WHEN STEERING WORKS: PREDICTING DEBIASING BENEFIT FROM DENOISER RESPONSES

Anonymous authors
Paper under double-blind review

ABSTRACT

Semantic steering can reduce stereotypes in text-to-image generation, but it can also change content that the prompt asks for. Can we predict which prompts will benefit before decoding their images? We introduce Directional Gain (G), which compares two denoiser responses at the same latent state: the change caused by steering and the change caused by replacing a stereotype prompt with its counterfactual. We measure agreement between these directions relative to random edits. Across SD3.5, FLUX, and Qwen-Image, with 1,295 test prompts per model, G predicts improvement better than response size. At the top 20%, selection by G gives larger stereotype reductions and usually smaller prompt-matching losses. Its advantage in combined-score improvement over response size is 4.7–20.7 points per selected prompt under Qwen3-VL; GPT-5.6 also favors G in all three models. Averaged over six single-probe choices, one probe retains 89–93% of the mean improvement obtained by six-probe selection. It uses fewer denoiser evaluations than generating a clean–steered image pair and requires no image decoding or judging. With the steering settings fixed, the ranking also selects prompts that benefit on new generation seeds.

1 INTRODUCTION

Text-to-image (T2I) models can reproduce social stereotypes even when a prompt leaves the relevant attribute unspecified (Naik & Nushi, 2023; Luccioni et al., 2023; Zhang et al., 2024). A growing line of work therefore seeks to steer generation by modifying text representations, attention, or denoising dynamics (Brack et al., 2023; Zhang et al., 2023; Friedrich et al., 2025; Park et al., 2025; Han et al., 2025). Such interventions, however, are not uniformly beneficial. The same steering rule may substantially reduce a stereotype for one prompt, have little effect on another, or alter content that the prompt was meant to preserve.

This variability raises a question that is largely separate from how a steering direction is constructed: *given a specified semantic intervention, can we tell in advance which prompts are likely to benefit from it?* Ideally, this decision should be made before generating and judging the final image. Applying steering indiscriminately can trade stereotype reduction for prompt fidelity, whereas evaluating every candidate intervention through full image generation undermines the appeal of lightweight inference-time control.

A natural signal is the size of the model's response to an intervention. Yet a large response is not necessarily a useful one. The denoiser may change substantially while moving generation in a direction unrelated to the intended counterfactual, or while disrupting other prompt content. More broadly, identifying a concept in an internal representation does not imply that intervening on that representation will reliably produce the desired behavioral change (Hase et al., 2023; Wu et al., 2025b). In diffusion models, an intervention must also propagate through the denoising dynamics before it affects the final image (Tinaz et al., 2025; Görgün et al., 2026). The direction of the denoiser's response may therefore help predict whether steering will improve the image.

We introduce *Directional Gain* (G), a pre-generation diagnostic of this directional response. For each neutral prompt, we assume a supplied stereotype–counterfactual pair that specifies the intended semantic change. At the same latent state and denoising step, we compare two changes in the denoiser output: the change induced by steering the neutral prompt, and the reference change induced by replacing the stereotype prompt with its counterfactual. G measures the alignment between these

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.

response directions, standardized against norm-matched random edits. It therefore asks a simple question: *does the model respond to the steering intervention in the same direction as it responds to the intended semantic counterfactual?* Computing G requires denoiser evaluations, but no image decoding or image-level judgment.

We evaluate G on Stable Diffusion 3.5 Medium (SD3.5) (Stability AI, 2024), FLUX.2-klein-9B (FLUX) (Black Forest Labs, 2026), and Qwen-Image (Wu et al., 2025a), using 1,295 held-out prompts per model from IMPLICIT-Bench (Dai et al., 2026). Across all three generators, G ranks subsequent steering benefit more reliably than response magnitude alone. Under our primary image judge, its Spearman correlation with held-out improvement ranges from 0.190 to 0.240, compared with -0.016 to 0.084 for response magnitude. This advantage also improves prompt selection. At 20% coverage, selecting by G gives larger mean gains than selecting by response magnitude in every model, with the largest advantage in Qwen-Image. The same ordering holds under an independent GPT-5.6 judge. The selected prompts show larger stereotype reductions and, in most settings, smaller losses in prompt fidelity.

Directional Gain can also be computed cheaply. Our main experiments average six denoiser probes. Using one probe retains roughly 90% of the full score's correlation with observed improvement and 89–93% of the mean gain from top-20% selection, averaged over the six possible single probes. In all three models, this single-probe variant requires fewer transformer forward calls than generating one clean–steered image pair, while avoiding image decoding and judging altogether. The ranking also remains useful when images are generated with new random seeds. Our code is available at `https://anonymous.4open.science/r/direction_gain-DDFE`.

Our contributions are:

- We study *selective semantic steering*: predicting which prompts will benefit from a fixed intervention before decoding or judging their images.
- We introduce *Directional Gain* (G), a decode-free diagnostic that compares the denoiser response induced by steering with the response induced by the intended stereotype–counterfactual change.
- Across SD3.5, FLUX, and Qwen-Image, we show that G consistently ranks steering benefit better than response magnitude and identifies prompts with a more favorable trade-off between stereotype reduction and prompt preservation under two independent image judges. A single denoiser probe retains 89–93% of the mean improvement from six-probe top-20% selection while costing less than evaluating one clean–steered image pair.

2 Related Work

Bias and debiasing in text-to-image generation. Text-to-image models have been shown to reproduce social stereotypes involving occupations, demographic appearance, gender presentation, and social context (Naik & Nushi, 2023; Luccioni et al., 2023; Zhang et al., 2024; Wang et al., 2024). Evaluating such behavior is itself non-trivial: measured bias can depend on prompt construction, caption–image mismatch, or errors in downstream attribute classifiers (Seshadri et al., 2024; Teo et al., 2023). Recent work has also broadened evaluation to multilingual prompts and human preferences (Klassert et al., 2026). We therefore evaluate both the targeted stereotype and whether the generated image still follows the original neutral prompt, so that stereotype reduction is considered together with prompt preservation. A number of methods intervene at inference time to reduce unwanted associations. SEGA modifies semantic guidance during denoising (Brack et al., 2023; Friedrich et al., 2025), ITI-GEN learns inclusive prompt embeddings (Zhang et al., 2023), and other approaches edit associations in cross-attention (Orgad et al., 2023). More recent methods explicitly address the tension between changing a targeted attribute and preserving other prompt content. Entanglement-Free Attention separates attributes to modify from those to retain (Park et al., 2025), while LightFair debiases the text encoder and restricts when modified conditioning is applied (Han et al., 2025). These approaches primarily study *how* to construct or apply a debiasing intervention. Our question is complementary: *once a steering rule is fixed, which prompts are likely to benefit from applying it?*

From representation to intervention. A separate line of work studies where semantic concepts and biases are represented inside generative models. Diffusion Lens examines how visual semantics emerge across text-encoder layers (Toker et al., 2024); GradBias attributes generated bias to individual prompt words (D'Incà et al., 2025); and sparse-autoencoder and mechanistic analyses identify internal features associated with editable concepts or biased generation (Surkov et al., 2025; Shi et al., 2025). Spectral analysis has identified concept-associated directions in cross-attention and tested their effects under intervention (Bae et al., 2026). However, identifying a concept internally does not necessarily imply that intervening on that representation will produce the desired behavioral change. This distinction has also been observed in language models, where localization and editability can diverge (Meng et al., 2022; Hase et al., 2023), and where seemingly meaningful steering features do not always translate into effective control (Wu et al., 2025b). We study whether the denoiser's response to a specified intervention predicts improvement in the final image.

When and where interventions affect diffusion models. The effect of an intervention in a diffusion model can depend strongly on when and where it is applied. Prior work studies when composition, style, and fine-grained concepts emerge over denoising time (Tinaz et al., 2025), how prompt-conditioned concepts persist or disappear across diffusion steps (Görgün et al., 2026), and whether internal probes can help localize generative decisions to particular layers (Zaleska et al., 2026). These studies help determine when and where to intervene. Our main experiments keep the intervention rule fixed and edit the text-encoder output. We then test whether the direction of the denoiser change predicts which prompts will benefit. We also compare the fixed output intervention with prompt-specific internal-layer selection in Appendix C. Because the internal-layer advantage does not generalize consistently across models, the main paper focuses on predicting prompt-level steering benefit at this common intervention site.

3 METHOD

Directional Gain measures how closely the denoiser's response to an edit follows the response specified by a stereotype–counterfactual prompt pair. We use this signal to choose which prompts to steer, before decoding or judging images.

3.1 STEERING SETUP

We use IMPLICIT-Bench triplets: a neutral prompt, a stereotype version, and a counterfactual version, with text-encoder outputs c^n, c^s, and c^c. We omit case and model indices for readability.

We form the steering direction d by averaging the counterfactual-minus-stereotype differences at the first two encoder rows whose change exceeds a fixed threshold, or the single row if only one changes. We retain its original length and add it to up to two selected rows $\mathcal{A}$ of the neutral representation:

$$\widetilde{c}^n_j = c^n_j + \alpha d, \qquad j \in \mathcal{A}. \tag{1}$$

To locate $\mathcal{A}$, we replace a candidate word in the neutral prompt with "thing" and select the first two encoder rows that change beyond the same threshold. The remaining rows stay unchanged. For each model, the weights, edit location, row-selection rule, and strength α are fixed. The threshold, candidate words, and model settings are specified in Appendix A.

