FAIR DIFFUSION MODELS VIA INITIAL-NOISE OPTIMIZATION IN H-SPACE

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion models can generate high-quality images, but it remains challenging to ensure fairness in the demographic composition of generated populations. Debiasing methods that optimize the initial noise before generation typically derives update directions without directly using demographic information encoded in the semantic features of denoiser, which potentially limits bias reduction or requires larger perturbations that compromise image fidelity. In this paper, we propose a new initial-noise optimization method in H-space that represents the semantic information of the denoiser. We define a batch-level fairness objective in H-space and propagate it back to the initial noise through the Jacobian, which leads to an update direction that reflects both the fairness objective and the model's feature response. Using a frozen terminal-time attribute probe, the proposed method performs prior-regularized stochastic updates on the initial noise and renders the optimized tensors with the original sampler. We provide a theoretical interpretation of this process by characterizing how initial-noise updates induce changes in H-space and establishing a gradient-alignment condition under which these updates decrease the image-space demographic distribution objective. Experiments on CelebA-HQ and FFHQ and multiple diffusion backbones, including P2, SD3, and SDXL, our method reduces fairness discrepancy by 35.8% while simultaneously improving FID and it consistently improves demographic distribution balance, demonstrating favorable fairness-fidelity trade-offs and generalization across datasets and model architectures.

1 INTRODUCTION

Diffusion models synthesize images by reversing a progressive noising process (Ho et al., 2020; Nichol & Dhariwal, 2021; Song et al., 2021a). Strong image fidelity does not by itself specify the composition of a generated dataset. When a collection is required to follow a prescribed demographic distribution, the objective concerns the aggregate attributes of its images. Fair generative modeling and distribution guidance address this population-level requirement (Choi et al., 2020; Parihar et al., 2024). We study it with a frozen diffusion model, binary perceived attributes, and a uniform target.

The central design choice is how an attribute objective changes generation. Internal features offer a semantic representation on which to define that objective (Kwon et al., 2023; Haas et al., 2023). Initial noise offers an input variable that can be optimized before invoking the original renderer (Wallace et al., 2023; Ben-Hamu et al., 2024). Connecting these two spaces requires more than relocating a loss: an input update changes the features according to the model's Jacobian. The resulting feature movement generally differs from directly stepping along the H-space gradient.

Our question is therefore how to translate a batch objective on H-space into an initial-noise update, and which properties of that update matter for the final image distribution. Answering it requires distinguishing the surrogate value, the direction obtained by backpropagation, and the output response measured after rendering. These quantities define the formulation, analysis, and experimental comparisons in this paper.

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.

The problem: translating feature objectives into input control. Distribution guidance estimates batch attributes from denoiser features and modifies those features within the sampling process (Parihar et al., 2024). This exposes a useful population objective, but its update is applied to an internal activation. In contrast, DOODL and D-Flow optimize initial inputs using gradients propagated from generated outputs (Wallace et al., 2023; Ben-Hamu et al., 2024). They provide direct access to output objectives through the generation path; DOODL also uses an invertible process to make backpropagation memory efficient. Neither comparison is resolved merely by stating that initial noise is optimizable. The issue here is what happens when the population objective is read from one internal feature map and its gradient is applied at the input.

Our formulation. We read a frozen attribute probe from the denoiser bottleneck at the terminal, highest-noise timestep. Its predictions define a batch distribution surrogate. For a stacked feature map $H(Z)$ and feature gradient g_H, the attribute direction in initial-noise space is $J_H^\top g_H$, where J_H is the input-to-feature Jacobian. The gradient is computed by a vector–Jacobian product. Prior-regularized stochastic updates optimize the input batch, and the ordinary deterministic sampler then renders it. The model weights and probe remain fixed throughout.

This choice has a concrete local interpretation. An input gradient step induces the H-space change $-\eta J_H J_H^\top g_H$ to first order. Thus model sensitivity weights the feature directions acted on by the attribute objective. The same step solves a linearized feature-loss problem with a quadratic penalty on input displacement. Recomputing the network from the updated input also recomputes its bottleneck and skip features together. These are structural properties of the procedure; whether they improve the distribution–fidelity tradeoff is a measurable question.

From representation to output response. A probe's classification accuracy measures its predictions, while control depends on the response to its gradient. We connect the two stages through a conditional descent result: if the propagated surrogate gradient is positively aligned with the gradient of an image-space distribution objective, a sufficiently small attribute step decreases that objective. The condition is stated explicitly and is not inferred from probe accuracy or batch correlation. It motivates local directional interventions and matched comparisons against direct feature steering and output-based input optimization. Full statements and proofs are in Appendix A.

