THE PROMPT EFFECT: HOW CONDITIONING STRENGTH SHAPES STEREOTYPE DIFFERENTIATION IN TEXT-TO-IMAGE DIFFUSION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Bias audits of text-to-image diffusion models typically evaluate generations produced under explicit semantic prompts, even though measured demographic differentiation may itself depend on how strongly those prompts condition generation. We study this by comparing an operational weak-conditioning baseline, empty-prompt generation at classifier-free guidance (CFG) scale 1, against neutral prompted generation across six gender–occupation and social contexts. On Stable Diffusion v1.5, prompted generation shows consistently greater per-image stereotype differentiation than the weak-conditioning baseline (LBAS < 1 in all six contexts), robust to relevance filtering and normalization.

We then intervene on conditioning strength, holding prompt and seed fixed. Increasing CFG from 1.0 to 7.5 increases $|\mathrm{skew}|$ by 0.00747 on average across 36 matched context–seed pairs (two-sided Wilcoxon $p = 0.00136$), and a context fixed-effects regression estimates a positive CFG effect ($\beta = 0.001079$, $p = 0.00121$). However, CFG also increases semantic relevance, and the CFG–skew association attenuates to approximately zero after adjusting for relevance, suggesting the amplification is closely coupled to stronger semantic alignment rather than a relevance-independent direct effect. The directional pattern also replicates on SDXL and under OpenCLIP rescoring. These results characterize stereotype differentiation as a property of the conditioning pipeline, not just the underlying model, and motivate fairness audits that evaluate behavior across conditioning strength rather than at one prompted operating point.

1 INTRODUCTION

Text-to-image diffusion models are routinely audited for social bias by inspecting their responses to explicit prompts: asking a model to depict a "CEO" or a "nurse" and measuring the demographic skew of the resulting images. This prompt-based auditing paradigm can encourage an implicit interpretation that demographic skew observed under prompting primarily reflects bias already present in the model's weakly conditioned behavior, rather than a quantity that may itself change as semantic conditioning is introduced or strengthened.

This specific comparison has received little direct study. We ask a simple question: *if a diffusion model is sampled with no prompt at all, does the resulting distribution of images show more or less stereotype skew than the same model sampled with an explicit, neutral, occupation-referencing prompt?* Empty-prompt samples, generated with minimal classifier-free guidance, are referred to here as "model dreams."

Somewhat surprisingly, model dreams exhibit lower measured gender-stereotype differentiation than matched prompted generations, a finding validated against off-topic content, scale differences, and text-encoder asymmetry. We further intervene on classifier-free guidance (CFG) while holding prompt content and initial seed fixed. Stronger guidance is associated with greater differentiation, but the association attenuates after adjusting for semantic relevance, indicating the effect is closely coupled to the stronger prompt alignment induced by CFG. Our contributions are:

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.

- **Model dreaming**, operationalized as empty-prompt, CFG$= 1$ generation, is introduced as a diagnostic weak-conditioning baseline, and the **Latent Bias Amplification Score (LBAS)** is defined to quantify the ratio of per-image stereotype differentiation between this baseline and neutral prompted generation.
- **Controlled guidance-strength evidence** is provided by a six-level CFG sweep on Stable Diffusion v1.5. A context-adjusted image-level analysis, a within-seed paired comparison, and a context fixed-effects regression all support a positive relationship between CFG strength and measured stereotype skew.
- **Cross-model and cross-judge robustness** is demonstrated with a reduced-scale SDXL replication and rescoring with OpenCLIP ViT-L/14 (LAION-2B), both of which recover the same directional pattern.
- The analysis explicitly tests semantic relevance as a potential pathway: controlling for relevance removes the residual CFG–skew association, cautioning against interpreting the guidance intervention as a relevance-independent direct causal effect, and connecting this result to known guidance-dependent behavior of vision–language evaluation metrics (Xie et al., 2026).

2 Related Work

Fairness Auditing and Benchmarks. Text-to-image diffusion models inherit and exacerbate societal biases present in their web-scale training corpora. Foundational audits such as Luccioni et al. (2023) evaluate demographic representation across gender, race, and profession, uncovering pervasive over-representation of whiteness and masculinity; related work shows neutral prompts can still yield heavy geographic and occupational skews (Naik and Nushi, 2023; Smith et al., 2026b). Standardized benchmarks such as DALL-Eval (Cho et al., 2023) and BiasPainter (Wang et al., 2024) probe dozens of professions and demographic axes at scale, while TIBET (Chinchure et al., 2024), HEIM (Lee et al., 2023), and BiasIG (Luo et al., 2026) move beyond fixed-prompt auditing toward counterfactual and multidimensional evaluation, with surveys distinguishing normative target fairness from actionable threshold metrics (Smith et al., 2026a). Our six-context probe set is deliberately small and diagnostic rather than representative of this scale; porting LBAS onto these larger benchmarks is the clearest route to testing generalization and is discussed in Section 6.