3.2 COMPARING DENOISER RESPONSES

We run generation under the neutral prompt to a predefined denoising step and hold the latent state fixed. We then evaluate the denoiser under each of the four text conditions in Figure 1, forming a *probe*. For probe q, let $v_q(c) = v_\theta(x_q, t_q; c)$ be the conditional denoiser output, where x_q and t_q are the latent state and timestep immediately before that step. We compute

$$\begin{aligned} r_q &= v_q(c^c) - v_q(c^s), \\ \Delta v_q &= v_q(\widetilde{c}^n) - v_q(c^n). \end{aligned} \tag{2}$$

The reference r_q is the change produced by swapping the stereotype prompt for its counterfactual, and the steering change Δv_q is the change produced by our edit to the neutral prompt. If steering moves the denoiser the way the counterfactual does, the two vectors point in the same direction, so we compare them by cosine similarity.

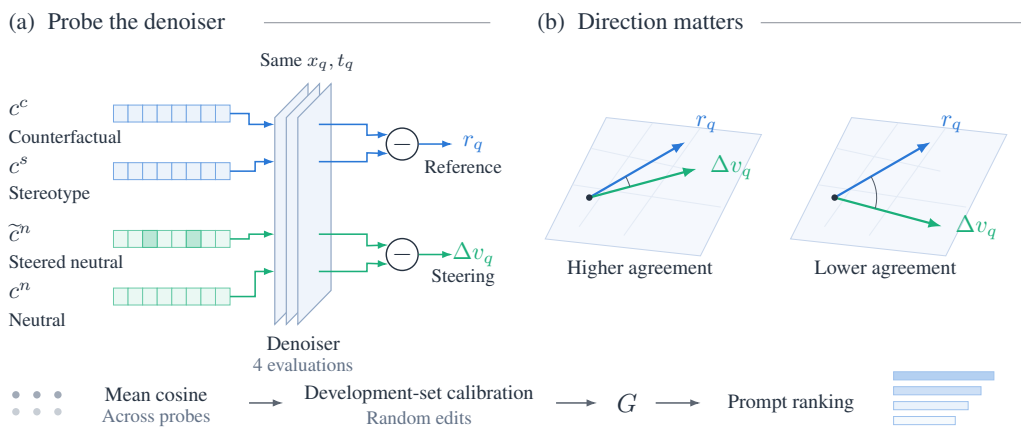

Figure 1: Directional Gain. (a) Reference and steering responses at a shared latent state. (b) Equal response sizes can have different directions.

3.3 Directional Gain and Prompt Selection

For a set of probes $\mathcal{Q}$, Directional Gain averages these cosine similarities and expresses the result relative to random edits:

$$G = \frac{\text{mean}_{q \in \mathcal{Q}}[\cos(\Delta v_q, r_q)] - \mu_0}{\sigma_0}. \tag{3}$$

Our main configuration uses six probes: two denoising steps per model, repeated across three seeds. To obtain μ_0 and σ_0, we apply random directions to development cases with the same direction length, edited rows, and strength as the semantic edit. We average each random edit's probe cosines, then compute the mean and sample standard deviation of these averages separately for each model.

We rank prompts by G, steer the highest-scoring fraction, and generate the remaining prompts without an edit. Standardization expresses G as the distance from random edits in units of their standard deviation, without changing the ranking within a model.

4 Experiments

4.1 Setup and Evaluation

We evaluate SD3.5, FLUX, and Qwen-Image on IMPLICIT-Bench. We split the prompt triplets into development and test sets, with no overlap after text normalization. Retaining only cases where steering changes the text representation leaves 498 development and 1,295 test cases per model. For each test case, we generate images with the same two random seeds under three conditions: no edit ("clean"), semantic steering, and a random-direction edit with the same magnitude and edited locations as semantic steering.

We use Qwen3-VL (Bai et al., 2025) and GPT-5.6 (OpenAI, 2026) to evaluate images with the scoring prompts from IMPLICIT-Bench (Dai et al., 2026). Each judge rates stereotype expression $b(I)$ from 0 to 5 and separately checks whether the image follows the neutral prompt. We give no credit to images that fail this check and measure improvement as

$$U(I) = \begin{cases} 100\left(1 - \dfrac{b(I)}{5}\right), & \text{if the image follows the neutral prompt,} \\ 0, & \text{otherwise,} \end{cases} \tag{4}$$

$$R = \text{mean}_s\left[U(I_s^{\text{steer}}) - U(I_s^{\text{clean}})\right].$$

The image score U ranges from 0 to 100, and R is its mean change across the two seeds, in score points. Changes in stereotype ratings and prompt-matching rates are measured in rating points and percentage points, respectively.

We compare G with response size and a text-only predictor to assess what its directional information contributes. Response size, $M = \text{mean}_{q \in \mathcal{Q}} \|\Delta v_q\|_2$, is computed from the same Δv_q as G and tests whether the magnitude of the denoiser change alone predicts benefit. A text-only predictor tests whether benefit can be predicted from the prompt representations and edit size, without probing the denoiser. Its inputs are the direction length, the distance between c^s and c^c, and the edit length relative to the neutral representation. We fit the text predictor on development cases and evaluate it on test cases using ratings from each judge.

4.2 Predicting Steering Benefit

Across the full test set, semantic steering raises the mean combined score over clean generation and random edits in all three models under both judges, as shown in Table 1. It reduces stereotypes but also lowers prompt-matching rates, with the largest changes in Qwen-Image.

Table 1: Mean image scores under two judges on the same 1,295 test prompts per model and two random seeds.

	Stereotype rating (0–5) ↓			Matching prompt (%) ↑			Combined score ↑		
Model	Clean	Random edit	Steer	Clean	Random edit	Steer	Clean	Random edit	Steer
Qwen3-VL									
SD3.5	2.95	2.91	**2.43**	**67.8**	66.0	65.1	24.7	24.5	**30.4**
FLUX	3.09	3.06	**2.40**	**80.0**	78.5	73.1	28.4	28.6	**34.8**
Qwen-Image	3.43	3.25	**2.30**	**81.8**	75.3	66.1	24.6	24.1	**31.5**
GPT-5.6									
SD3.5	2.50	2.46	**2.03**	**70.3**	68.8	68.1	32.1	31.3	**37.1**
FLUX	2.80	2.79	**2.11**	**79.8**	79.3	73.8	33.3	33.3	**39.6**
Qwen-Image	3.04	2.84	**1.97**	**81.6**	74.2	65.4	30.8	29.7	**35.6**

We measure how well each score ranks prompts by their steering benefit using the Spearman correlation with observed improvement R. The correlations in Table 2 are highest for G in every model under both judges. The advantage over response size is largest in Qwen-Image, where M has almost no association with improvement. Under GPT-5.6, G and M have similar correlations in SD3.5 (paired steering gains, prediction differences, and intervals in Appendix B.1).

Table 2: Spearman correlation with steering improvement R.

	Qwen3-VL			GPT-5.6		
Model	M	Text	G	M	Text	G
SD3.5	0.084	0.092	**0.190**	0.167	0.058	**0.174**
FLUX	0.059	0.161	**0.211**	0.054	0.186	**0.190**
Qwen-Image	−0.016	0.103	**0.240**	0.009	0.183	**0.239**

4.3 Selecting Prompts to Steer

We select prompts with the highest G or response size M and compare the mean steering benefit at the same selection fraction.

Under Qwen3-VL, prompts selected by G benefit more at every tested fraction below 100% (Figure 2). When prompts are divided into five groups by G, the highest-scoring group benefits most in each model (Appendix B.2).

At 20% coverage, selection by G improves the combined score more than selection by M under both judges. The advantage in Table 3 is largest in Qwen-Image, at about 21 points per selected prompt, and smallest in SD3.5, at about 4–5 points. The selected prompts have larger stereotype reductions and usually smaller prompt-matching losses. In SD3.5 under GPT-5.6, the matching losses are similar.

Table 3: Outcomes among the top 20% selected by each score. Combined-score gains are in score points; matching loss is the clean matching rate minus the steered rate, in percentage points (pp).

	Combined-score gain ↑		Stereotype reduction ↑		Matching loss (pp) ↓	
Model	by G	by M	by G	by M	by G	by M
Qwen3-VL						
SD3.5	**17.26**	12.51	**1.45**	1.27	**4.2**	7.5
FLUX	**18.11**	6.80	**1.64**	1.45	**11.0**	22.4
Qwen-Image	**27.61**	6.91	**2.86**	2.08	**24.1**	33.8
GPT-5.6						
SD3.5	**17.92**	13.55	**1.42**	1.14	4.2	**3.9**
FLUX	**20.12**	5.79	**1.61**	1.52	**7.5**	21.8
Qwen-Image	**27.45**	6.45	**2.76**	2.09	**22.4**	30.7

The results above count only the selected prompts. In practice the remaining prompts are generated without an edit, so Figure 3 scores the whole test set under that policy, using Qwen3-VL ratings. We also compare against random selection, which uniformly samples the same fraction of prompts and applies semantic steering to them. At the same prompt-matching loss, selection by G gives greater stereotype reduction than random selection in every model, and steering more prompts reduces stereotypes further at the cost of more matching loss. The trade-off is similar to that of the clean-image predictor B (Section 4.4), which predicts steering benefit from two generated and judged clean images per prompt. Appendix B.2 adds joint stereotype-and-matching success rates and negative-outcome risk.