Contributions.

- A formulation of population control that differentiates a terminal-time H-space attribute objective with respect to initial noise and renders the optimized tensors with the frozen sampler.
- A local characterization of the induced feature update and its input penalty, together with an output-descent condition and a corresponding comparison under shared stochastic perturbations.
- An evaluation of output discrepancy, image quality, and attribute response, with explicit matched protocols for testing the Jacobian and gradient-alignment interpretation.

Observed evidence. The reported CelebA-HQ gender summary gives mean discrepancy 0.018, versus 0.134 for vanilla generation and 0.035 for tuned distribution guidance. A separate direct-feature comparison attains lower discrepancy through initial-noise optimization with a higher FID. We retain both measurements and use them to motivate comparisons at matched distribution fit, rather than claiming simultaneous superiority in all metrics.

2 Related Work

Population objectives and feature guidance. Fair generative modeling seeks control over the distribution of generated data. Choi et al. (2020) use weak supervision from a reference dataset to reduce dataset bias in generative modeling. Balancing Act (Parihar et al., 2024) instead performs inference-time distribution guidance using attribute predictions from denoiser features. Its population matching objective and confidence-based sample selection are foundations for the present method. Our focus is the initial-input update induced by backpropagating such an objective through a terminal-time feature map.

Classifier guidance and Universal Guidance provide broader mechanisms for conditioning diffusion generation through attribute or content objectives (Dhariwal & Nichol, 2021; Bansal et al., 2023). Their scope underscores the need to specify the objective, representation, and update variable separately. A method can use the same attribute predictor and target distribution while changing which model state receives the gradient. This distinction defines the direct-H-space comparison in our evaluation.

Initial-noise optimization. DOODL optimizes diffusion inputs using gradients on generated pixels and an invertible sampling construction (Wallace et al., 2023). D-Flow differentiates through generative flows to optimize source points (Ben-Hamu et al., 2024), and diffusion-noise optimization has also been used as a motion prior (Karunratanakul et al., 2024). These methods establish input optimization as a control mechanism. Our surrogate reads one terminal-time bottleneck, so its backward pass ends at the initial input without traversing the rendering trajectory. This changes the feedback used by the optimizer and requires evaluating its output response.

Internal objectives for initial-noise optimization are also established. InitNO uses attention-based scores (Guo et al., 2024); FIND learns an initial-noise distribution with policy optimization (Chen et al., 2024). Noise Level Guidance refines initialization using a likelihood-based construction without backpropagation (Mannering et al., 2025). Consequently, our contribution is not the general use of internal signals or optimized initialization. We study a batch attribute objective, its input Jacobian, and the resulting image-distribution response within one formulation and evaluation protocol.

Semantic representations and initialization changes. H-space editing and interpretable diffusion directions reveal semantic structure within diffusion networks (Kwon et al., 2023; Haas et al., 2023). Latent Directions studies latent modifications for bias mitigation (Lopez Olmos et al., 2024), while Boost-and-Skip changes initialization variance and the sampling start for minority generation (Um et al., 2025). These approaches motivate studying the relation between a representation-level direction and its effect on generated content. The existence of semantic structure alone does not identify which terminal-time directions improve a population objective; we make that relationship an explicit target of analysis.

Score regularization and the comparison of update paths. Denoising score matching connects a noise predictor to the score of its noisy-data distribution (Song et al., 2021b). We distinguish this score from the exact Gaussian prior score when regularizing initial-noise updates. Our analysis first isolates the attribute direction, then relates it to a shared-noise comparison of the stochastic procedure. This separates the effect of the H-space objective from prior drift, injected noise, and rendering. The corresponding controls include direct feature updates, prior-only transitions, norm-matched random directions, and gradients obtained from an output objective.

3 H-SPACE-GUIDED INITIAL-NOISE OPTIMIZATION

We define the attribute objective in a terminal-time feature representation and apply its gradient to the initial input. Figure 1 separates the forward representation, the backward attribute path, and the final rendering stage.

