_{i,j}$ as an implicit visual prior if:

$$a \notin \text{Entailed}(t_i), \tag{1}$$

where $\text{Entailed}(t_i)$ denotes the set of attributes semantically entailed by t_i. In practice, obtaining the set of all visual attributes and the prompt-entailed attributes is not feasible. We therefore estimate $\text{Entailed}(t_i)$ using a natural language inference (NLI) model, and approximate $\mathcal{A}_{i,j}$ via a VLM, yielding the estimated set $\tilde{\mathcal{A}}_{i,j}$, as detailed below.

Visual Attribute Extraction. We employ a VLM $\mathcal{V}$ to extract attributes from each generated image using a two-stage querying strategy. First, we prompt $\mathcal{V}$ with the image $I_{i,j}$ and a structured instruction to generate a comprehensive caption $C_{i,j}$. Second, using both the image and its caption $C_{i,j}$ as context, we instruct $\mathcal{V}$ to extract a list of visual attributes, yielding a raw attribute set $\mathcal{A}_{i,j}^{\text{raw}}$. This two-stage approach ensures the extracted attributes are both semantically rich and visually grounded. Full prompt templates are provided in Appendix A1.

Attribute Refinement and Verification. The raw attribute sets may contain noise, restate information already specified in the prompt, or omit relevant visual details. We refine these into the estimated sets $\tilde{\mathcal{A}}_{i,j}$ via a five-stage process (filtering, candidate expansion, cross-image verification, bidirectional deduplication, and guarded hypernym propagation) using a pre-trained NLI model $\mathcal{M}$.

First, the filtering stage isolates implicit visual priors by removing attributes directly entailed by the prompt. For each attribute $a \in \mathcal{A}_{i,j}^{\text{raw}}$, we discard a if $\mathcal{M}(t_i, a)$ predicts `entailment`. Second, the candidate expansion stage breaks down complex attributes into more granular components. Using part-of-speech tagging, we extract valid noun-phrase skip-bigrams from the raw attributes (e.g., deriving "wooden table" from "old wooden dining table"). We retain a candidate only if it is entailed by its source attribute and not explicitly implied by the prompt. This yields a denser, more granular filtered set $\mathcal{A}_{i,j}^{\text{fltr}}$.

Next, the verification stage addresses the fact that a single VLM pass may miss relevant attributes. We aggregate all unique attributes across the m images into a prompt-level pool $\mathcal{A}_i^{\text{fltr}} = \bigcup_{j=1}^{m} \mathcal{A}_{i,j}^{\text{fltr}}$. We then verify each pooled attribute against every image's caption $C_{i,j}$. If $\mathcal{M}(C_{i,j}, a)$ predicts `entailment`, we add a to a new verified set $\mathcal{A}_{i,j}^{\text{ver}}$. This recovers attributes that the VLM described in the caption but omitted from the explicit attribute list.

Finally, we consolidate semantically equivalent and hierarchical attributes over $\mathcal{A}_i^{\text{ver}}$. The deduplication stage merges synonymous phrasings (e.g., "cloudy sky" and "overcast weather") via bidirectional entailment: $a_x \equiv a_y \iff \mathcal{M}(a_x, a_y) = \texttt{ent.} \wedge \mathcal{M}(a_y, a_x) = \texttt{ent.}$

We construct an equivalence graph and merge connected components, selecting the shortest phrasing as the representative. Subsequently, the hypernym propagation stage leverages unidirectional entailment to enforce hierarchical consistency (e.g., ensuring "jacket" is present if "blue jacket" is). To avoid unnecessary NLI computations, we only test pairs where the lemmas of one attribute form a subset of the other. We then propagate these entailment chains so that higher-level concepts are added to the image sets. The resulting deduplicated and propagated prompt-level pool is denoted $\tilde{\mathcal{A}}_i$, with the final per-image estimated sets denoted $\tilde{\mathcal{A}}_{i,j}$.

3.2 Implicit Visual Prior Analysis

Prompt-Level Analysis. The prevalence of an implicit visual prior is measured by its frequency $f(a \mid t_i)$ across the m generated images. A prior appearing in every image reflects a strong model tendency, while one appearing rarely indicates a weak one. For each attribute $a \in \tilde{\mathcal{A}}_i$, we compute its frequency as:

$$f(a \mid t_i) = \frac{1}{m} \sum_{j=1}^{m} \mathbf{1}\left[a \in \tilde{\mathcal{A}}_{i,j}\right]. \tag{2}$$

Ranking attributes by their frequency yields the prompt-level prior profile. To quantify the prevalence of the model's most frequent prior for a given prompt, we compute the mode concentration (MC):

$$\text{MC}(t_i) = \max_{a \in \tilde{\mathcal{A}}_i} f(a \mid t_i), \tag{3}$$

which captures the frequency of this top prior. A model that consistently injects the same visual attribute for at least one concept will have $\text{MC}(t_i)$ close to 1. To discover structured associations between attributes, we compute their Lift, which measures the ratio of their observed co-occurrence to their expected co-occurrence under independence:

$$\text{Lift}(a_x, a_y \mid t_i) = \frac{f(a_x, a_y \mid t_i)}{f(a_x \mid t_i) \cdot f(a_y \mid t_i)}, \tag{4}$$

where $f(a_x, a_y \mid t_i)$ denotes the co-occurrence frequency of attributes a_x and a_y across the m generated images. A Lift value of 1.0 indicates independence. Values significantly greater than 1.0

suggest a strong positive association, while values less than 1.0 indicate a negative association. To prevent spurious associations from rare pairs, we restrict our analysis to attributes appearing in at least a τ_{freq} fraction of the images, and only consider pairs with a co-occurrence frequency of at least τ_{co}. Finally, to avoid capturing trivial or near-identical pairs, we exclude those with an embedding cosine similarity exceeding τ_{triv}.

Dataset-Level Analysis. Comparing priors across prompts requires a shared attribute vocabulary. We pool all attributes into a global set $\tilde{\mathcal{A}}^{\text{global}} = \bigcup_i \tilde{\mathcal{A}}_i$, embed each using a sentence embedding model, and apply greedy centroid clustering with a cosine similarity threshold τ_{cluster}. This produces clusters $\{g_1, \ldots, g_K\}$, where we select the member closest to the centroid as the representative label.

For each cluster g_k, we compute its prevalence as the fraction of prompts where it appears as a strong prior:

$$\text{Prev}(g_k) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}[f(g_k \mid t_i) \geq \theta]\,. \tag{5}$$

Here, $f(g_k \mid t_i)$ is the fraction of the m generated images that contain at least one attribute belonging to cluster g_k. Additionally, to summarize how constrained a model's generations are for an individual prompt, we count its number of strong priors:

$$N_{\text{strong}}(t_i) = \left| \left\{ a \in \tilde{\mathcal{A}}_i : f(a \mid t_i) \geq \theta \right\} \right|. \tag{6}$$

The distribution of N_{strong} across prompts indicates how pervasive strong priors are across the dataset.

Cross-Model Comparison. We compare models using the shared global vocabulary. For each prompt t_i, we construct an attribute frequency distribution $P_{\mathcal{G}}^{(i)}$ over the global clusters for a given model $\mathcal{G}$. We measure the distributional divergence between two models, $\mathcal{G}_1$ and $\mathcal{G}_2$, via the Jensen-Shannon Divergence (JSD):

$$\text{JSD}\left(P_{\mathcal{G}_1}^{(i)} \| P_{\mathcal{G}_2}^{(i)}\right) = \frac{1}{2} D_{\text{KL}}\left(P_{\mathcal{G}_1}^{(i)} \| M^{(i)}\right) + \frac{1}{2} D_{\text{KL}}\left(P_{\mathcal{G}_2}^{(i)} \| M^{(i)}\right), \tag{7}$$

where $M^{(i)} = \frac{1}{2}(P_{\mathcal{G}_1}^{(i)} + P_{\mathcal{G}_2}^{(i)})$ is the mixture distribution. Averaging the JSD across all prompts yields an overall prior divergence score between two models.