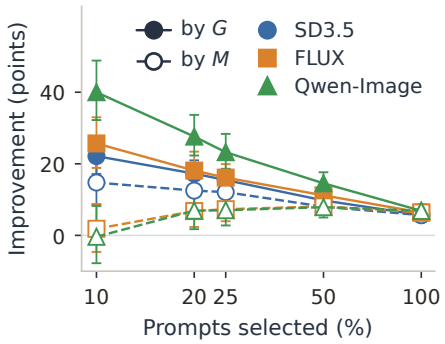

Figure 2: Mean improvement per selected prompt. Bars show 95% bootstrap intervals.

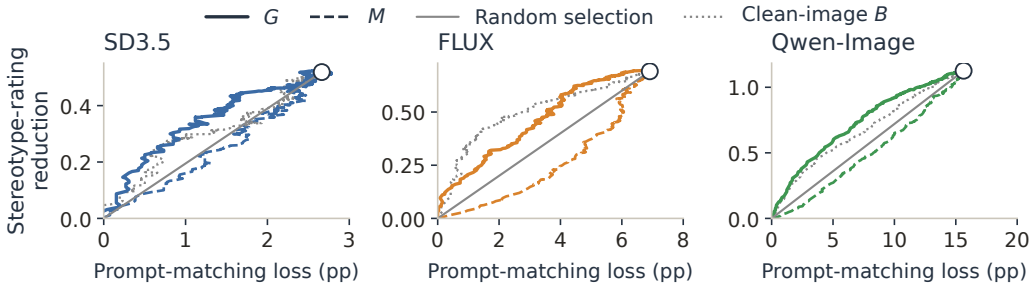

Figure 3: Mean changes over the full test set (Qwen3-VL). Open circles mark steering every prompt.

4.4 Understanding What G Predicts

Response size is associated with two opposing changes: greater stereotype reduction and greater prompt-matching loss. These changes can offset each other in the combined score. G is more strongly associated with stereotype reduction than M, and more weakly with matching loss (Table 4).

Table 4: Spearman correlations with stereotype reduction and prompt-matching loss. Matching loss is the clean matching rate minus the steered rate.

	Stereotype reduction ↑		Matching loss ↓	
Model	G	M	G	M
SD3.5	**0.34**	0.26	**0.05**	0.09
FLUX	**0.33**	0.32	**0.01**	0.20
Qwen-Image	**0.46**	0.35	**0.13**	0.30

Figure 4 shows Qwen-Image prompts from the top and bottom 5% of G, with the same random seed for each clean–steered pair. In the high-G examples, the prompt pair changes how a person is described ("a man"/"a woman", "rock"/"classical" guitarist), and steering changes the person while preserving the scene. In the low-G examples, the pair changes a trait or context ("dropout"/"over-educated", "factory"/"office", expertise in "the Vedas"/"rock and roll"), and the images change little.



Figure 4: High- and low-G examples in Qwen-Image. Ratings show clean→steered stereotype scores.

Pairs that differ only in a gendered subject word tend to have higher G and larger steering benefits than other contrasts. They make up 30–36% of the top 20% selected by G, compared with 6–19% for M. Within the other contrasts, G still selects prompts with gains above their group mean in every model, with the largest advantage over M in Qwen-Image. Within subject-swap pairs, the two scores select groups with similar gains (group definitions, full results, and intervals in Appendix B.4).

To assess whether G adds information beyond how well the model performs before steering, we compare the clean-image predictor B with $B + G$. Predictor B combines the text features from Section 4.1 with two features of two independently generated clean images: their headroom $H = 100 - \bar{U}^{\text{clean}}_{\text{ind}}$ and their prompt-matching rate. $B + G$ adds G as an input. We fit both predictors on development cases and evaluate them on test cases.

Adding G lowers test mean squared error in all three models in Table 5, with paired intervals excluding zero in FLUX and Qwen-Image. Top-20% selection gains also rise in SD3.5 and Qwen-Image, but the selection-gain point estimate falls in FLUX despite the lower prediction error. Full results, paired intervals, and a separate comparison against a generated clean–steered image pair are reported in Appendix B.4.

Table 5: Adding G to the clean-image predictor on all 1,295 test cases per model (Qwen3-VL). Selection gain is mean R among the top 20%, in score points.

	Test MSE ↓		Selection gain ↑	
Model	B	$B+G$	B	$B+G$
SD3.5	474.11	**468.59**	13.40	**19.00**
FLUX	700.04	**660.72**	**20.77**	19.31
Qwen-Image	983.92	**900.06**	24.44	**30.81**

4.5 Score and Steering Ablations

We vary the reference direction, scoring rule, steering strength, and probe timing on 300 test cases per model. Reference and scoring variants are compared within the same ablation run using Qwen3-VL ratings. The stereotype-to-counterfactual reference gives higher correlations than the neutral-to-counterfactual reference in all three models (Table 6). Weighting the cosine by response size slightly raises the FLUX correlation and lowers the Qwen-Image correlation.

Table 6: Reference and scoring choices: correlation with R on the same 300-case subset per model, measured within one ablation run using Qwen3-VL ratings.

Score / reference	SD3.5	FLUX	Qwen-Image
G, stereotype-to-counterfactual	**0.144**	0.190	**0.199**
Neutral-to-counterfactual reference	0.123	0.129	0.174
Cosine weighted by response size	0.142	**0.199**	0.144
Response size alone	0.089	0.116	−0.047

Across the tested steering strengths, G remains positively correlated with improvement in SD3.5 and FLUX when recomputed at each strength. The strength with the largest mean improvement differs between models (Appendix B.5). The sweep covers SD3.5 and FLUX only, with Qwen-Image evaluated at its main strength.

Prediction quality also depends on when the denoiser is probed. In FLUX, correlation with improvement is highest at step 1 and falls to 0.027 at the final sampler step. Appendix B.5 gives per-step and averaged results and checks whether recomputing a probe reproduces the same values.

4.6 Efficient Probing and Reuse Across Seeds

A single probe retains 89–93% of the six-probe selection gain, averaged over the six single-probe choices under Qwen3-VL. In Table 7, single-probe G also selects prompts with larger gains than response size measured at the same probe. Individual configurations and paired intervals are reported in Appendix B.3.

Table 7: Probe cost and top-20% selection benefit under Qwen3-VL. Single-probe gains average the six configurations. Retention is the single-probe mean gain divided by the six-probe mean gain.

	Forward calls ↓			One-probe gain ↑		Gain retained (%) ↑
Model	One probe	Six probes	Image pair	G	M	
SD3.5	32	108	56	**15.84**	11.52	91.8
FLUX	12	48	16	**16.08**	6.83	88.8
Qwen-Image	38	126	200	**25.65**	5.76	92.9

One probe uses fewer transformer forward calls than a clean–steered image pair in every model. Six probes cost more calls than an image pair in SD3.5 and FLUX, and fewer in Qwen-Image. Most of the six-probe cost comes from neutral generation across three seeds to reach the probe steps. Both probe configurations avoid image decoding and judging, whereas an image pair requires two decodes and four judge calls (costs and retained correlations for every tested probe budget in Appendix B.3).

Once computed, G can be reused for new generation seeds with the same prompt, model, reference pair, and steering settings. On 300 cases per model, we reuse the original top-20% cutoffs and generate images with six new random seeds. Selected prompts gain 14.4 points in SD3.5, compared with 4.1 when the full subset is steered. The corresponding gains are 13.5 versus 4.7 in FLUX and 21.9 versus 6.3 in Qwen-Image. Averaging improvement over more generated image pairs also increases the correlation between G and measured improvement (Appendix B.3).

5 LIMITATIONS

We evaluate IMPLICIT-Bench prompts with stereotype–counterfactual pairs. Broader prompt distributions and steering rules remain to be tested. Image outcomes are measured by automated judges; human evaluation would help assess perceived stereotype reduction and image quality. Prompt-matching losses remain even among selected prompts, leaving room to combine G with steering methods that better preserve prompt content (Park et al., 2025).

6 CONCLUSION

The direction of a denoiser's response helps identify which prompts will benefit from a specified semantic edit. Directional Gain measures this response before image decoding. Across three generators and two image judges, it selects prompts with larger stereotype reductions and greater combined-score gains than response size. Averaged over the single-probe choices, one probe retains 89–93% of the six-probe selection gain while using fewer forward calls than a clean–steered image pair. The same ranking also selects useful prompts on new generation seeds, allowing the measurement to be reused.

AI USE STATEMENT

We used generative AI tools to assist with manuscript editing and analysis code. Sections 3 and 4.1 describe the models used in the experiments. The authors are responsible for the paper's content.

ETHICS STATEMENT

This study examines how to reduce social stereotypes in generated images. The benchmark's stereotype–counterfactual pairs specify the associations being evaluated, and the examples are included to study these associations. These paired contrasts simplify identities and social contexts that are more varied in practice. In particular, binary gender and race categories can exclude people whose identities do not fit those categories and overlook how different aspects of identity intersect.