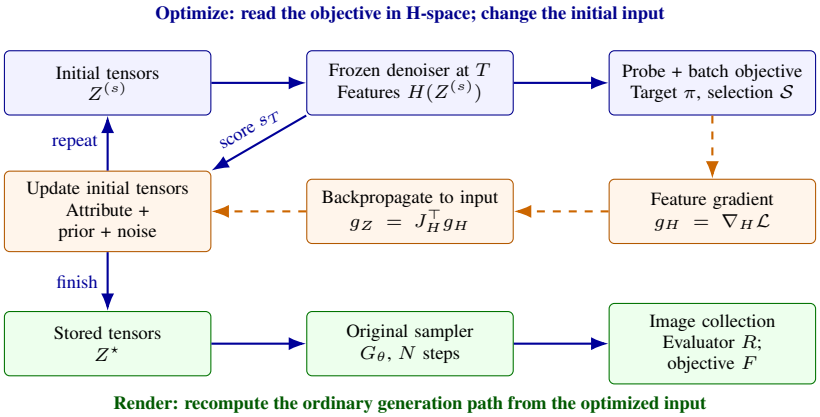


Figure 1: **The proposed forward and backward paths.** The dashed path translates an H-space objective into an initial-noise gradient. The renderer is evaluated after optimization. Gradients through that renderer are used only in the proposed diagnostic comparison, not in the method.

3.1 POPULATION OBJECTIVE AND FEATURE REPRESENTATION

Let G_θ be a frozen deterministic N-step DDIM sampler (Song et al., 2021a), and $Z = (z_1, \ldots, z_B)$ a stacked input batch with prior $p_0(Z) = \prod_i \mathcal{N}(z_i; 0, I_d)$. An image evaluator R_a supplies attribute probabilities, defining the output objective

$$r_a(Z) = \frac{1}{B} \sum_i R_a(G_\theta(z_i)), \qquad F_a(Z) = \frac{1}{2} \| r_a(Z) - \pi_a \|_2^2. \tag{1}$$

This objective specifies what is measured; it is not differentiated during the proposed optimization. Instead, a frozen linear probe C_a reads the terminal-time bottleneck $h_T(z_i)$:

$$p_{i,a} = \mathrm{softmax}(C_a(h_T(z_i))/\tau), \quad \bar{p}_a = \frac{1}{B} \sum_i p_{i,a}, \quad H(Z) = (h_T(z_1), \ldots, h_T(z_B)). \tag{2}$$

The evaluated target is $\pi_a = (1/2, 1/2)$ and T is the highest-noise timestep. On P2 the probe maps a flattened $512 \times 8 \times 8$ bottleneck to two logits. All samples contribute to the population statistic, although the implemented attribute gradient is allocated through a selected subset. Probe training and evaluation details appear in Appendix C.

3.2 FROM A FEATURE OBJECTIVE TO AN INPUT DIRECTION

For one attribute, the full-batch reference surrogate is $\mathcal{L}_{\text{batch}} = \kappa \sum_j (\bar{p}_{a,j} - \pi_{a,j})^2 / \pi_{a,j}$. The implementation selects samples from the overrepresented proxy class in increasing confidence order and uses their mean probabilities:

$$\mathcal{L}_a^{\text{prob}} = \kappa \sum_{j=0}^{1} \frac{(\bar{p}_{a,j}^{\mathcal{S}_a} - \pi_{a,j})^2}{\pi_{a,j}}, \qquad \bar{p}_a^{\mathcal{S}_a} = \frac{1}{|\mathcal{S}_a|} \sum_{i \in \mathcal{S}_a} p_{i,a}. \tag{3}$$

The majority and membership are fixed during differentiation and recomputed after each update. The selected mean determines the gradient; the full-batch proxy mean and the image mean in (1) are separate measurements. For multiple attributes we sum losses with fixed nonnegative weights. Selection size, stopping, and the margin-loss alternative are specified in Appendix B as operating choices of the evaluated method.

The backward path. Write $g_H = \nabla_H \mathcal{L}$ and $J_H = DH(Z)$ for the current fixed selection. The attribute gradient and a deterministic input step are

$$g_Z = \nabla_Z \mathcal{L}(H(Z)) = J_H^\top g_H, \qquad \Delta Z = -\eta g_Z. \tag{4}$$

Here η denotes an attribute-step length; DDIM stochasticity is fixed to zero throughout the main experiments. A vector–Jacobian product computes g_Z without forming J_H. The feature map acts sample by sample, while the aggregate objective couples the feature gradients across selected samples.