While JSD captures distributional differences, it does not reveal whether two models default to the exact same attributes. To measure this, we compare the content of each model's strongest priors. Let $\mathcal{S}_k(\mathcal{G}, t_i)$ denote the set of top-k highest-frequency attributes from model $\mathcal{G}$ for prompt t_i. The content overlap between two models is measured via Jaccard similarity:

$$J_k(t_i) = \frac{|\mathcal{S}_k(\mathcal{G}_1, t_i) \cap \mathcal{S}_k(\mathcal{G}_2, t_i)|}{|\mathcal{S}_k(\mathcal{G}_1, t_i) \cup \mathcal{S}_k(\mathcal{G}_2, t_i)|}. \tag{8}$$

Together, JSD quantifies how differently models distribute their attribute frequencies, while J_k quantifies whether they default to the same visual priors.

4 Experiments

4.1 Experimental Setup

Models. We evaluate two text-to-image models with different architectures: Stable Diffusion XL (SDXL) (Podell et al., 2023), a widely deployed latent diffusion model, and FLUX.1 [schnell] (Black Forest Labs, 2024), a recent diffusion transformer designed for fast inference. For attribute extraction, we employ Gemma 4 12B (Team, 2026) as our VLM with greedy decoding for deterministic, reproducible outputs. For attribute refinement, we use a DeBERTa-v3 model fine-tuned for natural language inference (NLI) [1] (He et al., 2021; Sileo, 2024) with an entailment threshold of 0.80 for filtering, verification, and bidirectional deduplication. Finally, for global attribute clustering, we use the all-MiniLM-L6-v2 sentence embedding model (Reimers & Gurevych, 2019; Wang et al., 2020) alongside our greedy centroid clustering algorithm with a distance threshold of τ_{cluster}=0.3.

[1] `sileod/deberta-v3-base-tasksource-nli`

Prompt Datasets. To ensure our findings generalize across different visual domains, we curate 1,000 prompts from MS-COCO (Lin et al., 2014), sampling 200 prompts from each of five distinct supercategories: People, Vehicles, Animals, Food, and Furniture. For the non-human domains, we explicitly exclude images containing people to isolate domain-specific priors. To ensure semantic diversity within each domain, we apply K-Means clustering ($K = 200$) on the text embeddings of the filtered captions, selecting the caption closest to each cluster center.

Image Generation. For each of the 1,000 prompts, we generate m=20 images per model at 1024×1024 resolution using fixed seeds shared across models to enable fair cross-model comparison. We use the default inference parameters recommended by the model authors for all experiments, i.e., SDXL runs the Euler Discrete sampler (30 steps, CFG (Ho & Salimans, 2022)=7.5), and FLUX.1 [schnell] runs the Flow Matching Euler Discrete sampler (4 steps). This process yields $20{,}000$ images per model, providing a robust visual corpus for our analysis.

Analysis Setup. We set the strong prior frequency threshold to $\theta = 0.8$, meaning an attribute must appear in at least 80% of the generated images to be considered a strong prior. For co-occurrence analysis, we set the minimum attribute frequency to $\tau_{\text{freq}} = 0.4$ and the minimum co-occurrence frequency to $\tau_{\text{co}} = 0.2$. The trivial pair similarity threshold for filtering near-identical attributes is set to $\tau_{\text{triv}} = 0.75$. Additional implementation details are provided in Appendxi A1.

4.2 Prompt-Level Analysis

Prior Profiles Reveal Pervasive Defaults. Figure 3 shows the prior profiles for several representative prompts across SDXL and FLUX.1. For each prompt, we display the top attributes ranked by frequency across the m=20 generated images. The most immediate observation is that models exhibit strong priors across multiple visual properties simultaneously. For example, for the prompt "a little girl that is eating some food," SDXL predominantly defaults to indoor settings, while depicting the subject seated, having dark hair, and east Asian racial features, despite none of these attributes being specified in the prompt. These priors span environmental settings, demographic characteristics, and compositional conventions.

Co-occurrence Analysis Reveals Visual Templates. Beyond individual attribute frequencies, our method reveals structured associations between attributes, visualized in Figure 3 using two complementary heatmaps. The raw co-occurrence matrix reveals dominant visual templates (i.e., specific combinations of attributes that frequently co-occur) by counting the absolute number of images sharing two attributes, while the Lift matrix quantifies their statistical association to expose the model's conditional dependencies and distinct visual attribute subgroups.

For instance, the Lift heatmap reveals that for the prompt "a man that is playing a wii game," the model strongly associates a framed wall picture and a bookshelf with books (Lift = 1.5), implicitly defaulting to a standard living room layout.

4.3 Aggregate and Cross-Model Analysis

Aggregate Prior Statistics. To characterize the overall prevalence of implicit visual priors, we compute the mode concentration (MC) and the count of strong priors (N_{strong}) across all prompts for each model. Table 1 summarizes the mean and median values across our five distinct prompt domains, as well as the overall aggregate.

The results reveal that T2I models exhibit a remarkably high degree of constraint. Across almost all domains, both models display median MC scores of 1.0, meaning that for the vast majority of prompts, at least one visual attribute appears in all of the generated images. Furthermore, the N_{strong} metric indicates that models do not just constrain a single visual dimension, but rather impose 5-7 simultaneous strong priors per prompt on average.

Comparing models, the newer FLUX.1 model consistently exhibits higher MC and N_{strong} numbers than SDXL across all domains, suggesting that models capable of higher image fidelity may actually enforce a narrower set of implicit priors. This effect is most pronounced in the People and Vehicles domains, where FLUX.1 imposes significantly more simultaneous priors than SDXL. Full histograms detailing the distributions of these metrics across prompts are provided in Appendix A2.1.