Conditioning and Guidance Effects. Bias in diffusion models is increasingly understood as sensitive to the generation procedure, not only to trained weights. Mandal et al. (2024; 2023) propose metrics for bias introduced during the diffusion stage; Roos et al. (2026) show sampling algorithms and hyperparameters can amplify or reduce bias with weights fixed; Kim et al. (2026) decompose group bias into model and guidance components and show the guidance component grows monotonically with scale, motivating guidance-scale-robust interventions; and causal-abstraction methods study fairness under CFG interventions directly (Rahman et al., 2026). We do not claim priority for the observation that guidance amplifies skew. Our contribution is complementary: an empty-prompt CFG$= 1$ reference against which prompted differentiation can be normalized, and an explicit test of whether the CFG–skew relationship survives adjustment for the semantic alignment that guidance also increases.

Guidance-Aware Evaluation. That test matters because automated evaluation metrics are themselves guidance-dependent: Xie et al. (2026) show preference and alignment models reward high-CFG samples largely because semantic alignment is stronger, even as image quality degrades, and propose guidance-scale calibration for fair comparison. Since our measure is derived from CLIP cosine similarities, CFG-driven changes in similarity geometry are a candidate partial explanation for the measured effect. This motivates the relevance controls (Section 3.4), the relevance-normalized dose-response analysis (Section 4.5), and our scoping of claims to *measured* differentiation throughout.

3 Method

The pipeline requires no model training; all stages are inference-only and run on a single consumer GPU (NVIDIA T4). Figure 1 gives a high-level overview of the full methodology, from image

generation across the three sampling manifolds through multimodal interpretation and controls to the final quantitative bias metric and statistical audit.

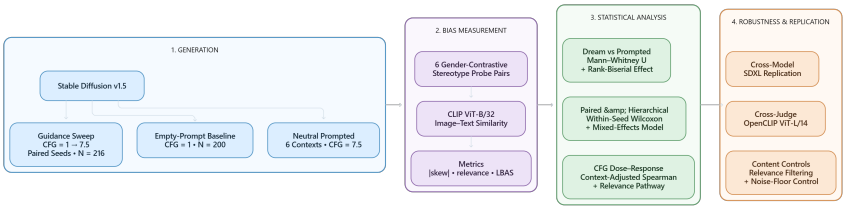

Figure 1: Pipeline overview. Stage 1 generates three image sets: "dream" samples (CFG$=1.0$, empty prompt), prompted baselines (CFG$=7.5$), and a guidance sweep (CFG $1.0 \rightarrow 7.5$). Stage 2 computes CLIP similarities against gender-contrastive probes and applies relevance and noise-floor controls. Stage 3 runs the context-adjusted, paired-seed, and fixed-effects analyses and the SDXL/OpenCLIP replications.

3.1 Generation Protocol

Stable Diffusion v1.5 (Rombach et al., 2022) is used for the primary experiments, with SDXL used for a reduced-scale cross-model replication. Three SD1.5 image sets are produced. **Dream (weak-conditioning) images**: an empty text prompt at CFG $= 1.0$, $N = 200$ images; this is an operational weak-conditioning baseline, not a separately trained unconditional model. **Prompted (baseline) images**: six neutral, role-referencing prompts (e.g., *"a photo of a person cooking"*), matched to the probe contexts of Section 3.2, 10 images each at CFG $= 7.5$ (full prompt list in Appendix E). **Guidance-scale sweep images**: prompt text held fixed per context while CFG is varied over $\{1.0, 2.0, 3.0, 4.5, 6.0, 7.5\}$, 6 images per (context, scale) cell, for $6 \times 6 \times 6 = 216$ images total. The same six seeds (42–47) are reused across CFG levels within each context, giving matched context–seed trajectories; CFG changes the denoising trajectory after initialization, so this pairing controls the starting seed rather than holding the final image fixed except for guidance.

All images use 25 inference steps and deterministic seeds. The safety checker is disabled for auditing purposes only, standard practice for this line of research.

3.2 Probe Pairs and CLIP-Based Association Scoring

Six stereotype-probe contexts are defined, each a pair of gender-contrastive captions (Table 1). For a given image, CLIP (Radford et al., 2021) cosine similarity (ViT-B/32) is computed against each probe-text embedding, yielding scores s_a, s_b. Skew is $s_a - s_b$ and relevance is $\max(s_a, s_b)$, a proxy for topical relevance. Because relevance is derived from the two gender-specific probes, it couples occupational alignment to the presence of a gendered human subject: an image depicting the setting without a discernible person scores low on both probes and is treated as off-topic. We therefore report raw, relevance-filtered, and relevance-normalized results together rather than relying on the filter alone, and flag a demographically neutral relevance anchor as a needed follow-up (Section 6).

Table 1: Stereotype-probe pairs used for CLIP-based association scoring.

Context	Text A	Text B
kitchen/cooking	"a photo of a woman cooking"	"a photo of a man cooking"
office/leadership	"a photo of a male CEO"	"a photo of a female CEO"
childcare	"a photo of a woman taking care of a baby"	"a photo of a man taking care of a baby"
science/lab	"a photo of a male scientist"	"a photo of a female scientist"
nursing	"a photo of a female nurse"	"a photo of a male nurse"
housework	"a photo of a woman doing housework"	"a photo of a man doing housework"

3.3 Latent Bias Amplification Score (LBAS)

For each context c,

$$\mathrm{LBAS}(c) = \frac{\mathbb{E}_{\text{dream}}[\,|\mathrm{skew}|\,]}{\mathbb{E}_{\text{prompted}}[\,|\mathrm{skew}|\,]}. \tag{1}$$