The induced feature change. Recomputing the bottleneck from the updated input gives, locally,

$$\Delta H = -\eta J_H J_H^\top g_H + O(\eta^2), \qquad \Delta Z = \arg\min_u \left\{ g_H^\top J_H u + \frac{\|u\|_2^2}{2\eta} \right\}. \tag{5}$$

Proposition 1 states the differentiability assumptions and proves both identities. Direct H-space descent would instead apply $-\eta g_H$. Through the input, feature directions are weighted by $J_H J_H^\top$, a positive semidefinite matrix determined by the current model sensitivity. It is generally not an orthogonal projector. This formulation identifies a quadratic cost on input movement; it does not assert optimal image quality or preservation of a data manifold.

The updated input also determines the skip features. If $U(Z)$ stacks the bottleneck and skips, the network recomputes $U(Z + \Delta Z)$ from that single input. A direct bottleneck intervention instead changes one activation while holding the remaining input-derived features fixed for that evaluation. Whether this difference improves content preservation is tested at matched output discrepancy, rather than assumed from the architecture.

The condition needed at the output. The input direction is useful for population control when it decreases $F = \sum_a F_a$. Theorem 2 provides a local sufficient condition: $\langle \nabla F, g_Z \rangle \geq c \|g_Z\|^2$ for $c > 0$, with a sufficiently small step. This condition concerns gradients, not classification accuracy. Section 4.4 specifies its direct measurement and the corresponding finite-difference intervention.

3.3 PRIOR-REGULARIZED OPTIMIZATION AND RENDERING

For step size $\epsilon > 0$, the evaluated stochastic update is

$$Z^{(s+1)} = Z^{(s)} + \frac{\epsilon}{2} s_T(Z^{(s)}) - \frac{\epsilon\lambda}{2} g_Z(Z^{(s)}) + \sqrt{\epsilon}\,\Xi^{(s)}, \quad \Xi^{(s)} \sim \mathcal{N}(0, I_{Bd}). \tag{6}$$

The attribute component has the form (4) with $\eta = \epsilon\lambda/2$. The denoiser-based score is $s_T(z) = -\epsilon_\theta(z, T)/\sqrt{1-\bar{\alpha}_T}$, which estimates the noisy-data score at T (Song et al., 2021b). The alternative $s_T(z) = -z$ is the exact score of the Gaussian initialization. The score and bottleneck can be read from the same model forward pass.

The prior drift and noise update every latent during an active iteration, including samples receiving zero attribute gradient. Appendix B specifies the stopping convention and the fixed energy reference underlying this regularization. The local attribute analysis isolates $-\eta g_Z$; Corollary 3 instead compares a guided step and a prior-only step under the same injected noise. It states the additional alignment condition at the perturbed point required for an output-descent conclusion.

Algorithm 1 H-space-guided initial-noise optimization

1: Draw $Z^{(0)} \sim p_0$; initialize the loss mode as unset.
2: **for** $s = 0, \ldots, S-1$ **do**
3: Evaluate the frozen denoiser at T to obtain $H(Z^{(s)})$ and s_T.
4: Read probe probabilities and compute batch statistics and selection.
5: **if** the stopping criterion in Appendix B is met **then**
6: **return** $Z^{(s)}$.
7: **end if**
8: If unset, choose probability or margin mode; retain it for this batch.
9: Compute g_H with the selected indices fixed.
10: Backpropagate $g_Z = J_H^\top g_H$ to the initial tensors.
11: Apply (6) with fresh Gaussian noise.
12: **end for**
13: **return** $Z^{(S)}$.

What is computed and stored. Each active iteration uses a terminal-time forward–backward pass. Once optimization ends, G_θ renders $Z^\star$ with N ordinary denoising steps. If an iteration costs C_T^{fb} and rendering costs $C_G(N)$, total computation is approximately $S_{\mathrm{act}} C_T^{\mathrm{fb}} + C_G(N)$ for $S_{\mathrm{act}} \le S$. This does not include a backward pass through G_θ. Runtime and peak memory must nevertheless be measured under matched implementations; the shorter gradient path alone is not a total runtime comparison against in-loop guidance.

Operating configuration. The reported configuration uses $B = 20$, $S = 50$, $\epsilon = 10^{-2}$, $\lambda = 1$, $\tau = 8$, $\kappa = 300$, and $N = 50$. Optimized tensors are stored with the model checkpoint, sampler, and preprocessing configuration. They reproduce the same collection under the same deterministic pipeline. New independent batches require new optimization. Transfer across rendering configurations is a separate measurement from replay and is included in the extended evaluation design in Appendix F.