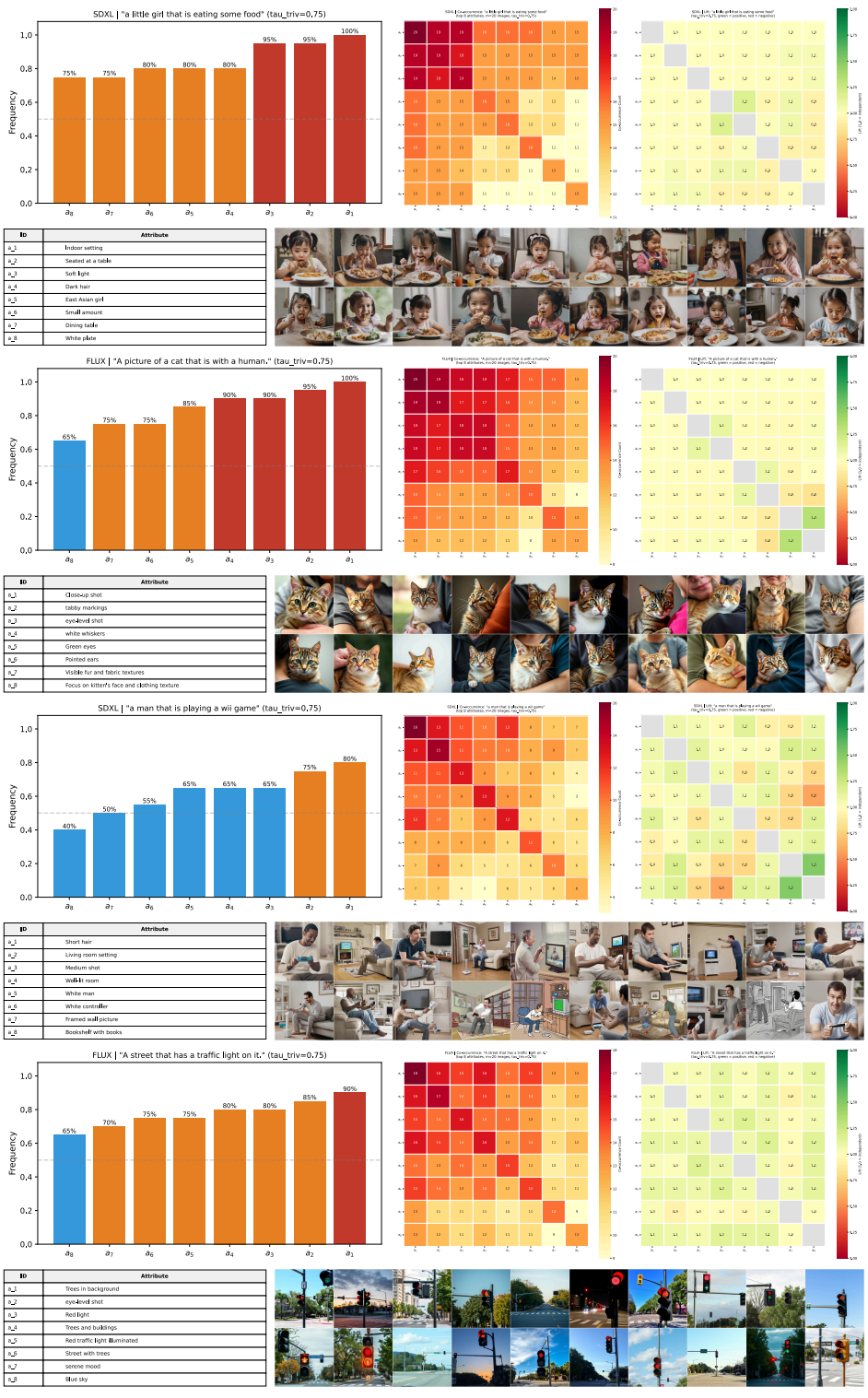

Figure 3: Implicit visual priors for four MS-COCO prompts across SDXL and FLUX.1. *Prior frequency profile* (Left): frequency of unprompted attributes across m=20 images; *Co-occurrence heatmap* (Middle): dominant visual templates; *Lift heatmap* (Right): statistical associations (green: attracting; red: repelling). Generated images (bottom) confirm the homogeneity found by our pipeline.

Table 1: Aggregate prior statistics across five distinct prompt domains and overall. Higher MC and N_{strong} values indicate a more constrained generation process.

	MC_{mean}		$\text{MC}_{\text{median}}$		$N_{\text{strong, mean}}$		$N_{\text{strong, median}}$	
Domain	**SDXL**	**FLUX.1**	**SDXL**	**FLUX.1**	**SDXL**	**FLUX.1**	**SDXL**	**FLUX.1**
Animals	0.966	0.971	1.000	1.000	7.3	7.7	6	7
Food	0.971	0.968	1.000	1.000	6.0	6.0	6	5
Furniture	0.946	0.958	0.950	1.000	4.6	5.0	4	5
People	0.945	0.963	1.000	1.000	4.7	5.7	4	5
Vehicles	0.952	0.959	1.000	1.000	5.2	6.3	5	5
All	0.956	0.964	1.000	1.000	5.6	6.1	5	5

Table 2: Top 3 most prevalent implicit visual priors for each prompt domain. Prevalence is measured as the fraction of prompts where the prior appears in $\geq 80\%$ of the generated images.

Domain	Model	Top 1	Top 2	Top 3
Animals	SDXL	Soft lighting (31.5%)	Fur texture (23.5%)	Light green grass (21.5%)
	FLUX.1	Fur texture (35.0%)	Medium-shot, eye-level photograph (25.0%)	Facing left, head tilted toward camera (21.5%)
Food	SDXL	Close-up, high-angle shot (83.0%)	Bright lighting (32.0%)	Simple white plate (22.5%)
	FLUX.1	High-angle, eye-level shot (89.5%)	Bright lighting (40.0%)	Simple white plate (29.0%)
Furniture	SDXL	High-angle, eye-level shot (74.0%)	Bright lighting (17.0%)	Simple white plate (13.0%)
	FLUX.1	Close-up, high-angle shot (44.5%)	Bright lighting (22.0%)	Eye-level shot (19.0%)
People	SDXL	Eye-level shot (20.5%)	Full-length shot (14.0%)	Bright, overcast sky (12.0%)
	FLUX.1	Eye-level shot (29.5%)	Medium shot (21.0%)	Pale blue sky with light clouds (14.5%)
Vehicles	SDXL	Eye-level shot (43.5%)	Bright, overcast sky (29.0%)	High-angle shot (20.5%)
	FLUX.1	Eye-level shot (48.5%)	Bright, cloudy sky (33.0%)	High-angle shot (24.5%)
All	SDXL	High-angle, wide-angle shot (26.9%)	Eye-level shot (22.5%)	Soft lighting (11.1%)
	FLUX.1	Eye-level shot (27.3%)	Close-up, high-angle shot (21.0%)	Bright lighting (17.6%)

Prior Prevalence Across Prompts. Using the global attribute clusters g_k, we identify the most prevalent implicit visual priors across the entire dataset to reveal a model's global prior profile. Table 2 shows the top 3 most prevalent attribute clusters for SDXL and FLUX.1 across our five prompt domains, ranked by the fraction of prompts for which the cluster constitutes a strong prior (i.e., $f \geq 0.8$). Full top-10 lists for all domains are provided in Appendix A2.2.

The results reveal pervasive compositional and environmental assumptions in both models. Compositional priors dominate the Food and Furniture domains, where both models almost deterministically generate high-angle shots for up to 89.5% of prompts. Conversely, for People and Vehicles, both models heavily favor eye-level shots. Cross-domain priors are also prevalent, i.e., "Bright lighting" and "Soft lighting" are consistently ranked in the top three across Animals, Food, and Furniture domains. Demographically, when human subjects are left underspecified, both models exhibit strong priors for light skin tones and dark hair (see Appendix A2.2). Environmental priors are equally rigid, with both models consistently placing vehicles against cloudy or overcast skies (up to 33.0%) and animals on grassy fields. These prevalent priors act as the foundational visual templates the models rely on when they lack explicit instruction.

Cross-Model Comparison. We compare SDXL and FLUX.1 using prior divergence (JSD) and overlap (J_k) metrics, as detailed in Section 3. Results are summarized across the five prompt domains in Table 3, and the full distributions across all prompts are visualized in Appendix A2.3.

The results demonstrate substantial divergence between the two models across all domains. Overall, the models exhibit a mean JSD of 0.49, indicating that they resolve visual underspecification in different ways. This divergence is also reflected in the low content overlap of their strongest priors. The overall median J_5 score is 0.25, meaning that for the median prompt, SDXL and FLUX.1 share only two attributes among their five strongest implicit priors.

This divergence is not uniform across domains. People and Vehicles exhibit the highest divergence (mean JSD up to 0.50) and the lowest overlap ($J_5 \approx 0.20$). The median J_5 overlap drops to 0.111 for these domains, meaning SDXL and FLUX.1 frequently share only a single attribute among their top five priors. This exposes significant differences in how the models depict human subjects. Conversely, Food and Furniture show the lowest divergence (mean JSD ≈ 0.45) and highest overlap ($J_5 \approx 0.27$). This suggests that while these models disagree on demographic and vehicle priors,

Table 3: Cross-model comparison across five domains, where higher JSD indicates greater divergence and higher J_5 indicates greater overlap.

Domain	JSD_{mean}	JSD_{med}	$J_{5,\text{mean}}$	$J_{5,\text{med}}$
Animals	0.436 ± 0.088	0.421	0.213 ± 0.140	0.250
Food	0.463 ± 0.087	0.456	0.277 ± 0.181	0.250
Furniture	0.435 ± 0.079	0.424	0.269 ± 0.178	0.250
People	0.488 ± 0.088	0.486	0.193 ± 0.145	0.111
Vehicles	0.503 ± 0.072	0.504	0.205 ± 0.153	0.111
All	0.488 ± 0.085	0.483	0.214 ± 0.164	0.250

Table 4: Increasing visual diversity of the Food domain images by using the top 3 priors as negative prompts, measured via DINOv2 distance.

Metric	Baseline	Diversified
Mean Diversity	0.2452	0.2725
Rel. Improvement	–	+11.12%
t-stat	–	10.68
p-value	–	< 0.0001
Prompts (N)	200	

they have learned similar compositional templates for inanimate objects, leaning toward the same high-angle shots and backgrounds.

4.4 Applications of ProbeGen: Enhancing Visual Diversity

To demonstrate that the insights extracted by ProbeGen are actionable for downstream tasks, we conduct a simple experiment to improve visual diversity. We focus on the Food domain for SDXL. As shown in Table 2, the top 3 most prevalent priors for this domain are a "Close-up, high-angle shot" (83.0%), "Bright lighting" (32.0%), and a "Simple white plate" (22.5%). To enhance the overall visual diversity, we apply these priors as negative prompts to 50% of the generations. We then measure the pairwise cosine distance of DINOv2 (Oquab et al., 2023) embeddings to quantify semantic diversity.