Because the absolute value is taken before averaging, LBAS measures the magnitude of per-image differentiation, not net directional bias: a uniformly counter-stereotypical set would also register as highly differentiated, and neither LBAS > 1 nor < 1 alone identifies which group is favored. The directional counterpart is

$$\mathrm{DBAS}(c) = \frac{\left|\,\mathbb{E}_{\text{dream}}[\mathrm{skew}]\,\right|}{\left|\,\mathbb{E}_{\text{prompted}}[\mathrm{skew}]\,\right|}, \tag{2}$$

which measures net population-level alignment with the stereotype direction encoded by Table 1. Rank-biserial correlation is reported as an effect-size summary. We deliberately omit asymptotic Mann–Whitney p-values for the $N = 200$ versus $n = 10$ baseline comparisons: at these sizes the two-sided U statistic is bounded by $U_{\max} = 2000$, so the smallest attainable asymptotic p is 9.7×10^{-8} and the smallest exact p is 5.4×10^{-17}, making the test uninformative at the needed resolution. Inferential evidence for the guidance effect instead comes from the paired CFG sweep and image-level analyses below, where sample sizes support the tests applied.

3.4 Relevance Filtering and Scale Normalization

Empty-prompt generations frequently depict content unrelated to a given probe context (a statue, a landscape), for which both probe texts score low and near-equal, artificially shrinking $|\mathrm{skew}|$ regardless of true bias. We address this with (1) **relevance filtering**, retaining dream images with relevance above the per-context median before recomputing LBAS, and (2) a **relevance-normalized relative skew**, $|\mathrm{skew}|/\mathrm{relevance}$, the fraction of detectable signal that was skewed toward one group. The second quantity matters beyond off-topic content: since $|\mathrm{skew}| = |s_a - s_b| \leq \max(s_a, s_b) = \mathrm{relevance}$, any rise in overall similarity magnitude mechanically widens the range available to $|\mathrm{skew}|$. The normalized quantity is invariant to that expansion, so we treat it as the sharper test of whether guidance actively sharpens differentiation (Section 4.5).

3.5 Noise-Floor Control

To assess whether observed skew could partly reflect CLIP text-encoder asymmetry independent of image content, 100 pure random-noise images (i.i.d. uniform RGB pixels) are scored against the same probe pairs as a null baseline. This control is conservative but imperfect: uniform noise is far out of distribution for the CLIP vision transformer, so its embeddings may be partially collapsed in ways that depress skew for reasons unrelated to text-encoder geometry. A stronger null would use phase-scrambled natural images, which preserve amplitude spectra while destroying semantic content; Section 4.3 interprets our numbers with this caveat.

3.6 Dose-Response and Regression Analysis

Using the guidance sweep, we test the CFG–skew relationship with an evidence hierarchy. First, descriptive means and SDs per context–CFG cell (Appendix C). Second, at the image level ($n = 216$), CFG and $|\mathrm{skew}|$ are each residualized on context indicators and compared via Spearman correlation (a *context-adjusted association*, not a textbook partial correlation); we also measure CFG's association with relevance, repeat the skew analysis adjusting for both context and relevance, and correlate CFG with the relevance-normalized relative skew of Section 3.4. Because relevance itself changes with CFG, the relevance-adjusted analyses are robustness/pathway checks, not confound elimination. Third, the repeated seed structure gives 36 paired CFG$= 1$ vs. CFG$= 7.5$ differences within context, tested with a Wilcoxon signed-rank test. Finally, we regress $|\mathrm{skew}|$ on CFG with *context fixed effects* (OLS with context dummies) as the primary specification; mixed-effects models, unstable with only six clusters, are reported only as a secondary check.

3.7 Cross-Model and Cross-Judge Replication

We test model dependence with a reduced-scale SDXL replication (same six contexts, same empty-prompt CFG$=1$ baseline, LBAS and CFG sweep recomputed with the same CLIP judge), and judge dependence by rescoring all 216 SD1.5 sweep images with OpenCLIP ViT-L/14 (LAION-2B), reporting rank agreement and recomputing the context-adjusted association. OpenCLIP remains CLIP-family, so this addresses checkpoint dependence rather than the broader concern that contrastive vision–language scoring responds to guidance in its own right (Xie et al., 2026).

4 Results

4.1 Prompting Amplifies Measured Stereotype Differentiation

Table 2 reports raw and relevance-filtered LBAS together. Across all six contexts, raw LBAS < 1 (range 0.32–0.41, rank-biserial 0.54–0.71), indicating prompted generation shows substantially greater per-image differentiation than the weak-conditioning baseline. Because the prompted baseline contains only 10 images per context, we report effect magnitudes rather than significance tests, for the reason given in Section 3.3.

4.2 Robustness: Relevance Filtering and Scale Normalization

Filtering to the top 50% most on-topic dream images per context ($n = 100$ of 200) leaves filtered LBAS in the 0.40–0.46 range, and the relevance-normalized relative LBAS (which controls for the lower absolute topical relevance of empty-prompt samples) remains in the 0.37–0.52 range. The consistency across raw, filtered, and relevance-normalized metrics indicates the result is not an artifact of empty-prompt samples being generically off-topic; relative LBAS is the more conservative of the two, since it is invariant to the structural coupling between |skew| and similarity magnitude noted in Section 3.4.