Automated ratings reflect both the benchmark rubric and the judge models' own biases and errors. We report prompt matching alongside stereotype reduction so that changes to the content requested by the user remain visible in the evaluation.

REPRODUCIBILITY STATEMENT

We define the steering intervention and Directional Gain in Section 3. The data splits, edit construction, generator settings, seed assignments, calibration, judging procedure, and uncertainty calculations are specified in Appendix A. Full numerical comparisons and probe-budget results are available in Appendix B. We also document numerical sensitivity when probes are recomputed in Appendix B.5.

REFERENCES

Jun-Hyun Bae, Wonyong Jo, Jaehyup Lee, and Heechul Jung. Mechanistic dissection of cross-attention subspaces in text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 19657–19665, 2026. doi: 10.1609/aaai.v40i24.39046.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, et al. Qwen3-VL Technical Report, 2025. URL `https://arxiv.org/abs/2511.21631`.

Black Forest Labs. FLUX.2 [klein] 9B. Model card, 2026. URL https://huggingface.co/black-forest-labs/FLUX.2-klein-9B.

Manuel Brack, Felix Friedrich, Dominik Hintersdorf, Lukas Struppek, Patrick Schramowski, and Kristian Kersting. SEGA: Instructing text-to-image models using semantic guidance. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-1102.

Yue Dai, Ziyang Liu, Marc Cheong, and Caren Han. IMPLICIT-Bench: Measuring implicit bias in text-to-image models under neutral prompts, 2026. URL https://arxiv.org/abs/2609.24228.

Moreno D'Incà, 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*, 47(11):9863–9875, 2025. doi: 10.1109/TPAMI.2025.3592901.

Felix Friedrich, Manuel Brack, Lukas Struppek, Dominik Hintersdorf, Patrick Schramowski, Sasha Luccioni, and Kristian Kersting. Auditing and instructing text-to-image generation models on fairness. *AI and Ethics*, 5:2103–2123, 2025. doi: 10.1007/s43681-024-00531-5.

Ada Görgün, Fawaz Sammani, Nikos Deligiannis, Bernt Schiele, and Jonas Fischer. Temporal concept dynamics in diffusion models via prompt-conditioned interventions. In *International Conference on Learning Representations*, 2026.

Boyu Han, Qianqian Xu, Shilong Bao, Zhiyong Yang, Kangli Zi, and Qingming Huang. LightFair: Towards an efficient alternative for fair T2I diffusion via debiasing pre-trained text encoders. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-0765.

Peter Hase, Mohit Bansal, Been Kim, and Asma Ghandeharioun. Does localization inform editing? surprising differences in causality-based localization vs. knowledge editing in language models. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-0774.

Thomas Klassert, Adrian Ulges, and Biying Fu. BAFIS: Dataset + framework to assess occupational bias and human preference in modern text-to-image models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2168–2177, 2026. doi: 10.1109/wacv61042.2026.00215.

Alexandra Sasha Luccioni, Christopher Akiki, Margaret Mitchell, and Yacine Jernite. Stable bias: Evaluating societal representations in diffusion models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 56338–56351, 2023. doi: 10.52202/075280-2458.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, 2022. doi: 10.52202/068431-1262.

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. ACM, 2023. doi: 10.1145/3600211.3604711.

OpenAI. GPT-5.6 System Card, July 2026. URL https://deploymentsafety.openai.com/gpt-5-6.

Hadas Orgad, Bahjat Kawar, and Yonatan Belinkov. Editing implicit assumptions in text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7053–7061, 2023. doi: 10.1109/iccv51070.2023.00649.

Jeonghoon Park, Juyoung Lee, Chaeyeon Chung, Jaeseong Lee, Jaegul Choo, and Jindong Gu. Fair generation without unfair distortions: Debiasing text-to-image generation with entanglement-free attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17567–17576, 2025. doi: 10.1109/iccv51701.2025.01632.

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*, pp. 6367–6384. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.naacl-long.353.

Yingdong Shi, Changming Li, Yifan Wang, Yongxiang Zhao, Anqi Pang, Sibei Yang, Jingyi Yu, and Kan Ren. Dissecting and mitigating diffusion bias via mechanistic interpretability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8192–8202, 2025. doi: 10.1109/cvpr52734.2025.00767.

Stability AI. Stable Diffusion 3.5 Medium. Model card, 2024. URL `https://huggingface.co/stabilityai/stable-diffusion-3.5-medium`.

Viacheslav Surkov, Chris Wendler, Antonio Mari, Mikhail Terekhov, Justin Deschenaux, Robert West, Caglar Gulcehre, and David Bau. One-step is enough: Sparse autoencoders for text-to-image diffusion models. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-3342.

Christopher Teo, Milad Abdollahzadeh, and Ngai-Man Cheung. On measuring fairness in generative models. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-0467.

Berk Tinaz, Zalan Fabian, and Mahdi Soltanolkotabi. Emergence and evolution of interpretable concepts in diffusion models. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-5566.

Michael Toker, Hadas Orgad, Mor Ventura, Dana Arad, and Yonatan Belinkov. Diffusion lens: Interpreting text encoders in text-to-image pipelines. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pp. 9713–9728. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.524.

Wenxuan Wang, Haonan Bai, Jen-tse Huang, Yuxuan Wan, Youliang Yuan, Haoyi Qiu, Nanyun Peng, and Michael R. Lyu. New job, new gender? measuring the social bias in image generation models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 3781–3789. ACM, 2024. doi: 10.1145/3664647.3681433.

Chenfei Wu, Jiahao Li, Jingren Zhou, et al. Qwen-Image Technical Report, 2025a. URL `https://doi.org/10.48550/arXiv.2508.02324`.

Zhengxuan Wu, Aryaman Arora, Atticus Geiger, Zheng Wang, Jing Huang, Dan Jurafsky, Christopher D. Manning, and Christopher Potts. AxBench: Steering LLMs? even simple baselines outperform sparse autoencoders. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 67035–67080. PMLR, 2025b.

Katarzyna Zaleska, Łukasz Popek, Monika Wysoczańska, and Kamil Deja. Attention, may i have your decision? localizing generative choices in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 30854–30863, 2026.

Cheng Zhang, Xuanbai Chen, Siqi Chai, Chen Henry Wu, Dmitry Lagun, Thabo Beeler, and Fernando De la Torre. ITI-GEN: Inclusive text-to-image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3969–3980, 2023. doi: 10.1109/iccv51070.2023.00367.

Yanzhe Zhang, Lu Jiang, Greg Turk, and Diyi Yang. Auditing gender presentation differences in text-to-image models. In *Proceedings of the 2024 ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pp. 1–10. ACM, 2024. doi: 10.1145/3689904.3694710.

A Implementation and evaluation details

A.1 Case counts and seed assignments

We split the IMPLICIT-Bench prompt triplets into a development set of 513 and a test set of 1,314. For each model, 498 development cases and 1,295 test cases have valid edits. We fit predictors on the valid development cases. After text normalization, no ordered prompt triplet appears in both sets.

Clean generation, semantic steering, and random edits of the same size use random seeds 3 and 4. Clean images from independent seeds 0 and 1 provide the baseline's headroom and prompt-matching features. Seeds 6, 7, and 8 are used for denoiser probes. Including baseline images for cases without valid edits, the evaluation across three models contains 31,194 images and 62,388 individual judge scores. We analyze results at the prompt level.

A.2 Constructing the text-encoder output edit

Let a and b be the stereotype and counterfactual token-representation matrices. We compare rows up to the shorter sequence length. A row counts as changed when its difference norm exceeds $\max(10^{-3}, 0.05 \max_k \|b_k - a_k\|_2)$. We average the differences for the first two such rows, or use the single row if there is only one. Subtraction uses the representation's original data type before conversion to float. We do not normalize this semantic direction to unit length.

We choose which rows to edit by trying candidate words in the neutral prompt. We first check each case's `head` and `target` fields for words longer than two characters. We try up to two such words from each field. We then try a fixed list: person, man, woman, people, girl, boy, child, worker. For each candidate in the neutral prompt, we replace its first occurrence with "thing" and encode the prompt again. We stop at the first candidate that changes any rows under the threshold above. We select up to two of those rows for editing. We add the direction multiplied by the model's steering strength at these rows, as in Eq. 1. An edit is invalid if the two prompt representations are identical or the averaged direction is zero. It is also invalid if no rows are selected or applying the edit produces no change.

Table 8: Fixed generator settings. Probe indices are zero-based forward indices. Strengths multiply the original direction vectors, so edit sizes differ across models.

Model	Steps	Strength	Probe indices	Guidance
SD3.5	28	2	6, 10	CFG 4
FLUX	8	1	2, 3	Native distilled
Qwen-Image	50	2	10, 16	Normalized CFG 4

All outcome images are 1024×1024. SD3.5 uses float16, while FLUX and Qwen-Image use bfloat16. We run neutral generation with each model's standard settings to reach the probe steps. In Eq. 2, v_θ is the conditional transformer output before guidance is applied. SD3.5 keeps the neutral pooled projection fixed when comparing prompt token representations. FLUX allows text tensor shapes to vary with the prompt, while Qwen-Image uses the text mask for each prompt.