As shown in Table 4, avoiding these prevalent priors successfully increases the semantic diversity of the outputs. The mean pairwise cosine distance increased from 0.2452 to 0.2725, representing a statistically significant ($p < 0.0001$), relative improvement of 11.12%. While we demonstrate this utility through a simple negative prompting experiment, the results validate that the priors discovered by our pipeline are highly actionable. For future work, we expect that integrating these insights into more sophisticated generation methods will enable a broader range of applications.



Figure 4: *Top*: Baseline defaults to simple white plates. *Bottom*: Using top-3 priors as negative prompts breaks this template, yielding diverse plate colors and shapes.

5 Conclusion

We proposed ProbeGen, a bottom-up framework that extracts, refines, and analyzes implicit visual priors directly from generated images. Across five distinct visual domains, our experiments revealed that T2I models frequently gravitate toward rigid visual patterns, imposing multiple simultaneous defaults. Co-occurrence analysis further demonstrated that these priors form coherent visual templates rather than occurring independently. Cross-model comparisons highlighted substantial divergence in how different architectures resolve visual underspecification. Crucially, we demonstrated that these extracted insights are highly actionable, using the top discovered priors to successfully increase the visual diversity of the outputs by 11.12%, establishing ProbeGen as a robust tool for auditing and improving T2I models.

Limitations

Our method's perceptiveness is bounded by the VLM used for attribute extraction, i.e., the attributes we can discover are limited to what this model can visually perceive and articulate. Furthermore, the NLI-based attribute refinement and verification, while robust, rely on the NLI model's semantic understanding. However, the modular design of our pipeline allows these components to be seamlessly upgraded as more capable foundation models become available.

AI USE STATEMENT

In this work, generative AI tools were used to assist in writing code for our experiments and editing the manuscript to improve readability. All AI-assisted code was manually inspected and verified for correctness. The authors reviewed all AI-assisted text and take full responsibility for the final content of this work.

REFERENCES

Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan. Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. In *Proceedings of the 2023 ACM conference on fairness, accountability, and transparency*, pp. 1493–1504, 2023.

Black Forest Labs. Announcing Black Forest Labs. Web Page, August 2024. URL `https://bfl.ai/blog/24-08-01-bfl`.

Huanqia Cai, Sihan Cao, Ruoyi Du, Peng Gao, Steven Hoi, Zhaohui Hou, Shijie Huang, Dengyang Jiang, Xin Jin, Liangchen Li, et al. Z-image: An efficient image generation foundation model with single-stream diffusion transformer. *arXiv preprint arXiv:2511.22699*, 2025.

Aditya Chinchure, Pushkar Shukla, Gaurav Bhatt, Kiri Salij, Kartik Hosanagar, Leonid Sigal, and Matthew Turk. Tibet: Identifying and evaluating biases in text-to-image generative models. In *European Conference on Computer Vision*, pp. 429–446. Springer, 2024.

Jaemin Cho, Abhay Zala, and Mohit Bansal. Dall-eval: Probing the reasoning skills and social biases of text-to-image generation models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3043–3054, 2023.

Sepehr Dehdashtian, Gautam Sreekumar, and Vishnu Naresh Boddeti. Oasis uncovers: High-quality t2i models, same old stereotypes. *arXiv preprint arXiv:2501.00962*, 2025.

Moreno D'Incà, Elia Peruzzo, Massimiliano Mancini, Dejia Xu, Vidit Goel, Xingqian Xu, Zhangyang Wang, Humphrey Shi, and Nicu Sebe. Openbias: Open-set bias detection in text-to-image generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12225–12235, 2024.

Moreno D'Inca, Elia Peruzzo, Massimiliano Mancini, Xingqian Xu, Humphrey Shi, and Nicu Sebe. Gradbias: Unveiling word influence on bias in text-to-image generative models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*, 2021.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Dengyang Jiang, Dongyang Liu, Zanyi Wang, Qilong Wu, Liuzhuozheng Li, Hengzhuang Li, Xin Jin, David Liu, Zhen Li, Bo Zhang, et al. Distribution matching distillation meets reinforcement learning. *arXiv preprint arXiv:2511.13649*, 2025.

Kimmo Karkkainen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 1548–1558, 2021.

Lijun Li, Zhelun Shi, Xuhao Hu, Bowen Dong, Yiran Qin, Xihui Liu, Lu Sheng, and Jing Shao. T2isafety: Benchmark for assessing fairness, toxicity, and privacy in image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13381–13392, 2025.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Dongyang Liu, Peng Gao, David Liu, Ruoyi Du, Zhen Li, Qilong Wu, Xin Jin, Sihan Cao, Shifeng Zhang, Hongsheng Li, et al. Decoupled dmd: Cfg augmentation as the spear, distribution matching as the shield. *arXiv preprint arXiv:2511.22677*, 2025.

Yunbo Lyu, Zhou Yang, Yuqing Niu, Jing Jiang, and David Lo. Do existing testing tools really uncover gender bias in text-to-image models? In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 11687–11696, 2025.

Ranjita Naik and Besmira Nushi. Social biases through the text-to-image generation lens. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 786–808, 2023.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

Sedat Porikli and Vedat Porikli. Hidden bias in the machine: Stereotypes in text-to-image models. *arXiv preprint arXiv:2506.13780*, 2025.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 3982–3992, 2019.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

Preethi Seshadri, Sameer Singh, and Yanai Elazar. The bias amplification paradox in text-to-image generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6367–6384, 2024.

Pushkar Shukla, Aditya Chinchure, Emily Diana, Alexander Tolbert, Kartik Hosanagar, Vineeth N Balasubramanian, Leonid Sigal, and Matthew A Turk. Biasconnect: Investigating bias interactions in text-to-image models. *arXiv preprint arXiv:2503.09763*, 2025.

Damien Sileo. tasksource: A large collection of nlp tasks with a structured dataset preprocessing framework. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 15655–15684, 2024.

Gemma Team. Gemma 4 technical report, 2026. URL https://arxiv.org/abs/2607.02770.

Jordan Vice, Naveed Akhtar, Richard Hartley, and Ajmal Mian. Exploring bias in over 100 text-to-image generative models. *arXiv preprint arXiv:2503.08012*, 2025a.

Jordan Vice, Naveed Akhtar, Richard Hartley, and Ajmal Mian. Quantifying bias in text-to-image generative models. *IEEE Transactions on Dependable and Secure Computing*, 2025b.

Jordan Vice, Naveed Akhtar, Leonid Sigal, Richard Hartley, and Ajmal Mian. On the fairness, diversity and reliability of text-to-image generative models. *Artificial Intelligence Review*, 59(2): 57, 2026.

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

Jialu Wang, Xinyue Liu, Zonglin Di, Yang Liu, and Xin Wang. T2iat: Measuring valence and stereotypical biases in text-to-image generation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 2560–2574, 2023.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in neural information processing systems*, 33:5776–5788, 2020.

A1 Extended Implementation Details

A1.1 VLM Prompt Templates

As described in Section 3 of the main text, we employ a two-stage querying strategy using the Gemma 4 12B Vision-Language Model (Team, 2026) to extract rich, visually-grounded attributes. All VLM inferences are performed using greedy decoding for perfect reproducibility.

Stage A: Detailed Captioning. The image is passed to the VLM along with the following prompt:

```
You are an AI image analyst. Your task is to write a
detailed visual description for the image as a
paragraph.
Rules:
- Be specific but not verbose
- Avoid redundancy and don't repeat information
- Include visible racial cues in your people
descriptions
```

Stage B: Structured Feature Extraction. Using both the image and the generated caption as context, we extract the final raw attributes using:

```
Extract key visual features from this description as a
concise bulleted list. Remove filler words and keep
phrases short.
Description: {caption}
Format as:
- feature 1
- feature 2
etc.
```

A1.2 NLI Premise and Hypothesis Formulations

A core contribution of our pipeline is the zero-shot application of Natural Language Inference (NLI) to filter, verify, and deduplicate attributes. For all NLI tasks, we utilize the `sileod/deberta-v3-base-tasksource-nli` (He et al., 2021; Sileo, 2024) model. This model

takes two text strings, a Premise and a Hypothesis, and outputs probabilities for three classes: `entailment`, `neutral`, and `contradiction`. We define an entailment as successful if the `entailment` class probability exceeds 0.80 (unless otherwise specified). The exact formulation of the Premise and Hypothesis for each pipeline stage is detailed below.