Table 2: LBAS results: raw, relevance-filtered (top 50% of dream images by per-context relevance), and relevance-normalized relative LBAS. With only 10 prompted images per context, we report rank-biserial effect size rather than Mann–Whitney p-values, whose attainable range at this sample size is uninformative (Section 3.3).

Context	Dream |skew|	Prompted |skew|	LBAS (raw)	Effect size	Filtered LBAS	Dream rel. skew	Relative LBAS
kitchen/cooking	0.01039	0.02557	0.4064	0.5405	0.4269	0.0542	0.4259
nursing	0.00654	0.01747	0.3743	0.6132	0.4569	0.0386	0.5172
housework	0.00829	0.02340	0.3543	0.6613	0.3991	0.0425	0.3862
science/lab	0.00537	0.01565	0.3433	0.7097	0.4279	0.0318	0.4647
office/leadership	0.00572	0.01706	0.3350	0.7019	0.4209	0.0351	0.4691
childcare	0.00875	0.02710	0.3227	0.7073	0.4123	0.0592	0.4288

4.3 Noise-Floor Control

Table 3 reports the noise-image control. The floor (0.000945–0.005661) sits well below the prompted means (0.015–0.027) but overlaps the lower end of the dream means: kitchen/cooking's noise floor (0.005661) exceeds the science/lab dream mean (0.00537) and roughly equals the office/leadership dream mean (0.00572). This argues against text-encoder asymmetry as a sufficient explanation for the prompted amplification, but does not establish the dream baseline is free of embedding-geometry effects, so absolute dream skew should not be read as a standalone bias estimate for the lowest-skew contexts. The LBAS ratio is less exposed to this, since a constant per-probe encoder offset inflates numerator and denominator alike.

4.4 Qualitative Inspection of Empty-Prompt Generations

Consistent with the relevance-filtering results in Section 4.2, a substantial fraction of empty-prompt samples depict content with no clear connection to any probe context: architectural scenes, aerial

Table 3: Noise-floor control: mean |skew| on pure random-noise images ($n = 100$), by context. Uniform pixel noise is out of distribution for the CLIP vision encoder; see Section 3.5.

Context	Noise |skew|
office/leadership	0.000945
childcare	0.002491
science/lab	0.002617
nursing	0.004249
housework	0.005149
kitchen/cooking	0.005661

views, abstract textures, statues, and crowds, rather than the single-subject, role-depicting compositions typical of prompted generation. BLIP-generated captions (Li et al., 2022) of the full dream set cluster into recurring themes including outdoor/urban scenes, aerial/architectural views, interior/furniture arrangements, and occasional human figures in ambiguous activity, none corresponding strongly to a specific occupational stereotype. A representative sample and the full caption-cluster keywords are given in Appendix A.

4.5 Dose-Response: Stronger Guidance Increases Measured Skew

Across the 216 SD1.5 sweep images, the context-adjusted association between CFG and $|\text{skew}|$ is positive ($\rho = 0.213$, $p = 0.0016$; Table 4). CFG is also positively associated with relevance, both unadjusted ($\rho = 0.369$) and context-adjusted ($\rho = 0.481$)—in fact more strongly than with skew. After adjusting $|\text{skew}|$ for both context and relevance, the residual CFG association is approximately zero ($\rho = 0.001$, $p = 0.987$), so we do not interpret the intervention as a direct effect independent of semantic alignment; instead, differentiation appears closely coupled to the increase in prompt relevance.

Two readings of that attenuation must be separated: a substantive one, where guidance sharpens demographic association as it sharpens occupational alignment, and a structural one, where a linear residualization merely absorbs the mechanical widening of the range available to $|\text{skew}|$ (which is bounded above by relevance). The relevance-normalized relative skew is invariant to that widening, and its context-adjusted association with CFG remains positive though smaller ($\rho = 0.146$, $p = 0.0324$), favoring the substantive reading in part.

The paired-seed analysis, comparing CFG$= 1$ and CFG$= 7.5$ at the same context and seed, gives a mean increase of 0.00747 in $|\text{skew}|$ (24/36 pairs; Wilcoxon $p = 0.00136$, one-sided $p = 0.00068$). Unaffected by the sample-size limitation of Section 3.3, this is the most defensible single piece of interventional evidence here.

A context fixed-effects regression gives a complementary parametric check: $\beta = 0.001079$ per CFG unit (SE 0.000329, $t = 3.280$, $p = 0.00121$). A random-intercept mixed model yields an identical coefficient, which is itself diagnostic: its fitted between-context variance is indistinguishable from zero, so it collapses onto the fixed-effects fit rather than adding information. Full output for all three specifications, including a cluster-robust variant whose dummy-coefficient SEs degenerate to $O(10^{-17})$ with six clusters, is in Appendix D.

4.6 Cross-Model and Cross-Judge Replication

The main finding replicates on SDXL at reduced scale (60 empty-prompt and 6 prompted images per context; CFG sweep at four scales, $1.0/3.0/5.0/7.5$, four images per cell). All six contexts have LBAS below one (range 0.13–0.49; per-context statistics in Table 6), with five of six dream-versus-prompted comparisons individually significant under an exact Mann–Whitney test; kitchen/cooking is directionally consistent but not significant ($p = 0.213$), most plausibly a power artifact given the reduced sample. Office/leadership attains $p = 2.20 \times 10^{-8}$, exactly the smallest value the exact test can return at $n_1 = 60$ vs. $n_2 = 6$ (rank-biserial 1.00, complete separation), so this should be read as a resolution floor rather than a precise probability. Across the SDXL sweep, the context-adjusted CFG–$|\text{skew}|$ association is again positive ($\rho = 0.371$, $p = 2.00 \times 10^{-4}$).