A.3 Calibration, predictors, and uncertainty

For each development case, we sample four random directions of the same length as the semantic direction and apply them with the same rule for selecting rows and the same steering strength. We use their six-probe mean cosines to compute the random-edit baseline. Probe seeds are 6, 7, and 8, with the two model-specific forward indices listed above. We assign no response score when the edit is invalid or zero, or when cosine similarity is undefined. Standardization subtracts a fixed mean and divides by a positive standard deviation within each model, leaving the ranking and selected prompts unchanged. The sample standard deviation uses $\mathrm{ddof} = 1$. We use these development-set constants unchanged on the test set to compute G as defined in Eq. 3.

Each learned predictor uses standardized features and ridge regression with penalty 1 and an SVD solver. We fit the scaling and coefficients on 498 valid development cases per model. The baseline

uses three text features and two image features. The text features are the input direction length, the distance between the prompt representations over their common rows, and the relative edit length. We compute relative edit length by dividing the edit length by the length of the neutral representation. The image features are the headroom H and the prompt-matching rate, both from independently generated clean images. Adding G or M changes only the features. The image features require decoded and judged clean images. The text-only baseline uses the same fitting procedure with just the three text features. We compare its test correlation with R against that of G using paired bootstrap.

Each selection method takes the first $\lceil qN \rceil$ ranked cases for $q \in \{0.10, 0.20, 0.25, 0.50, 1.00\}$, breaking ties by case key. For the top 20%, we report stereotype reduction and prompt matching separately, as well as the fraction of prompts for which both steered images match the prompt and the mean stereotype rating decreases. At every selection fraction, choosing prompts uniformly at random gives the full-set mean improvement in expectation.

The main and selection analyses each use 10,000 paired prompt-bootstrap resamples, with selection intervals reported at each fraction. Each resampled prompt keeps all its scores and its two-seed mean outcome, allowing paired method comparisons. These intervals describe uncertainty from resampling prompts while holding the observed outcomes, fitted predictors, calibration, and judge fixed. MSE reductions use squared score units, while mean improvements use score points.

A.4 Image judging

We use the stereotype-rating and prompt-matching prompts from IMPLICIT-Bench for both judges. The stereotype-rating question asks for an integer from 0 (the specified stereotype is absent) to 5 (an extreme expression of it). The prompt-matching question asks whether the image shows the main subjects, action, setting, and specified attributes of the neutral prompt, ignoring artistic style and minor visual flaws. Qwen3-VL returns at most eight tokens using greedy decoding, with one retry for invalid formatting. We do not fill in missing labels. We chose the combined score and its components from development results.

A.5 Prediction reliability and baseline comparisons

Reliability and corrected correlation. Let R_3 and R_4 be the per-seed improvements whose mean is R, and let r_{34} be their Spearman correlation. We estimate the reliability of R using the Spearman–Brown formula, $r_{RR} = 2r_{34}/(1 + r_{34})$. We estimate the reliability of G by splitting its six probe measurements into two groups (200 random three-versus-three splits, Spearman–Brown corrected). Standardization does not change this estimate. The reliability-corrected correlation is $\rho^* = \rho_{\text{obs}}/\sqrt{r_{RR}\, r_{GG}}$. We recompute all three terms in every bootstrap resample to obtain its interval.

Image-pair baseline and partial correlation. The image-pair comparison uses the Spearman correlation between R_3 and R_4. Each R_s measures improvement from one clean and one steered image generated with seed s. We evaluate this baseline and G against the same single-seed outcome. We also compute the partial Spearman correlation of G with R_4 given R_3. We first regress the ranks of G and R_4 separately on the ranks of R_3. We then correlate the two sets of residuals.

A.6 Threshold selection

The selection rule steers prompt i when its score is at least τ and returns the clean image otherwise. Coverage is the fraction of prompts steered, and risk is the fraction of steered prompts with $R < 0$. The area under the risk–coverage curve (AURC) averages risk over the coverage values k/n, $k = 1, \ldots, n$. For random selection, the expected risk is constant at $P(R < 0)$, which we use as a reference. Intervals come from 5,000 paired case bootstrap resamples.

B ADDITIONAL EXPERIMENTAL RESULTS

B.1 PREDICTION AND JUDGE COMPARISONS

Table 9 gives paired intervals for the advantage of G over M in correlation and top-20% selection. Table 10 reports prediction reliability and the text-feature comparison using the stored per-seed image scores and per-probe cosines, with 10,000 paired prompt-bootstrap resamples. Both tables use all 1,295 test cases per model. Prediction differences compare Spearman correlations with R, while selection differences compare mean improvement among the top 20%. In Table 10, r_{34} is the correlation between improvements measured with seeds 3 and 4, r_{RR} and r_{GG} are the reliabilities of R and G, and ρ^* is the reliability-corrected correlation. B_{text} uses text features only.

Table 9: Advantages of G over M in prediction and top-20% selection. Brackets show 95% paired bootstrap intervals.

Model	Prediction: correlation difference	Selection: improvement difference
SD3.5	0.106 [0.047, 0.163]	4.75 [1.78, 8.11]
FLUX	0.152 [0.086, 0.218]	11.31 [6.72, 16.91]
Qwen-Image	0.255 [0.190, 0.321]	20.69 [15.44, 27.14]

Table 10: Prediction reliability and the text baseline. Brackets show 95% paired bootstrap intervals.

	SD3.5	FLUX	Qwen-Image
$\rho(G, R)$, two seeds	**0.190**	**0.211**	**0.240**
$\rho(G, R_s)$, one seed (mean over s)	0.154	0.182	0.189
r_{34}	0.166	0.318	0.350
r_{RR}	0.284 [0.184, 0.373]	0.483 [0.406, 0.550]	0.518 [0.447, 0.581]
r_{GG}	0.914 [0.907, 0.925]	0.936 [0.925, 0.939]	0.912 [0.898, 0.920]
ρ^*	0.374 [0.263, 0.506]	0.314 [0.232, 0.398]	0.349 [0.270, 0.431]
$\rho(B_{\text{text}}, R)$	0.092	0.161	0.103
$\rho(G, R) - \rho(B_{\text{text}}, R)$	0.098 [0.035, 0.161]	0.051 [−0.019, 0.121]	0.137 [0.068, 0.203]

Steering gains on the full test set Steering improves the combined score over both clean generation and random edits, with paired intervals above zero in Table 11. The corresponding mean stereotype ratings and prompt-matching rates appear in Table 1.

Table 11: Mean differences in observed improvement in score points. Brackets give 95% paired prompt-bootstrap intervals.

Model	Steering − clean	Steering − random
SD3.5	5.65 [4.46, 6.86]	5.86 [4.68, 7.06]
FLUX	6.39 [4.94, 7.86]	6.14 [4.69, 7.64]
Qwen-Image	6.86 [5.13, 8.63]	7.39 [5.68, 9.15]

Second judge We used GPT-5.6, with model ID `gpt-5.6-terra`, to re-score all 7,770 outcome images per model. These include clean, steered, and random-edit images from both random seeds. We used the exact stereotype-rating and prompt-matching questions from Appendix A, with one image and one question per request, high image detail, and reasoning disabled. Every response had a valid format. We recomputed U and R from the new ratings using Eq. 4 and compared them with Qwen3-VL's scores on the same 1,295 cases per model, as shown in Table 12. The text baseline is the predictor fitted on development cases in Section 4.1 and is evaluated against the GPT-5.6 improvement without refitting. Agreement is the per-image Spearman correlation of stereotype ratings, and offset is the mean GPT-5.6 rating minus the Qwen3-VL rating. Contrast is the decrease in stereotype rating after steering. The final row compares $\rho(G, R) - \rho(M, R)$ between judges.

Under GPT-5.6, steering raises the mean combined score by 4.9, 6.3, and 4.7 points over clean generation in SD3.5, FLUX, and Qwen-Image, respectively.

Table 12: Agreement and steering results under the two judges. Brackets show 95% paired bootstrap intervals.

	SD3.5	FLUX	Qwen-Image
Agreement ρ	0.684	0.745	0.695
Offset (0–5)	-0.43	-0.29	-0.38
Contrast, Qwen3-VL / GPT-5.6	0.52 / 0.47	0.69 / 0.70	1.13 / 1.07
Under GPT-5.6			
$\rho(M, R)$	0.167	0.054	0.009
$\rho(B_{\text{text}}, R)$	0.058 [0.00, 0.11]	0.186 [0.13, 0.24]	0.183 [0.13, 0.24]
$\rho(G, R)$	**0.174** [0.12, 0.23]	**0.190** [0.13, 0.25]	**0.239** [0.18, 0.29]
$\rho(G, R) - \rho(M, R)$	0.007 [-0.05, 0.06]	0.137 [0.07, 0.20]	0.230 [0.16, 0.30]
$\rho(G, R) - \rho(B_{\text{text}}, R)$	0.116 [0.05, 0.18]	0.004 [-0.06, 0.07]	0.056 [-0.01, 0.12]
Top-20% mean R, by G / by M	**17.92** / 13.55	**20.12** / 5.79	**27.45** / 6.45
Top-20% gain $G - M$	4.36 [0.2, 8.3]	14.32 [8.4, 19.7]	21.00 [14.5, 28.0]
Between-judge difference	0.099 [0.03, 0.17]	0.015 [-0.06, 0.09]	0.025 [-0.04, 0.08]

B.2 Prompt Selection

Five groups along the ranking. We order prompts by each score and divide them into five equal groups of 259 cases. The highest-G group has the greatest mean improvement in every model. Table 13 shows that the highest-M group does not have the greatest improvement in FLUX or Qwen-Image.