Step 3a: Filtering Prompt-Entailed Attributes. To ensure we only measure implicit priors, we remove any attribute directly requested by the user.

- **Premise:** The original user prompt (e.g., "a man waiting on the side of the road").
- **Hypothesis:** The extracted visual attribute (e.g., "Two-lane road").

If entailment is predicted, the attribute is discarded as explicit.

Step 3b: Cross-Image Attribute Verification. To correct for VLM omission errors, we check if an attribute extracted from one image is actually present in the detailed caption of another.

- **Premise:** The detailed VLM caption ($C_{i,j}$) for the target image.
- **Hypothesis:** An attribute (a) discovered in the master pool for that prompt.

If entailment is predicted, the attribute is added to the target image's refined list.

Step 3c: Bidirectional Entailment Deduplication. To merge semantically equivalent phrasings, we perform a strict bidirectional check between every pair of attributes in the pool.

- **Forward Check:** Premise = Attribute A; Hypothesis = Attribute B.
- **Backward Check:** Premise = Attribute B; Hypothesis = Attribute A.

Only if *both* checks yield entailment probabilities > 0.80 are the two attributes merged. This strictly merges synonyms (e.g., "overcast sky" $\equiv$ "cloudy weather") while rejecting one-way hypernyms (e.g., "blue jacket" $\rightarrow$ "jacket").

A core contribution of our pipeline is the zero-shot application of Natural Language Inference (NLI) to filter, expand, verify, deduplicate, and propagate attributes. For all NLI tasks, we utilize the `sileod/deberta-v3-base-tasksource-nli` (He et al., 2021; Sileo, 2024) model. This model takes two text strings, a Premise and a Hypothesis, and outputs probabilities for three classes: `entailment`, `neutral`, and `contradiction`. We define an entailment as successful if the `entailment` class probability exceeds a specified threshold. The exact formulation of the Premise and Hypothesis for each pipeline stage is detailed below.

Step 3a: Literal Prompt Filtering. (Threshold: 0.80) To ensure we only measure implicit priors, we remove any attribute directly requested by the user.

- **Premise:** The original user prompt (e.g., "a man waiting on the side of the road").
- **Hypothesis:** The extracted visual attribute (e.g., "Two-lane road").

If entailment is predicted, the attribute is discarded as explicit.

Step 3b: Candidate Expansion. (Threshold: 0.80) To break down complex, verbose attributes into atomic noun-phrase components, we use part-of-speech tagging to extract valid skip-bigrams from the raw attributes. We retain a candidate only if it passes two gates: (1) the source attribute entails the candidate, and (2) the prompt does not entail the candidate.

- **Gate 1 Premise:** The source attribute a.
- **Gate 1 Hypothesis:** The extracted candidate n-gram c.
- **Gate 2 Premise:** The original user prompt t_i.
- **Gate 2 Hypothesis:** The extracted candidate n-gram c.

Step 3c: Cross-Image Attribute Verification. (Threshold: 0.80) To correct for VLM omission errors, we check if an attribute extracted from one image is actually present in the detailed caption of another.

- **Premise:** The detailed VLM caption ($C_{i,j}$) for the target image.
- **Hypothesis:** An attribute (a) discovered in the master pool for that prompt.

If entailment is predicted, the attribute is added to the target image's refined list.

Step 3d: Bidirectional Entailment Deduplication. (Threshold: 0.85) To merge semantically equivalent phrasings, we perform a strict bidirectional check between every pair of attributes in the pool. To optimize, the backward check is only performed if the forward check succeeds.

- **Forward Check:** Premise = Attribute A; Hypothesis = Attribute B.
- **Backward Check:** Premise = Attribute B; Hypothesis = Attribute A.

Only if *both* checks yield entailment probabilities > 0.85 are the two attributes merged. This strictly merges synonyms (e.g., "overcast sky" $\equiv$ "cloudy weather") while rejecting one-way hypernyms.

Step 3e: Guarded Hypernym Propagation. (Threshold: 0.85) To enforce hierarchical consistency (e.g., ensuring "jacket" is present if "blue jacket" is), we leverage unidirectional entailment. To avoid unnecessary NLI computations, we only test pairs where the lemmas of one attribute form a subset of the other.

- **Premise:** The specific attribute (e.g., "blue jacket").
- **Hypothesis:** The broader hypernym (e.g., "jacket").

If entailment is predicted, we then propagate these entailment chains so that higher-level concepts are added to the final sets accordingly.

FairFace Demographic Translation (Appendix A4). To map our open-ended textual extractions to FairFace's rigid taxonomy for benchmarking, we constructed explicit, descriptive hypotheses for each FairFace class.

- **Premise:** The concatenated list of all refined attributes for a given image.
- **Hypothesis (Gender):** We test two hypotheses: "The person is male" and "The person is female."
- **Hypothesis (Race):** We test seven hypotheses corresponding to FairFace categories (e.g., "The person appears to be Black," "The person appears to be East Asian," etc.).

For this specific translation task, we lower the entailment threshold to 0.50 and take the argmax of the resulting probabilities. If no hypothesis exceeds the threshold, the demographic attribute is marked as "Unknown."

A2 Additional Results

A2.1 Aggregate Metric Distributions

In Section 4.3 of the main text, we reported the mean and median Mode Concentration (MC) and Strong Prior Count (N_{strong}) across different prompt domains. Figure A1 provides the full histograms of these metrics for both SDXL and FLUX.1. The strong skew toward 1.0 in the MC distributions, and the bell curves centering around 5 to 7 in the N_{strong} distributions, visually confirm the pervasiveness of implicit visual priors across the vast majority of textual prompts.

A2.2 Detailed Prior Prevalence Profiles

While Table 2 in the main text summarized the top 3 most prevalent priors, we provide the complete top-10 prevalence profiles here. Figure A2 visualizes the full rankings for both SDXL and FLUX.1 across all five visual domains, showing the fraction of prompts where each prior appears in $\geq 80\%$ of the generated images.