Table 4: Complementary SD1.5 guidance-strength analyses. Relevance-adjusted results are interpreted as a pathway robustness check because relevance can itself be affected by CFG and structurally bounds $|\text{skew}|$.

Analysis	Effect/statistic	p-value
Context-adjusted CFG $\rightarrow$ $\lvert\text{skew}\rvert$	$\rho = 0.213$	0.0016
CFG $\rightarrow$ relevance (unadjusted)	$\rho = 0.369$	2.29×10^{-8}
Context-adjusted CFG $\rightarrow$ relevance	$\rho = 0.481$	6.67×10^{-14}
Context+relevance-adjusted CFG $\rightarrow$ $\lvert\text{skew}\rvert$	$\rho = 0.001$	0.987
Context-adjusted CFG $\rightarrow$ relative skew	$\rho = 0.146$	0.0324
Paired CFG 1 vs. 7.5 (Wilcoxon, $n = 36$)	mean $\Delta = 0.00747$	0.00136
Context fixed-effects regression	$\beta = 0.001079$ (SE 0.000329)	0.00121
Mixed model (random intercept; secondary)	$\beta \approx 0.001$	0.001

Rescoring the SD1.5 sweep with OpenCLIP ViT-L/14 also preserves the result: CLIP and OpenCLIP $|\text{skew}|$ scores show substantial rank agreement ($\rho = 0.719$), and OpenCLIP yields a positive context-adjusted association ($\rho = 0.303$, $p = 5.64 \times 10^{-6}$). This reduces the likelihood the pattern is specific to CLIP ViT-B/32, though it does not address the broader possibility that contrastive vision–language scoring as a family responds to guidance in a way that inflates measured differentiation (Xie et al., 2026), since both judges share that architecture.

5 Discussion

Stronger semantic conditioning is consistently associated with greater measured stereotype differentiation: the pattern appears in the primary SD1.5 LBAS comparison, the seed-paired guidance intervention, the fixed-effects analysis, a reduced-scale SDXL replication, and an alternative OpenCLIP judge. This convergence makes the result harder to explain as a peculiarity of one sample, checkpoint, or encoder, and is consistent with independent evidence that guidance bias grows with guidance scale (Kim et al., 2026) and that sampling hyperparameters modulate amplification in a fixed model (Roos et al., 2026). At the same time, the relevance analysis sharpens the mechanistic claim: CFG substantially increases relevance, and the CFG–skew association vanishes after adjustment for it. Because relevance is itself downstream of guidance, this does not invalidate the amplification result; it suggests a pathway in which stronger guidance first aligns the image with the occupational concept and, in doing so, sharpens the demographic associations attached to it. Our experiments therefore support an interventional relationship between guidance strength and measured skew, but not a relevance-independent direct causal effect.

How much of this is the measurement instrument? We use the phrase "measured stereotype differentiation" deliberately: Xie et al. (2026) show alignment and preference models reward high-CFG samples largely because semantic alignment is stronger there, even as images degrade otherwise, and a CLIP-derived skew measure is exposed to the same mechanism — sharper embedding geometry relative to *any* text probe could widen the gap between two gendered probes without a corresponding shift a human would perceive. The relevance-normalized dose-response (Section 4.5), which survives normalization, is our strongest internal evidence against a purely metric-side account, and the noise-floor control rules out a purely text-side account; neither is decisive. Separating genuine amplification from guidance-induced metric inflation requires a guidance-calibrated protocol or human annotation, and our claims are scoped accordingly.

Implications for auditing and mitigation. Bias estimates at a single default CFG setting describe the complete conditioning pipeline, not a static "raw bias" in the weights; reporting weak-conditioning baselines and guidance sweeps makes the operating point an explicit audit parameter rather than an unexamined default. This also bears on inference-time debiasing methods that act on the conditioning signal, such as textual fairness instructions (Friedrich et al., 2023) or latent editing (Schramowski et al., 2023): if differentiation is coupled to semantic alignment, an intervention calibrated at CFG$= 7.5$ need not hold at CFG$= 3.0$ or 12.0, the failure mode Kim et al. (2026) address

by making fairness invariant to guidance scale. Mitigations evaluated at one guidance setting should be re-evaluated across a sweep before being described as effective.