Table 13: Mean improvement in five equal score-ranked groups (Qwen3-VL). Bold marks the highest group mean in each column.

	SD3.5		FLUX		Qwen-Image	
Group of prompts	G	M	G	M	G	M
lowest 20%	1.2	2.2	-2.1	2.6	-0.7	3.6
20–40%	2.7	3.9	3.3	4.1	-0.7	6.4
40–60%	2.3	3.9	4.7	**9.7**	0.9	**8.8**
60–80%	4.9	5.8	8.0	8.8	7.3	8.7
top 20%	**17.3**	**12.5**	**18.1**	6.8	**27.6**	6.9

All evaluated selection fractions Table 14 reports all tested selection fractions. H ranks prompts only by headroom in independently generated clean images. B and $B+G$ rank by improvement predicted by models fitted on development cases. These predictors and H are defined in Appendix B.4. Entries give mean improvement per selected prompt at the same selection fraction. Figure 2 shows the intervals for G and M.

Stereotype reduction and prompt matching. In the top 20% selected by G, both steered images follow the prompt for 156/259, 162/259, and 131/259 cases in SD3.5, FLUX, and Qwen-Image. The corresponding counts under M are 128/259, 122/259, and 97/259. We check prompt matching after generation and report selection gains for the entire selected group. Table 15 reports the fraction of prompts that satisfy this matching condition and have a lower mean stereotype rating than their clean images.

Selective steering The threshold rule and its uncertainty calculation are defined in Appendix A. Figure 5 shows the share of selected prompts with negative improvement under Qwen3-VL. This share stays at or below that of selection by M across the full range. Table 16 summarizes the curves for the same 1,295 cases per model under both judges, including the clean-image predictor defined in Appendix B.4. All steers every prompt, while Best ranks by observed improvement and gives

Table 14: Mean improvement per selected prompt at each selection fraction.

Model	Top (%)	N	G	M	H	B	$B+G$
SD3.5	10	130	22.15	14.77	3.85	16.31	**24.46**
	20	259	17.26	12.51	3.20	13.40	**19.00**
	25	324	15.46	12.10	3.06	13.27	**16.02**
	50	648	9.77	8.02	3.70	9.20	**9.91**
	100	1295	**5.65**	**5.65**	**5.65**	**5.65**	**5.65**
FLUX	10	130	25.69	1.85	8.31	**31.15**	28.38
	20	259	18.11	6.80	7.34	**20.77**	19.31
	25	324	16.14	7.31	6.11	16.88	**17.41**
	50	648	11.19	8.09	8.52	11.30	**11.88**
	100	1295	**6.39**	**6.39**	**6.39**	**6.39**	**6.39**
Qwen-Image	10	130	40.00	-0.38	12.46	32.00	**42.92**
	20	259	27.61	6.91	12.32	24.44	**30.81**
	25	324	23.24	7.04	12.19	21.76	**26.11**
	50	648	14.55	7.89	11.23	13.98	**14.68**
	100	1295	**6.86**	**6.86**	**6.86**	**6.86**	**6.86**

Table 15: Top-20% prompts with lower stereotype ratings and both steered images matching the prompt. Brackets show 95% paired bootstrap intervals.

Model	Selected by G (%)	Selected by M (%)	Difference (pp)
SD3.5	**35.91**	25.87	10.04 [5.02, 15.83]
FLUX	**39.38**	30.12	9.27 [3.09, 16.99]
Qwen-Image	**40.15**	25.48	14.67 [8.49, 21.24]

the lowest achievable risk. Figure 3 in the main text separates stereotype reduction from prompt-matching loss over the full test set. At 20% coverage under Qwen3-VL, the fraction with negative improvement is 11.5% for G-selected SD3.5 prompts, compared with 16.2% across all prompts. The corresponding comparison is 17.6% versus 17.9% in FLUX and 18.8% versus 25.6% in Qwen-Image. Selection reduces the frequency of negative outcomes in SD3.5 and Qwen-Image, while the FLUX frequency changes little.

Table 16: Area under the risk–coverage curve (lower is better) and paired differences for G. Bold compares G, M, B, and All; Best is the outcome-based lower bound.

		AURC					Difference for G	
Model	Judge	G	M	B	All	Best	vs. M	vs. All
SD3.5	Qwen3-VL	**0.141**	0.182	0.145	0.162	0.014	−0.041 [−0.061, −0.020]	−0.021 [−0.037, −0.004]
	GPT-5.6	**0.184**	0.218	0.210	0.223	0.027	−0.034 [−0.055, −0.013]	−0.039 [−0.057, −0.020]
FLUX	Qwen3-VL	0.157	0.240	**0.133**	0.179	0.017	−0.082 [−0.110, −0.054]	−0.022 [−0.039, −0.003]
	GPT-5.6	0.195	0.262	**0.162**	0.216	0.025	−0.066 [−0.096, −0.036]	−0.021 [−0.041, 0.000]
Qwen-Image	Qwen3-VL	0.214	0.336	**0.185**	0.256	0.036	−0.122 [−0.152, −0.093]	−0.042 [−0.061, −0.022]
	GPT-5.6	0.242	0.352	**0.219**	0.309	0.054	−0.110 [−0.139, −0.080]	−0.067 [−0.086, −0.047]

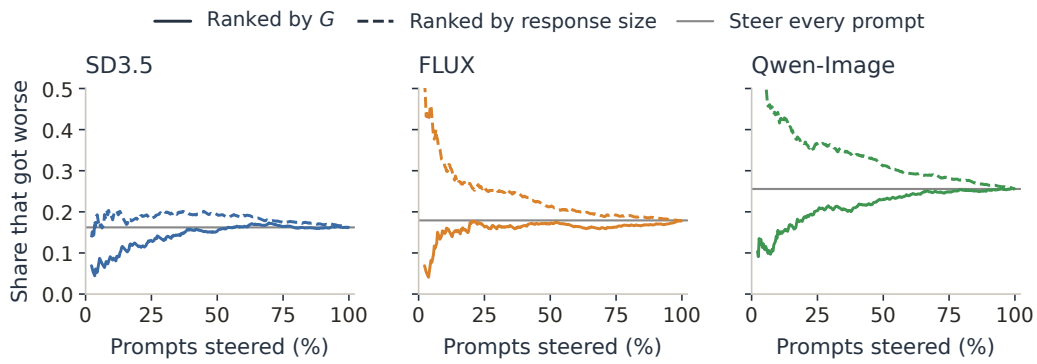

Figure 5: Fraction of selected prompts with negative improvement (Qwen3-VL). Horizontal lines show steering every prompt.

B.3 Efficient Probing and Reuse Across Seeds

Number of probes. For each number of probes, we take all subsets of the six stored probes with the required number of seeds and steps. We average the cosines within each subset and compute its Spearman correlation with R. Table 17 reports the mean correlation across subsets, divided by the correlation from all six measurements. The main configuration uses three seeds and two steps.

We count forward calls from the run records, including one neutral generation run per probe seed to reach the probe steps. Each probe measurement then adds four conditional forward calls. For SD3.5, a neutral run uses 28 calls and a complete image also takes 28. The corresponding counts are 8 and 8 for FLUX. Qwen-Image uses two calls per guidance step: its neutral probe run stops after 34 calls, whereas a complete image takes 100. A clean–steered pair requires two complete generations, two image decodes, and four judge calls. Table 17 gives the resulting totals.

The one-probe and six-probe cost ratios are 0.57 and 1.93 in SD3.5, 0.75 and 3.00 in FLUX, and 0.19 and 0.63 in Qwen-Image.

Table 17: Forward calls per prompt and correlation retained relative to six probes.

	Forward calls			Retained correlation (%)		
Seeds × steps	SD3.5	FLUX	Qwen-Image	SD3.5	FLUX	Qwen-Image
1 × 1	**32**	**12**	**38**	90.8	87.7	90.7
1 × 2	36	16	42	94.6	92.4	94.8
2 × 1	64	24	76	95.8	94.9	96.2
2 × 2	72	32	84	98.1	98.0	98.5
3 × 1	96	36	114	97.9	98.0	98.2
3 × 2	108	48	126	**100**	**100**	**100**
One image pair (clean, steered)	56	16	200	+2 decodes, 4 judge calls		

Selection with one probe. We rank the same 1,295 test prompts using each of the six single probes and select the top 259. Each probe is one combination of seeds 6, 7, and 8 with the two model-specific steps in Appendix A. We compare its cosine score with response size measured at that same probe. Table 19 gives the mean improvement for every configuration under Qwen3-VL. The single-probe average is the mean of these six separately selected groups.