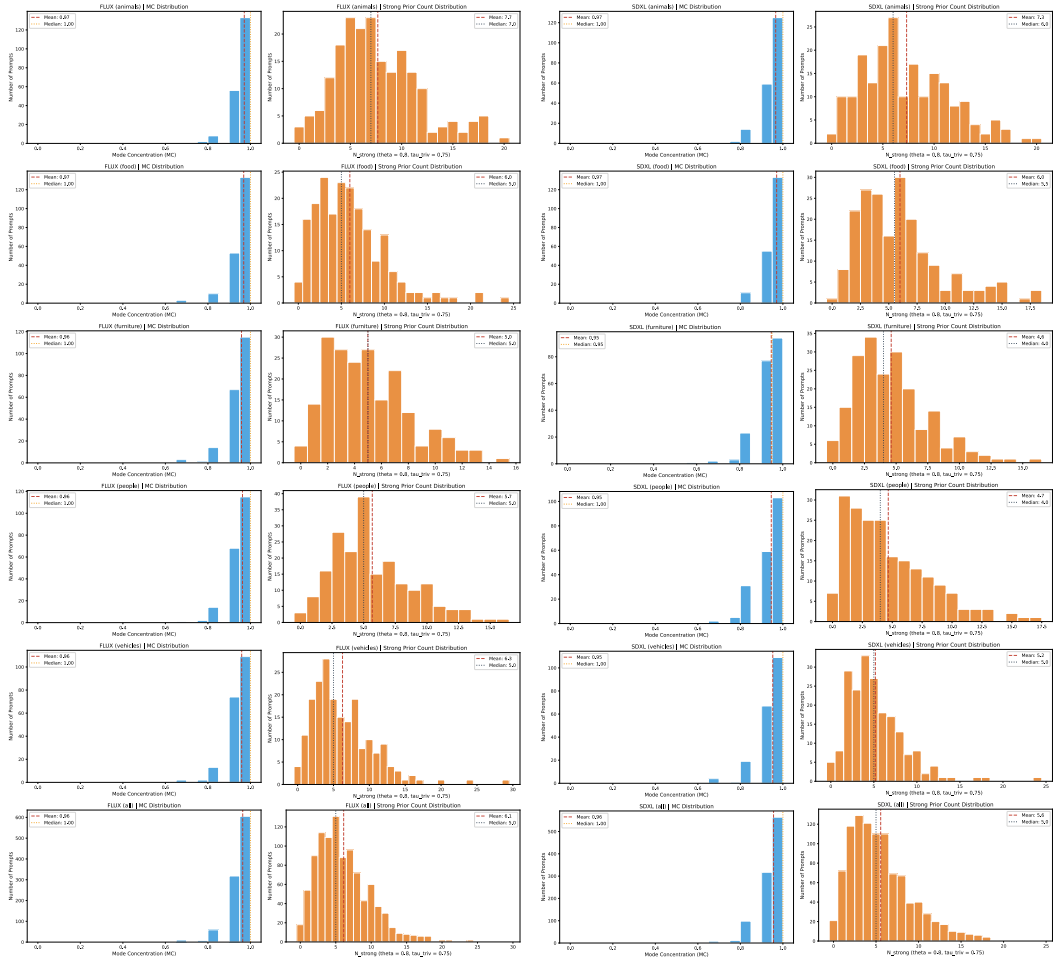

Figure A1: Distributions of MC and N_{strong}. Full distributions of Mode Concentration (blue) and Strong Prior Count (orange) across the evaluated models and domains.

A2.3 Cross-Model Divergence Distributions

While Table 3 summarized the aggregate mean and median cross-model metrics, we provide the full prompt-level distributions here. Figure A3 visualizes the spread of prior divergence (JSD) and top-5 overlap (J_5) across all prompts, highlighting the variance in how SDXL and FLUX.1 resolve visual underspecification across different domains.

A2.4 Additional Prior Profiles

In Figure A4 and Figure A5, we present additional examples of prompt-level implicit visual prior profiles, complete with generated image grids and co-occurrence heatmaps.

A3 Extended T2I Model Evaluations

In the main text, we contrasted SDXL (a latent diffusion model) with FLUX.1 (a flow-matching diffusion transformer). To demonstrate the generalizability of our method and track how implicit visual priors manifest across diverse modern architectures, we evaluate two additional text-to-image models:

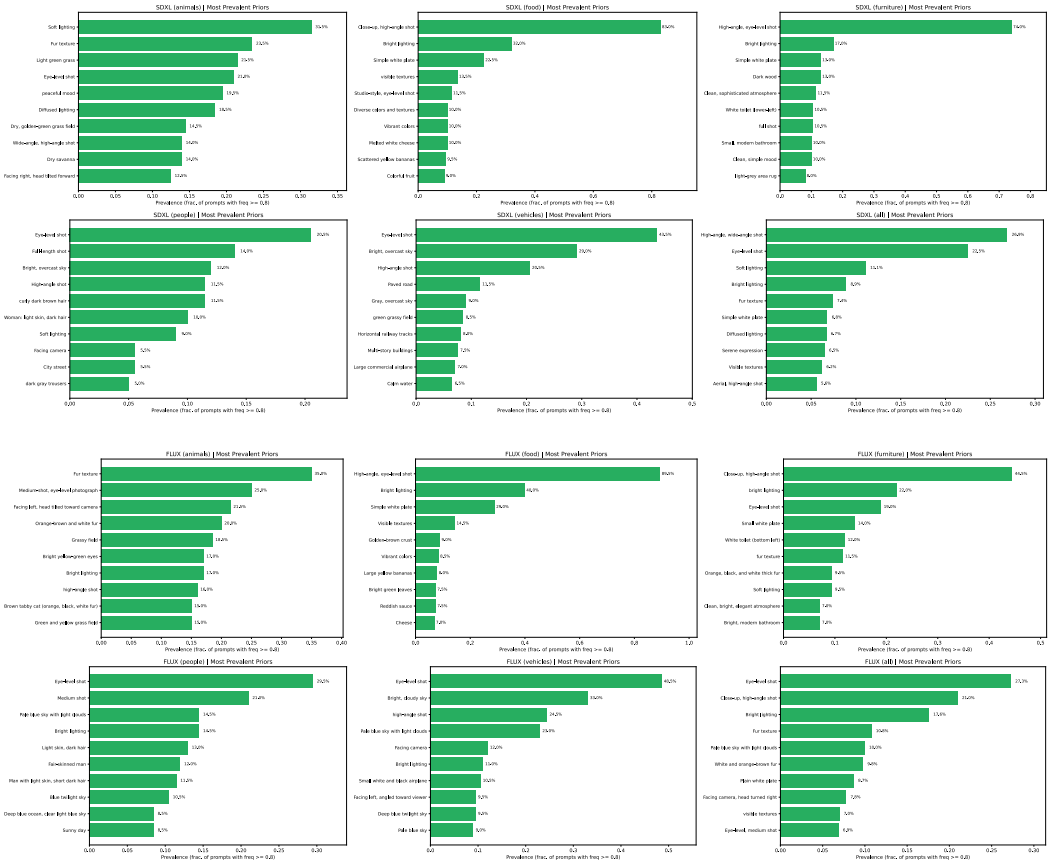

Figure A2: Full top-10 most prevalent implicit visual prior clusters for SDXL and FLUX.1 across five prompt domains. The x-axis represents the fraction of prompts where the prior appears in $\geq 80\%$ of the generated images.

- **Stable Diffusion 3.5 Large (SD-3.5 Large):** An 8.1-billion-parameter Multimodal Diffusion Transformer (MMDiT) (Esser et al., 2024) from Stability AI, designed for high-fidelity 1-megapixel image generation with advanced prompt adherence.
- **Z-Image Turbo:** A 6-billion parameter single-stream Diffusion Transformer (S3-DiT) from Tongyi-MAI (Cai et al., 2025), accelerated via Decoupled Distribution Matching Distillation (Liu et al., 2025) combined with reinforcement learning (Jiang et al., 2025), achieving sub-second inference in 8 function evaluations.

Using the exact same seeds and the 200 MS-COCO People prompts with m=20 images per prompt, we extracted and quantified the implicit visual priors for these additional models. Table A1 reports the aggregate prior statistics across these models.

Table A1: Aggregate Prior Statistics for additional T2I architectures on the 200 MS-COCO People prompts (m=20). Mode Concentration (MC) and the number of strong simultaneous priors (N_{strong}) remain remarkably high.

Model	$\mathbf{MC}_{\text{mean}}$	$\mathbf{MC}_{\text{median}}$	$N_{\text{strong, mean}}$	$N_{\text{strong, median}}$
SD-3.5 Large	0.965	1.000	6.3	6
Z-Image Turbo	0.989	1.000	10.0	10

Interestingly, the data reveals a striking pattern: Z-Image Turbo, the most aggressively distilled model via combined DMD and reinforcement learning, displays the highest MC_{mean} (0.989) and the most simultaneous strong priors ($N_{\text{strong, mean}} = 10.0$). This suggests that the distillation process,