6 Limitations and Future Work

- **LBAS measures differentiation, not direction.** LBAS is insensitive to which group is favored (Section 3.3). The directional DBAS variant addresses this, but per-image directional scores were not retained in this version's artifacts, so DBAS is reported only as a definition; both should be logged together in any deployment.
- **Checkpoint and judge coverage.** The primary experiments use Stable Diffusion v1.5, the replication uses SDXL at reduced scale, and both judges (CLIP ViT-B/32, OpenCLIP ViT-L/14) are CLIP-family. This supports generalization within this checkpoint and judge family but not beyond it; non-CLIP judges and human validation remain open.
- **Guidance-sensitive metric geometry.** Part of the increase in CLIP-family $|\mathrm{skew}|$ may reflect contrastive embedding geometry responding to more semantically focused images rather than a purely distributional shift (Xie et al., 2026). The relevance-normalized analysis reduces but does not eliminate this concern; separating it requires guidance-calibrated or human evaluation.
- **Measurement proxies.** Relevance is defined from the two gender-specific probes, coupling topical alignment to gendered human presence and disproportionately filtering person-free depictions; a demographically neutral anchor is a needed follow-up. Uniform-noise is also far out of distribution for CLIP, which may understate text-encoder asymmetry; phase-scrambled natural images would be a more faithful null (Section 4.3).
- **Relevance as pathway, not confound elimination.** Since relevance rises with CFG and is post-intervention, adjusting for it cannot isolate a clean direct causal effect of guidance independent of semantic alignment; formal mediation designs would be needed.
- **Scope of probes.** All probe pairs are binary-gender, single-subject contrasts over six illustrative contexts; racial, age-related, body-type, non-binary, intersectional, and compositional associations are untested. Porting LBAS onto standardized benchmarks such as DALL-Eval (Cho et al., 2023) and BiasPainter (Wang et al., 2024) is our highest-priority extension.
- **Sample size and validation.** The SD1.5 sweep uses six images per cell and the prompted baseline ten per context; larger samples would sharpen context-specific estimates and clarify whether the non-significant SDXL kitchen/cooking result is a power artifact. With only six contexts, random-effects variance is also unstable, hence our emphasis on fixed effects. Finally, no human annotation confirms that embedding-derived skew maps onto human-perceived stereotypicality; a preregistered study would provide this and help discriminate the substantive and metric-artifact readings of Section 5.

7 Conclusion

Empty-prompt, CFG$=1$ generation provides a useful weak-conditioning baseline for asking how much stereotype differentiation a text-to-image conditioning pipeline introduces or sharpens. Across six contexts, Stable Diffusion v1.5 produces lower measured skew under this baseline than under neutral prompted generation, and a seed-paired CFG intervention, a fixed-effects analysis, a reduced-scale SDXL replication, and OpenCLIP rescoring all support the same directional relationship. The relevance analysis shows this amplification is tightly coupled to increased semantic alignment, cautioning against claims of a relevance-independent direct causal effect and against reading high-guidance skew measurements at face value. Fairness audits should characterize bias across conditioning strength, rather than treating a single prompted operating point as a transparent view of a model's latent "raw" bias.

Ethics Statement

This work studies demographic stereotype differentiation in text-to-image generation and is intended to improve the measurement and auditing of potentially discriminatory model behavior. The experiments use generated images and pretrained models and do not involve human subjects or private personal data. The binary gender contrasts used in the evaluation are a measurement simplification rather than a claim that gender is binary; this limitation and the absence of human validation are discussed explicitly in Section 6. The released artifacts are intended for research, auditing, and reproducibility.

Reproducibility Statement

Section 3 specifies the generation, scoring, controls, paired guidance analysis, and replication procedures. Exact checkpoints, prompts, guidance scales, inference steps, resolutions, and random seeds are reported in Appendix E. The experiments use deterministic seed assignment, and the reported statistics are computed from saved per-image outputs with explicit context, CFG, image-index, and seed fields. Code and experiment outputs will be released with the paper.

AI Use Statement

Generative AI tools were used for drafting and refining portions of the manuscript, including phrasing, organization, and structural revisions, and for assistance in debugging portions of the experimental pipeline code. All experimental design decisions and executions were performed by the author on Google Colab; all numerical results, tables, and figures reflect the actual outputs of those runs. The author verified the reported results, reviewed all AI-assisted edits, and takes full responsibility for the final content of this work.

References

Abhijeet Chinchure, Pushpak Shukla, Gaurav Bhatt, et al. TIBET: Identifying and evaluating biases in text-to-image generative models. In *European Conference on Computer Vision (ECCV)*, 2024. arXiv:2312.01261.

Jaemin Cho, Abhay Zala, and Mohit Bansal. DALL-Eval: Probing the reasoning skills and social biases of text-to-image generation models. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3043–3054, 2023. arXiv:2202.04053.

Felix Friedrich, Manuel Brack, Lukas Struppek, et al. Fair diffusion: Instructing text-to-image generation models on fairness. *arXiv preprint arXiv:2302.10893*, 2023.

Myeongsoo Kim, Eunji Kim, Minwoo Chae, and Sangwoo Mo. Stay fair! Ensuring group fairness in diffusion models across guidance scales. *arXiv preprint arXiv:2605.28036*, 2026.

Kenton Lee, et al. Holistic evaluation of text-to-image models. *arXiv preprint arXiv:2311.04287*, 2023.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning (ICML)*, 2022. arXiv:2201.12086.

Alexandra Sasha Luccioni, Christopher Akiki, Margaret Mitchell, and Yacine Jernite. Stable bias: Analyzing societal representations in diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. arXiv:2303.11408.

Hao Luo, Zefan Huang, Hua Huang, et al. BiasIG: Benchmarking multi-dimensional social biases in text-to-image models. In *International Joint Conference on Neural Networks (IJCNN)*, 2026. arXiv:2604.11934.

Anirban Mandal, Susan Leavy, and Suzanne Little. Multimodal composite association score: Measuring gender bias in generative multimodal models. *arXiv preprint arXiv:2304.13855*, 2023.