Table 18 reports this average and its paired differences from six-probe G and single-probe M. We use 10,000 paired prompt-bootstrap draws with seed 20260913, reselecting the top 259 for every score in each draw. We average the six configuration estimates within each draw before computing the intervals. Mean improvement ranges from 14.48 to 17.18 across configurations in SD3.5, from 13.82 to 17.45 in FLUX, and from 22.47 to 27.68 in Qwen-Image. These ranges describe variation across configurations, whereas the intervals below describe prompt-resampling uncertainty.

Table 18: Mean improvement from single-probe selection and paired differences. Brackets give 95% prompt-bootstrap intervals.

Model	One-probe G	Difference from six-probe G	Difference from one-probe M
SD3.5	15.84 [12.70, 19.16]	−1.42 [−2.82, −0.13]	4.32 [1.78, 6.65]
FLUX	16.08 [12.36, 19.94]	−2.03 [−4.38, −0.64]	9.25 [5.66, 12.88]
Qwen-Image	25.65 [20.59, 30.64]	−1.96 [−4.03, −0.54]	19.89 [15.39, 24.56]

Table 19: Top-20% mean improvement for each single probe. Brackets give 95% prompt-bootstrap intervals.

Model	Probe seed	Step	Selected by direction	Selected by response size
SD3.5	6	6	15.06 [11.70, 18.57]	11.24 [7.64, 15.14]
	6	10	16.49 [12.90, 20.54]	11.04 [7.49, 14.94]
	7	6	14.48 [11.31, 18.30]	11.16 [7.64, 15.17]
	7	10	16.76 [12.97, 20.54]	10.93 [7.49, 15.17]
	8	6	15.10 [11.58, 18.76]	12.90 [9.23, 16.80]
	8	10	17.18 [13.55, 20.73]	11.85 [8.15, 15.91]
FLUX	6	2	17.34 [13.24, 21.70]	8.53 [3.86, 13.05]
	6	3	15.29 [10.93, 19.34]	6.53 [2.01, 10.97]
	7	2	16.76 [12.20, 21.00]	8.80 [4.59, 13.09]
	7	3	15.83 [11.51, 20.62]	6.29 [1.89, 10.35]
	8	2	17.45 [13.59, 22.05]	6.99 [2.86, 11.62]
	8	3	13.82 [9.58, 17.92]	3.82 [0.00, 8.57]
Qwen-Image	6	10	27.49 [21.85, 32.97]	5.98 [1.24, 11.27]
	6	16	25.75 [20.54, 31.24]	5.41 [0.15, 10.12]
	7	10	27.68 [21.39, 32.90]	9.88 [4.56, 14.86]
	7	16	25.75 [20.35, 31.47]	5.98 [0.93, 11.12]
	8	10	24.75 [19.07, 29.96]	3.44 [−1.24, 8.46]
	8	16	22.47 [17.45, 27.99]	3.86 [−0.77, 8.53]

Single-probe selection reduces stereotypes but still incurs prompt-matching losses. Averaged over the six configurations, SD3.5 has a stereotype reduction of 1.38 rating points and a matching loss of 4.63 percentage points. FLUX has a reduction of 1.48 and a loss of 10.26. Qwen-Image has the largest reduction, 2.71, and the largest loss, 23.46.

The main-text retention rates compare the selected groups' total mean improvements: single-probe mean R divided by six-probe mean R. Rounded to whole percentages, these are 92% for SD3.5, 89% for FLUX, and 93% for Qwen-Image. A second comparison subtracts the random-selection gain from both means before taking their ratio. On this basis, one probe retains 87.8% in SD3.5, 82.7% in FLUX, and 90.6% in Qwen-Image. The single-probe gains above random selection are 10.19 points for SD3.5, 9.69 for FLUX, and 18.79 for Qwen-Image.

Additional generation seeds. We randomly selected 300 test cases per model, with equal sampling probability across the 1,295 matched cases. We generated clean and steered images using six new random seeds (10–15) and the fixed edit from Appendix A. We scored them with the same judge and prompts. Together with seeds 3 and 4, this gives eight improvements per case.

Table 20 reports $\rho(G, R)$ when improvement is averaged over k seeds. We average these correlations over all seed subsets of size k when there are at most 40. Otherwise, we use that many random subsets. The table also reports the reliability of the k-seed mean, estimated using four-versus-four split-half correlation and the Spearman–Brown formula. Finally, we report the corrected correlation ρ^* from eight seeds, with a 95% paired prompt-bootstrap interval.

Table 20: G's correlation with improvement as more generation seeds are used; 300 cases per model.

	$\rho(G, R)$ by seeds averaged				Reliability of R			
Model	1	2	4	**8**	1 seed	2 seeds	**8** seeds	ρ^* (8 seeds)
SD3.5	0.151	0.176	0.202	**0.218**	0.10	0.18	**0.47**	0.33 [0.15, 0.52]
FLUX	0.122	0.152	0.179	**0.203**	0.20	0.33	**0.67**	0.26 [0.11, 0.40]
Qwen-Image	0.185	0.215	0.242	**0.245**	0.22	0.36	**0.70**	0.31 [0.16, 0.45]

Selection with new generation seeds. We fixed the top-20% cutoffs on the full test set before evaluating the 300-case subsets on six new random seeds. The selected group retains a higher mean gain than the full subset in every model. In SD3.5, selected prompts gain 14.4 points on the original seeds and 14.4 on the new seeds, compared with 4.1 points when steering the full subset on the new seeds. In FLUX, the selected group's gain changes from 14.8 to 13.5 points, compared with 4.7 for the full subset on new seeds. In Qwen-Image, it changes from 23.3 to 21.9 points, compared with 6.3 for the full subset on new seeds.

B.4 Understanding What G Predicts

Prompt types. Table 21 groups test pairs by the words that differ between their stereotype and counterfactual prompts. A pair is a *subject swap* when all differing words are gendered subject words (man/woman, boy/girl, male/female, he/she, and similar). Other pairs differ in a trait, activity, or setting associated with the subject. The groups are the same for all three models because they depend only on the prompts.

On the full test set, subject-swap pairs make up 30–36% of the top-20% prompts selected by G and 6–19% of those selected by M. The other prompts selected by G improve by 11.2, 8.1, and 15.1 points in SD3.5, FLUX, and Qwen-Image, respectively, two to four times their group mean. Among the bottom 20% by G, the mean stereotype reduction is 0.10–0.33 rating points and the matching loss is 1–9 percentage points, so low G mostly identifies prompts for which steering has little effect.

The component correlations in Table 4 use all 1,295 cases per model, with outcomes averaged over the two seeds. For Figure 4, we selected three prompts from each of the top and bottom 5% of G in Qwen-Image. Both images in each example pass the prompt-matching check, so their stereotype ratings are comparable.

Table 22 repeats the top-20% comparison within each group. This removes any advantage from placing subject-swap pairs above other pairs. Among the 125 subject-swap pairs, each method

Table 21: Mean scores and correlations within each prompt-pair type.

	Subject swap ($n = 125$)			Other contrast ($n = 1{,}170$)		
Model	mean G	mean R	$\rho(G, R)$	mean G	mean R	$\rho(G, R)$
SD3.5	3.89	24.3	0.31	1.47	3.7	0.13
FLUX	4.98	33.0	0.36	2.05	3.6	0.13
Qwen-Image	4.29	38.6	0.49	1.16	3.5	0.14

selects 25 prompts, with similar mean improvement for G and M. Among the 1,170 other pairs, each method selects 234 prompts. G gives an improvement above the group mean in all three models, with intervals above zero. For M, the interval is above zero only in SD3.5. Intervals use 10,000 paired prompt-bootstrap resamples within each group.

Table 22: Mean improvement from top-20% selection within each pair type.

		Mean R			Difference from group mean		
Group	Model	by G	by M	group	G	M	$G - M$
Subject swap	SD3.5	**47.6**	47.2	24.3	23.3 [7.4, 31.8]	22.9 [8.4, 36.2]	0.4 [−17.6, 11.2]
($n = 125$)	FLUX	**54.4**	46.4	33.0	21.4 [6.7, 32.5]	13.4 [−3.0, 27.7]	8.0 [−10.0, 24.8]
	Qwen-Image	**58.4**	57.2	38.6	19.8 [4.0, 32.2]	18.6 [2.2, 31.0]	1.2 [−12.0, 15.6]
Other contrast	SD3.5	**9.8**	6.6	3.7	6.1 [3.3, 8.9]	3.0 [0.4, 6.3]	3.2 [−0.6, 6.0]
($n = 1{,}170$)	FLUX	**7.3**	4.1	3.6	3.7 [0.8, 7.2]	0.6 [−3.0, 4.4]	3.1 [−1.0, 7.6]
	Qwen-Image	**12.0**	1.9	3.5	8.5 [4.8, 12.7]	−1.6 [−5.9, 2.2]	10.1 [5.3, 15.9]

Additional image-based predictors. Prompts with low scores before steering have more room to improve. To test whether G provides information beyond this headroom, we define $H = 100 - \bar{U}_{\text{ind}}^{\text{clean}}$, where $\bar{U}_{\text{ind}}^{\text{clean}}$ is the mean score of two images generated from the neutral prompt without steering. These images use seeds independent of those used to measure R. We train a linear model B on development cases to predict R from H, the prompt-matching rate of the same two images, and the three text features in Section 4.1. We then train another model with G as an additional input, denoted $B + G$.