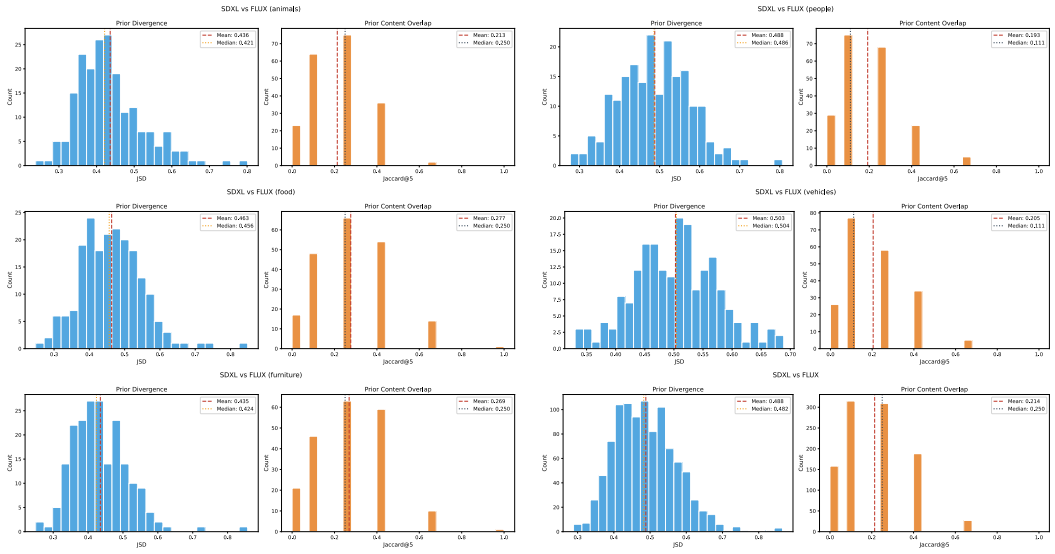

Figure A3: Distributions of cross-model prior divergence (JSD) and top-5 content overlap (J_5) across all prompts for SDXL vs. FLUX.1. Higher JSD indicates greater divergence, while higher J_5 indicates greater overlap in the top-5 strongest priors.

which compresses a model's sampling trajectory into fewer steps, may further amplify implicit visual biases by collapsing the generative distribution around dominant modes. Furthermore, a cross-model comparison between Z-Image Turbo and SD-3.5 Large reveals substantial divergence (mean JSD = 0.504), with a median J_5 overlap of only 0.111.

A4 Comparison with Closed-Set Classifiers

We benchmark ProbeGen against FairFace (Karkkainen & Joo, 2021), a widely adopted closed-set demographic classifier. While ProbeGen is designed to discover unconstrained visual priors, this comparison demonstrates that our pipeline maintains strong quantitative alignment with established methods when evaluated on standard social axes.

Table A2: Quantitative agreement between our method and the FairFace closed-set classifier on gender and race, evaluated on the 200 MS-COCO People prompts (m=20 images per prompt).

	Gender			Race		
Model	**Acc. (%) ↑**	**F1 (%) ↑**	**KL ↓**	**Acc. (%) ↑**	**F1 (%) ↑**	**KL ↓**
SDXL	76.94	75.32	0.0002	54.56	28.07	0.3537
FLUX.1	79.10	76.60	0.0059	60.53	25.53	0.3410

For a direct comparison, we use our NLI model to map the extracted attributes to FairFace's rigid taxonomy (2 genders, 7 racial categories). We conduct this evaluation on the 200 MS-COCO prompts from the People domain, utilizing the m=20 generated images per prompt to measure Accuracy, Macro F1, and Kullback-Leibler (KL) Divergence between the predicted attribute distributions. As shown in Table A2, our pipeline demonstrates strong agreement with FairFace on gender categorization. Both models achieve gender accuracies approaching 80% with near-zero KL divergence (down to 0.0002 for SDXL), indicating that our extracted text attributes faithfully mirror the underlying gender distribution identified by dedicated vision classifiers. In contrast, racial categorization exhibits lower Macro F1 scores (25-28%) and higher KL divergence (0.34). These scores align closely with recent open-set bias literature (D'Incà et al., 2024), which reports similar F1 ceilings for zero-shot racial classification. We attribute this discrepancy to an inherent taxonomy mismatch, i.e., FairFace forces generated images into mutually exclusive demographic categories (e.g., distin-

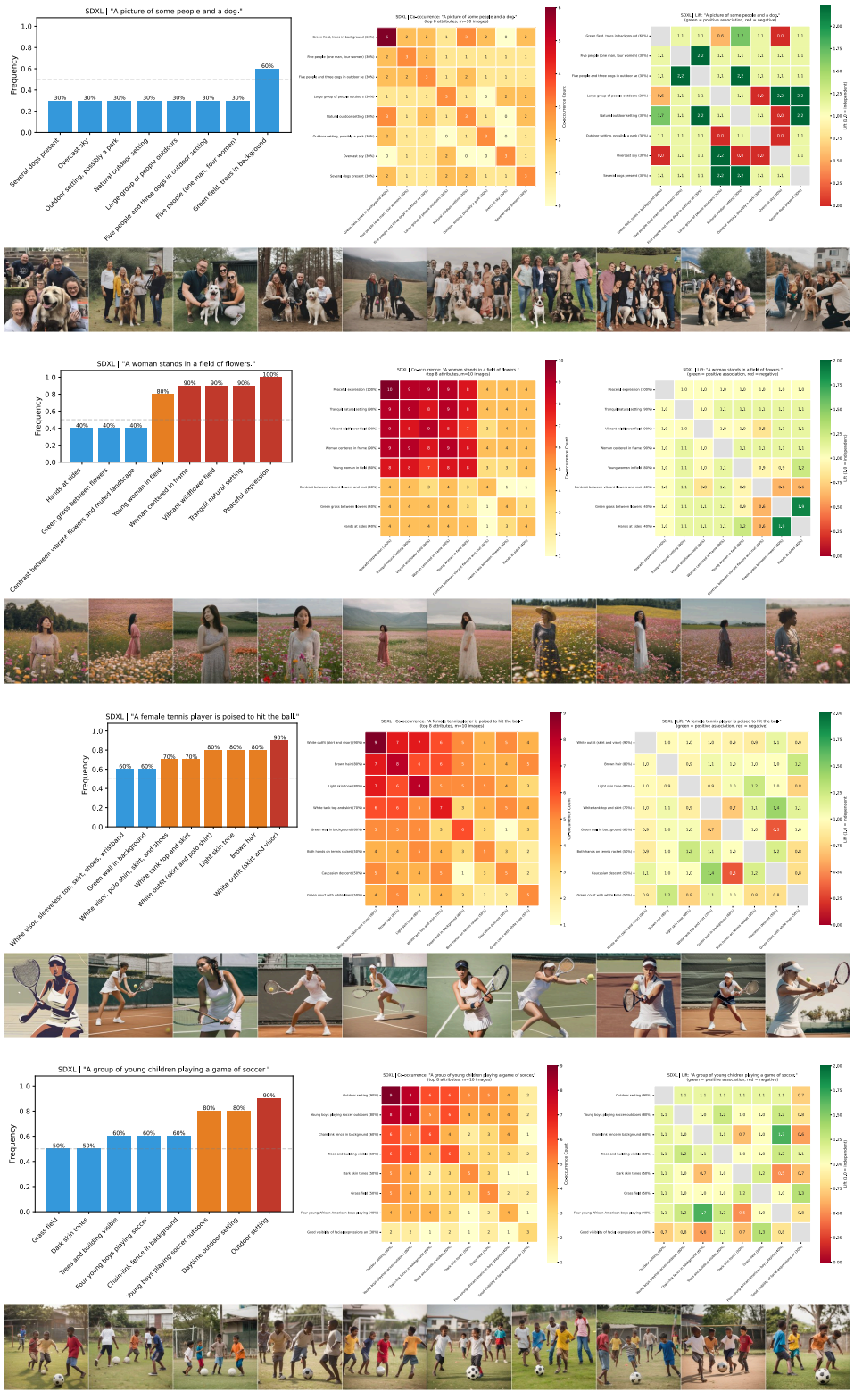

Figure A4: Additional Prior Profiles (SDXL). Image grids, frequency charts, and co-occurrence heatmaps for additional prompts.