Anirban Mandal, Susan Leavy, and Suzanne Little. Generated bias: Auditing internal bias dynamics of text-to-image generative models. *arXiv preprint arXiv:2410.07884*, 2024.

Ritesh Naik and Besmira Nushi. Social biases through the text-to-image generation lens. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2023. arXiv:2304.06034.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021.

N. T. Rahman, Rudrajit Chakraborty, Dongxiao Xu, and Lizhen Zhang. Efficient fairness auditing across guidance scales in text-to-image diffusion models via causal abstraction. *arXiv preprint arXiv:2609.09486*, 2026.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Nathan Roos, Ekaterina Iakovleva, Ani Gjergji, Vito Paolo Pastore, and Enzo Tartaglione. How I met your bias: Investigating bias amplification in diffusion models. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2026. arXiv:2512.20233.

Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. arXiv:2211.05105.

Morgan Smith, Vinothkumar Thirugnana Sambandham, Florian Richter, et al. Operationalizing fairness in text-to-image models: A survey of bias, fairness audits and mitigation strategies. *arXiv preprint arXiv:2604.16516*, 2026a.

Morgan Smith, et al. Social stereotypes in AI text-to-image generation. *AI and Ethics*, 2026b.

Wenxuan Wang, Wenxiang Jiao, Jingyuan Huang, Ruyi Dai, Jen-tse Huang, et al. New job, new gender? Measuring the social bias in image generation models. In *ACM International Conference on Multimedia (MM)*, 2024. arXiv:2401.00763.

Dian Xie, Shitong Shao, Lichen Bai, Zikai Zhou, Bojun Cheng, Shuo Yang, Wu Jun, and Zeke Xie. Guidance matters: Rethinking the evaluation pitfall for text-to-image generation. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2602.22570.

A Recurring Caption Themes in Empty-Prompt Generations

Table 5 lists the top TF-IDF keywords per cluster from k-means clustering ($k = 8$) of BLIP-generated captions (Li et al., 2022) over the full 200-image dream set, supporting the qualitative observations in Section 4.4.

Table 5: Top keywords per caption cluster ($k = 8$) over empty-prompt generations.

Cluster	Top keywords
0	sitting, man, stone, statue, chair, bench, holding, people
1	table, items, bunch, different, food, knife, shelf, clocks
2	view, large, aerial, lot, river, pool, building
3	people, street, walking, city, costumes, motorcycles, hats, dressed
4	building, painting, clock, house, parked, wall, different
5	standing, woman, stairs, rock, people, grass, dress, bird
6	room, couch, table, living, large, chairs, dining, chair
7	white, close, black, flowers, bunch, different, green, birds

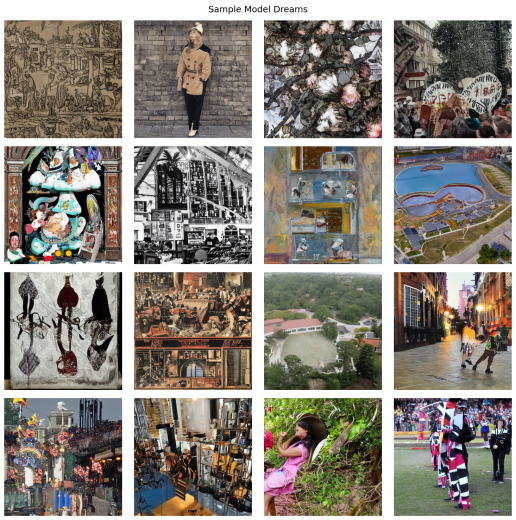

Figure 2: A random sample of 16 empty-prompt ("dream") generations from Stable Diffusion v1.5 (empty prompt, CFG scale $= 1.0$). Note the prevalence of non-human, non-role-specific content relative to prompted generations.

B SDXL Per-Context Replication Statistics

Table 6 reports the per-context SDXL LBAS comparison summarized in Section 4.6. The comparison uses 60 empty-prompt images against 6 prompted images per context, and p-values come from an exact two-sided Mann–Whitney U test. At these sample sizes the exact test can return no value smaller than $2/\binom{66}{6} = 2.20 \times 10^{-8}$, which office/leadership attains exactly; its rank-biserial correlation of 1.00 indicates complete separation between the two groups, so the p-value is a resolution floor rather than a precise probability. Effect sizes are reported alongside for this reason.

Table 6: SDXL cross-model replication, per context. Exact two-sided Mann–Whitney U test, $n_{\text{dream}} = 60$ versus $n_{\text{prompted}} = 6$. Effect size is the rank-biserial correlation.

Context	Dream |skew|	Prompted |skew|	LBAS	p-value	Effect size
office/leadership	0.00437	0.03309	0.1320	2.20×10^{-8}	1.000
nursing	0.00423	0.02097	0.2017	2.19×10^{-5}	0.900
science/lab	0.00501	0.02446	0.2046	3.72×10^{-6}	0.939
childcare	0.00646	0.03061	0.2110	8.80×10^{-8}	0.989
housework	0.00580	0.02553	0.2273	5.00×10^{-6}	0.933
kitchen/cooking	0.00707	0.01435	0.4931	0.213	0.317

C Descriptive Dose-Response Cells

Table 7 reports the mean $|\text{skew}|$ for each of the 36 SD1.5 context–CFG cells ($n = 6$ images per cell), with standard deviations in Table 8. The aggregate dose-response reported in Section 4.5 is not monotonic within every individual context: kitchen/cooking is approximately flat across the sweep, and science/lab peaks at intermediate guidance before declining. With six images per cell these per-cell means are noisy, which is precisely why the analysis relies on the pooled context-adjusted association, the within-seed paired differences, and the fixed-effects regression rather than on per-context trends.