Table 23 compares their prediction errors and the gains obtained by steering the top 20% of prompts ranked by each predictor. Adding G reduces test mean squared error (MSE) in all three models. It also selects prompts with larger gains in SD3.5 and Qwen-Image, while in FLUX, the error decreases but the selected group's gain falls. Adding M instead gives smaller error reductions in SD3.5 and FLUX and increases error in Qwen-Image, as detailed in Table 24.

Table 23: Adding G to the clean-image baseline. †: the paired 95% interval is above zero.

	Prediction: MSE ↓			Selection: mean R ↑		
Model	B	$B + G$	Reduction (%)	B	$B + G$	Gain
SD3.5	474.11	**468.59**	1.17	13.40	**19.00**	$5.60^{\dagger}$
FLUX	700.04	**660.72**	$5.62^{\dagger}$	**20.77**	19.31	−1.47
Qwen-Image	983.92	**900.06**	$8.52^{\dagger}$	24.44	**30.81**	$6.37^{\dagger}$

Predictors fitted on development cases Table 24 compares four feature sets using the same fitting procedure and development cases, evaluated on all 1,295 test cases per model. $B + G$ has the lowest test MSE in FLUX and Qwen-Image, whereas $B + G + M$ has the lowest in SD3.5.

Uncertainty in added predictive value Adding G to B reduces prediction error in FLUX and Qwen-Image, while the SD3.5 interval includes zero, as shown in Table 25. Positive values favor adding G, and absolute reductions use squared score units.

Table 24: Test MSE for four predictors fitted on development cases.

Predictor	SD3.5	FLUX	Qwen-Image
B	474.11	700.04	983.92
$B + G$	468.59	**660.72**	**900.06**
$B + M$	473.06	696.57	1088.66
$B + G + M$	**465.67**	662.44	938.01

Table 25: Reduction in test MSE from B to $B + G$. Brackets give 95% paired prompt-bootstrap intervals for the absolute reduction.

Model	Relative reduction (%)	Absolute reduction
SD3.5	1.17	5.524 [−15.871, 27.454]
FLUX	5.62	39.316 [17.496, 61.779]
Qwen-Image	8.52	83.865 [45.121, 122.199]

Selecting by $B + G$ instead of B changes mean improvement by 5.60 [2.36, 7.72] points in SD3.5, -1.47 [-5.10, 2.39] in FLUX, and 6.37 [3.28, 10.04] in Qwen-Image. Table 14 gives the corresponding means across selection fractions. Table 3 in the main text reports the gains and component changes for selection by G and M.

Correlations across seeds. The image-pair comparison measures the Spearman correlation between improvements on the two random seeds, evaluating each seed against the other on all 1,295 test cases per model. Table 26 also reports the partial correlations with G, computed as described in Appendix A.

Table 26: Cross-seed improvement correlations and partial correlations with G. Brackets show 95% paired bootstrap intervals.

	SD3.5	FLUX	Qwen-Image
$\rho(R_3, R_4)$, image-pair baseline	**0.166**	**0.318**	**0.350**
$\rho(G, R_4)$	0.142	0.199	0.189
Partial $\rho(G, R_4 \mid R_3)$	0.118 [0.062, 0.172]	0.156 [0.101, 0.210]	0.133 [0.077, 0.189]
Partial $\rho(G, R_3 \mid R_4)$	0.147 [0.091, 0.201]	0.110 [0.054, 0.167]	0.135 [0.080, 0.188]

B.5 Steering and Probe Settings

Steering strength For the 300-case subset described in Appendix B.3, we generated steered images at each strength with random seeds 3 and 4, and recomputed G using probes at that strength. We reused clean images and steered images at the main strength from the main evaluation. We did not vary the strength for Qwen-Image.

Probe design and numerical agreement Table 28 lists each single-step result by sampler-step index. On the same 300-case subset, we varied the probe timesteps, the reference direction, and how each probe response was scored. We compare variants within the same run because recomputing a probe as part of a different set of steps can change its numerical result. The new values match the stored main-evaluation probes exactly for all 300 Qwen-Image cases, whose steps were unchanged. They match for 59 of 300 SD3.5 cases and 16 of 300 FLUX cases, where we expanded the set to six steps.

The per-probe correlations between runs are $\rho = 0.77$, 0.76, and 1.00. After averaging six probes, the correlations between the resulting G scores are $\rho = 0.91$, 0.93, and 1.00, and $\rho(G, R)$ changes by at most 0.03. The stored probes reproduce the reported G exactly in all three models. Running the same code on a different accelerator changed cosine similarity by up to 0.46. Because Δv subtracts two nearly identical velocity fields, it is sensitive to hardware and the order of numerical operations.

Table 27: Mean improvement, its two components, and $\rho(G, R)$ at four steering strengths; 300 cases per model. Bold marks the highest value within each model.

Model	α	Mean R	95%	Stereotype red.	Matching (pp)	$\rho(G, R)$	Diff. vs. main
SD3.5	0.5	1.07	[−0.80, 2.93]	0.052	**0.8**	0.078	−0.066 [−0.209, 0.073]
	1	1.97	[−0.13, 4.10]	0.185	−1.3	**0.154**	0.010 [−0.117, 0.136]
	2*	4.40	[2.03, 6.77]	0.538	−5.2	0.144	—
	4	**5.80**	[2.93, 8.77]	**0.658**	−5.5	0.104	−0.040 [−0.167, 0.088]
FLUX	0.5	3.20	[0.60, 5.77]	0.382	**−4.2**	0.165	−0.024 [−0.160, 0.106]
	1*	**6.80**	[3.90, 9.77]	0.867	−10.2	**0.190**	—
	2	3.00	[−0.87, 6.97]	1.553	−26.7	0.104	−0.085 [−0.198, 0.024]
	4	−0.47	[−4.60, 4.10]	**1.853**	−35.5	0.148	−0.041 [−0.169, 0.077]

*Strength used throughout the paper.

Table 28: Correlation with R under different probe settings (300 cases per model).

Variant	$\rho(G, R)$ for SD3.5 / FLUX / Qwen-Image
Main pair of steps	0.144 / 0.190 / 0.199
All probed steps averaged	0.163 / 0.218 / 0.199
Single step	2: **0.165** / 0: 0.177 / 10: 0.191 6: 0.157 / 1: **0.250** / 16: **0.204** 10: 0.125 / 2: 0.185 / – 14: 0.154 / 3: 0.178 / – 18: 0.156 / 5: 0.136 / – 22: 0.109 / 7: 0.027 / –
Reference $v(c^c) - v(c^n)$	0.123 / 0.129 / 0.174
Cosine weighted by response size	0.142 / 0.199 / 0.144
Response size alone	0.089 / 0.116 / −0.047

C CHOOSING THE EDIT LAYER PER PROMPT

The main paper keeps the edit location fixed. We also tested whether the same kind of denoiser probe could help choose, for each prompt, between editing the text-encoder output and an internal encoder layer. These separate experiments used a different image score. We compare this rule with editing the encoder output on development and held-out cases.

Procedure. For a candidate layer L with neutral state h_L^n and proposed edit Δh_L, the score $N_L = \mathbb{E}_q\langle \Delta v_{L,q}, u_q\rangle$ measures how far the denoiser change moves along a target direction u_q. Computing N_L costs one denoiser evaluation per candidate layer. We therefore first narrow down the candidates with a gradient-based approximation. Linearizing the remaining computation $f_{L,q}(h) = v_\theta(x_q, t_q; D_L(h))$ around h_L^n gives

$$S_L = \mathbb{E}_q\big[\langle J_{L,q}^{\mathsf{T}} u_q, \Delta h_L\rangle\big] \approx N_L, \qquad J_{L,q} = \left.\frac{\partial f_{L,q}(h)}{\partial h}\right|_{h=h_L^n}, \tag{5}$$

where backpropagation gives $J_{L,q}^{\mathsf{T}} u_q$ and a single backward pass through the encoder scores all layers. We keep the K layers with the largest S_L and add the encoder output as a candidate. We then compute N_L for these candidates and edit the layer with the largest value, using the output when scores tie. Neither step decodes or judges images. Internal-layer edits used centered reflection at strength 2 along a unit direction from the prompt pair, while the output baseline edited the full positional output. These experiments used the image score $J = \mathrm{mean}_s(Y_s H_s)$, where Y is a binary counterfactual choice and H checks whether the subject and scene are preserved. This differs from the continuous R in Eq. 4.

Outcome. On 235 Qwen-Image development cases, the rule with $K = 2$ reached $J = 0.338$, compared with 0.272 for output editing. A $K = 1$ rule chosen after inspecting these results reached 0.347. Choosing a layer by the gradient approximation alone, without checking N_L, reached 0.251.

We then fixed the $K = 1$ rule, its threshold, and its fallback, and evaluated it on 1,592 unique triplets outside the development set per model. The difference in J from output editing was $+0.001$ $[-0.009, +0.011]$ for Qwen-Image, $+0.007$ $[-0.000, +0.014]$ for FLUX, and -0.002 $[-0.006, +0.001]$ for SD3.5. All three intervals include zero among cases assigned an internal layer as well. The main experiments therefore keep the edit location fixed and evaluate prompt selection.