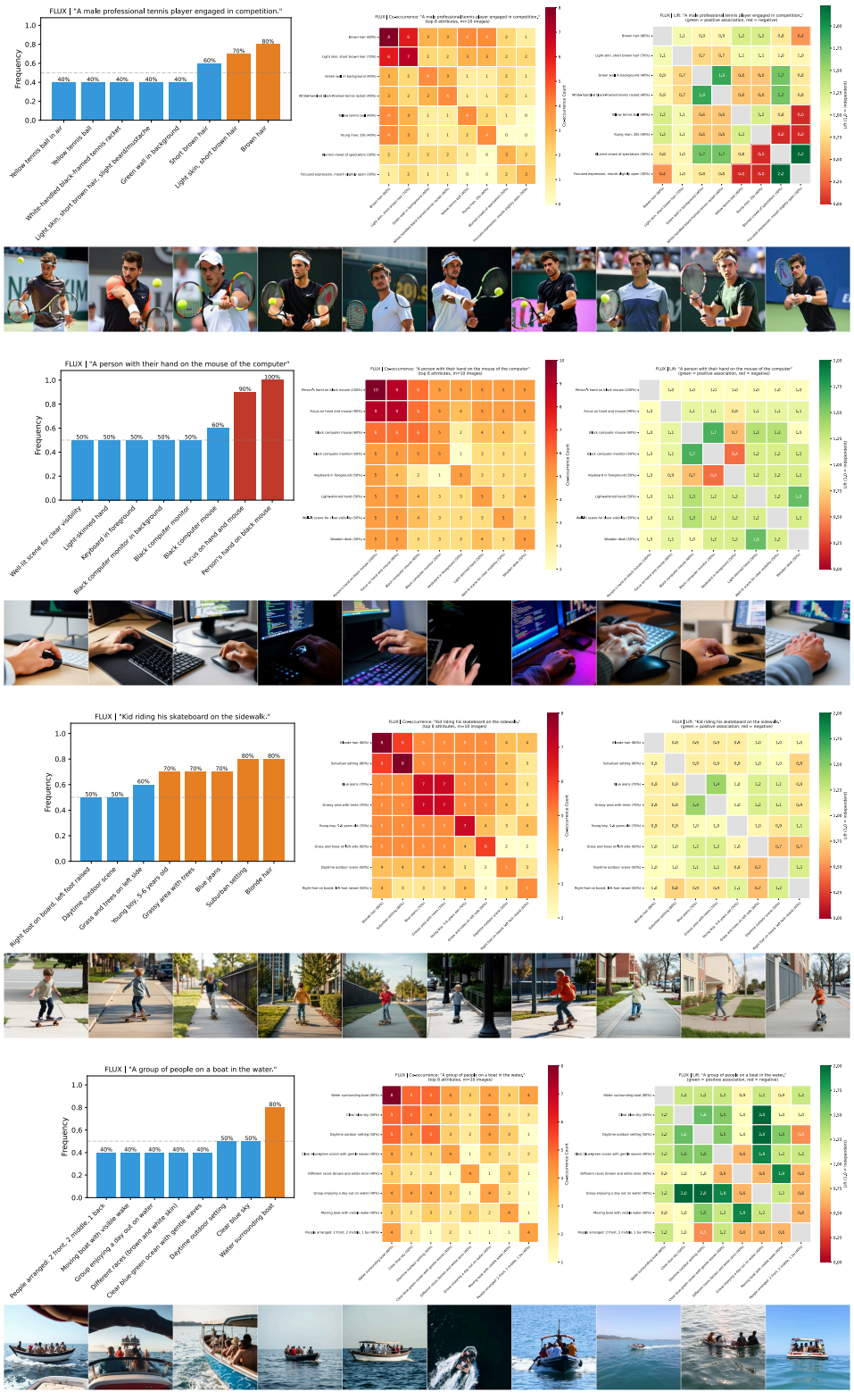

Figure A5: Additional Prior Profiles (FLUX.1). Image grids, frequency charts, and co-occurrence heatmaps for additional prompts.

guishing "East Asian" from "Southeast Asian"), whereas our method extracts broader descriptions that avoid such rigid categorization.

A5 Comparison with Open-Set Bias Detection

We compare our method against OpenBias (D'Incà et al., 2024), the primary state-of-the-art for open-set bias detection. OpenBias operates via a top-down, hypothesis-driven methodology, i.e., it queries an LLM to predict what biases might occur given a prompt, and then uses a VQA model to test those specific hypotheses.

To perform a direct comparison, we ran the OpenBias pipeline on the 200 MS-COCO prompts from the People domain, generating $m = 10$ images per prompt. We evaluated the two methods along two axes: *Discovery Coverage* and *Agreement on Shared Biases*.

A5.1 Discovery Scope Comparison

A fundamental difference between top-down and bottom-up approaches is the scope of attributes they can discover. Top-down methods evaluate hypotheses generated by an LLM based purely on the text prompt, whereas our bottom-up method extracts salient visual properties directly from the generated images.

This distinction highlights a critical vulnerability in top-down pipelines: LLMs frequently propose biases that turn out to be visually unverifiable in the generated outputs. For instance, an LLM might propose evaluating "person emotion" or "person gender" for a given prompt, but the generative model may happen to depict the subject from behind. Forcing a VQA model to answer these pre-defined questions on such images often leads to hallucinated categorizations and noisy measurements. Conversely, our approach only extracts attributes that are explicitly perceptible in the image, avoiding forced categorizations on ambiguous subjects.

To illustrate this discovery gap, Figure A6 displays the m=10 images generated for the representative prompt: *"A picture of a person on the side of the road."* Table A3 provides a side-by-side list of the findings from both methods for this exact image set. The comparison shows that our approach captures a broader profile without needing biases to be hypothesized in advance.



Figure A6: Generated Image Set for Comparison. The m=10 images generated by SDXL for the prompt *"A picture of a person on the side of the road."* which are evaluated by both methods in Table A3.

To formally quantify this difference in scope, we compared the coverage of both methods across all our prompts. On average, OpenBias proposed and detected 2.2 distinct biases per prompt. In contrast, our bottom-up method discovered 6.1 distinct strong priors ($f \geq 0.8$) per prompt, confirming that our framework provides a richer and more comprehensive profile.

Table A3: Comparison of Discovered Priors (Single Prompt). A side-by-side listing of the biases identified by OpenBias and the implicit visual priors discovered by our bottom-up framework for the images Figure A6.

OpenBias (Evaluated Bias Categories)	**Ours (Extracted Priors)**
• Person Gender • Person Clothing • Person Emotion	• Grayish tone • Isolated figure • Isolated, tranquil scene • Sense of scale and solitude • Isolated scene, no other people/vehicles • Sense of scale from lone figure • Solitary figure in a rural landscape • Two-lane road • Grassy areas on both sides

A5.2 Agreement on Shared Biases

We first evaluated whether both methods agree on biases they are both capable of detecting. Because OpenBias forces its evaluations into explicit categorical choices (e.g., Person Gender: `['Male', 'Female', 'Non-Binary']`), while our pipeline outputs free-form textual attributes, we bridged this formatting gap using the same NLI translation technique applied in our FairFace evaluation (Appendix A4).

For each bias category proposed by OpenBias, we used our zero-shot NLI model to map the free-form attributes extracted by our pipeline to the most logically entailed OpenBias choice, utilizing a confidence threshold of 0.50. We then computed the normalized entropy (H_{norm}) over our mapped distribution, allowing for a direct mathematical comparison with the entropy scores reported by OpenBias. A lower entropy indicates a more skewed, biased distribution.

Table A4: Agreement with OpenBias on shared bias categories. We used NLI to map our free-form extractions to OpenBias's categorical choices and compared the resulting normalized entropy distributions. The methods correlate positively on conspicuous attributes (Food, Gender, Breed) but diverge on ambiguous ones (Age) where forced-choice VQA models are prone to error.

Bias Category	**OpenBias Mean H_{norm}**	**Ours Mean H_{norm}**	**Pearson (r)**
Food Type (n=16)	0.218	0.213	0.727
Animal Breed (n=12)	0.305	0.226	0.444
Person Gender (n=118)	0.162	0.325	0.459
Vehicle Type (n=12)	0.145	0.117	0.370
Person Age (n=14)	0.428	0.146	-0.094
Overall	-	-	0.327 ($p < 0.0001$)

As shown in Table A4, the overall agreement between the two methodologies is statistically significant (Overall Pearson $r = 0.327, p < 0.0001$), though the correlation strength varies by attribute. The methods exhibit strong positive correlation on visually conspicuous traits like Food Type ($r = 0.727$), Person Gender ($r = 0.459$), and Animal Breed ($r = 0.444$). However, on inherently ambiguous traits like Person Age, the correlation drops to near zero ($r = -0.094$). This discrepancy highlights the vulnerability of top-down pipelines: OpenBias forces its VQA model to output an age category for every image, even when the subject is distant or obscured, introducing significant measurement noise. Conversely, our method only extracts age descriptors when they are visually salient, leading to cleaner, though differently distributed, prior measurements.