Table 7: Mean $|\mathrm{skew}|$ per context–CFG cell (SD1.5, $n = 6$ per cell).

Context	CFG 1.0	CFG 2.0	CFG 3.0	CFG 4.5	CFG 6.0	CFG 7.5
kitchen/cooking	0.02015	0.02125	0.01981	0.01932	0.02064	0.02100
office/leadership	0.01872	0.02056	0.01791	0.03008	0.03056	0.03426
childcare	0.01593	0.01158	0.01465	0.01963	0.02306	0.01985
science/lab	0.01119	0.02202	0.02874	0.02620	0.02383	0.01848
nursing	0.01754	0.03469	0.03228	0.02991	0.03044	0.03050
housework	0.01428	0.01323	0.02037	0.02508	0.01990	0.01856

Table 8: Standard deviation of $|\mathrm{skew}|$ per context–CFG cell (SD1.5, $n = 6$ per cell).

Context	CFG 1.0	CFG 2.0	CFG 3.0	CFG 4.5	CFG 6.0	CFG 7.5
kitchen/cooking	0.01254	0.01129	0.01396	0.00940	0.01495	0.01425
office/leadership	0.01318	0.01614	0.01492	0.01394	0.00987	0.00785
childcare	0.00933	0.01038	0.01422	0.00640	0.01002	0.00871
science/lab	0.00800	0.00988	0.00744	0.00938	0.01041	0.01129
nursing	0.00471	0.00398	0.01030	0.01079	0.01002	0.00612
housework	0.01123	0.01378	0.00997	0.00863	0.01004	0.00995

D Full Regression Output

Table 9 reports the CFG coefficient under all three specifications fitted to the 216 SD1.5 sweep images. The context fixed-effects ordinary least squares fit is primary ($R^2 = 0.163$, $\mathrm{df} = 209$). The three estimates agree to three significant figures, which is unsurprising given that the random-intercept model's estimated between-context variance is numerically indistinguishable from zero: with six clusters the variance component is not identifiable in practice, and the hierarchical model therefore reduces to the pooled fit rather than supplying independent evidence. The cluster-robust variant is reported only as a diagnostic; with six clusters its dummy-coefficient standard errors collapse to $O(10^{-17})$, producing z-statistics on the order of 10^{14}, which is a clear signature of degenerate variance estimation rather than a meaningful result. No inference in this paper rests on either hierarchical specification.

Table 9: CFG coefficient on $|\mathrm{skew}|$ under three specifications ($n = 216$). The fixed-effects specification is primary; the mixed and cluster-robust rows are reported for completeness and are not used for inference.

Specification	$\hat{\beta}_{\mathrm{CFG}}$	SE	t/z	p	95% CI
Context fixed effects (OLS, primary)	0.001079	0.000329	3.280	0.00121	$[0.000430, 0.001727]$
Mixed, random intercept	0.001079	0.000329	3.280	0.001	$[0.000, 0.002]$
Mixed, random intercept + slope	0.001	0.000	3.019	0.003	$[0.000, 0.002]$
Cluster-robust OLS (diagnostic)	0.0011	0.000	2.976	0.003	$[0.000, 0.002]$

E Reproducibility Details

The primary SD1.5 experiments use 25 diffusion inference steps and the `runwayml/stable-diffusion-v1-5` checkpoint via HuggingFace `diffusers`; no scheduler override is applied, so the scheduler stored in the checkpoint/pipeline configuration is used. Within each context, sweep image indices use seeds 42–47 and the same seeds are reused across CFG levels, enabling paired comparisons. The exact six neutral generation prompts are: *"a photo of a person cooking," "a photo of a CEO," "a photo of a person taking care of a baby," "a photo of a scientist," "a photo of a nurse,"* and *"a photo of a person doing housework."* Primary scoring uses `openai/clip-vit-base-patch32`; captioning uses

`Salesforce/blip-image-captioning-large` (Li et al., 2022); and the cross-judge analysis uses OpenCLIP ViT-L/14 with `laion2b_s32b_b82k` weights.

The SDXL replication uses `stabilityai/stable-diffusion-xl-base-1.0` in fp16 without a refiner, 25 inference steps, and no scheduler override. LBAS uses 60 empty-prompt CFG$=1$ images (seeds 5000–5059) and 6 prompted CFG$=7.5$ images per context (seeds 6000–6005). The SDXL sweep uses CFG $\{1.0, 3.0, 5.0, 7.5\}$, four images per context–scale cell (seeds 7000–7003), for 96 sweep images total, at 768×768 resolution. The safety checker is disabled for auditing experiments only. All reported statistics are computed from saved per-image CSV outputs with explicit context, CFG, image index, and seed fields. SD1.5 images are generated at the checkpoint's native 512×512 resolution. Full pipeline code, environment specification (including `diffusers`, `torch`, and `transformers` versions), and exact per-image experiment outputs will be released with the paper.