4 EXPERIMENTS

We organize the evaluation around three questions: whether the initial-input objective controls the rendered distribution, how its operating point compares with direct H-space steering, and whether measured output responses support the proposed gradient interpretation. Existing measurements are reported with their protocol scope. New alignment and matched-operating-point measurements are marked as pending, rather than inferred from aggregate FD.

4.1 EVALUATION PROTOCOL

We evaluate P2 diffusion models (Choi et al., 2022) trained on CelebA-HQ and FFHQ at 256^2 resolution. Rendering uses 50-step DDIM with zero stochasticity and batch size 20. The main image evaluator is a ResNet-18 trained on CelebA-HQ; the optimization uses a separate linear bottleneck probe. FD is $\|r_a - \pi_a\|_2$, the distance between the mean image probability vector and the uniform target. The smooth objective in (1) equals $\mathrm{FD}^2/2$ for one attribute. FID (Heusel et al., 2017), hard-label proportions, and non-target drift provide complementary measurements.

The tuning seed is 1234 and the protocol lists evaluation seeds $\{5678, 9012, 3456, 7890\}$, with 1000 images per seed. We retain the reported evaluation-seed summaries without recomputing them from unavailable per-seed outputs. Appendix C records a conflicting seed-count caption in the supplied draft and the metadata needed to resolve it. The in-loop H-Distribution baseline uses its tuned $\gamma = 50$; direct H-space steering in the component comparison is a separately reported arm.

4.2 OUTPUT DISTRIBUTION UNDER THE ORIGINAL RENDERER

Table 1: **CelebA-HQ gender, reported evaluation-seed summaries.** FD mean $\pm$ standard deviation, 1000 images per seed. Rollout and unselected-batch rows are variants of this study. Additional comparisons and seed provenance appear in Appendix D.

method	FD $\downarrow$
Vanilla	0.134 ± 0.028
H-Distribution guidance, $\gamma = 50$	0.035 ± 0.015
Rollout input optimization, $K = 5$	0.015 ± 0.004
Unselected batch probability objective	0.025 ± 0.004
H-space-guided initial-noise optimization	0.018 ± 0.015

The terminal-time method has a lower reported mean discrepancy than vanilla generation. Its images are produced by the ordinary sampler after the attribute updates have ended, so this is an output-distribution measurement of the optimized input collection. The tuned in-loop baseline and the rollout variant also attain low discrepancy. These summaries do not establish a ranking among those control procedures or demonstrate that the terminal-time gradient approximates the output gradient.

The available run-level comparisons extend beyond the main gender setting. On CelebA-HQ race, vanilla FD 0.428 becomes 0.268; on FFHQ, gender FD 0.048 becomes 0.037 and race FD 0.410 becomes 0.348. The joint-attribute run reports gender/race FD $0.006/0.271$ against vanilla $0.120/0.428$. Appendix D preserves the baseline values and separates these rows from the evaluation-seed summaries. Marginal discrepancies do not specify the joint attribute distribution, which remains an additional evaluation target.

4.3 DIRECT FEATURE STEERING, IMAGE QUALITY, AND COST

The most direct comparison for the proposed interpretation changes the variable receiving the attribute update. Table 2 records the available H-space steering and input-optimization measurements. It also includes the unselected objective to distinguish update location from gradient allocation. Each row is a reported operating point; it is not a comparison at identical FD or identical latent displacement.

Table 2: **Component operating points.** CelebA-HQ gender, seed 1234, 1000 images. These are run-level ablation measurements; the H-space steering arm is distinct from the tuned baseline in Table 1.

configuration	class-1 (%)	FD ↓	FID ↓
Vanilla	41.8	0.114	43.5
Direct H-space steering	45.3	0.067	43.6
Input update, no boundary selection	53.0	0.043	41.9
Full initial-noise procedure	49.9	0.000	46.5

The full procedure obtains a lower FD than the reported direct-feature arm, with a higher FID. This supports evaluating the two update paths jointly on distribution fit and quality. It does not show that input-induced feature changes preserve image quality better. To test that hypothesis, the proposed comparison sweeps both procedures and measures FID, diversity, and non-target drift at common target-FD levels. The probe, target, and initial batches are shared; wall time is reported for each operating point.

Evaluator agreement. Table 3 scores the same run-level images using three evaluators. The full method improves on vanilla under all three, while its ordering relative to H-Distribution guidance depends on the evaluator. The gradient-alignment analysis therefore uses separately specified image evaluators and records their responses individually.

Table 3: **Same images under three evaluators.** FD on CelebA-HQ gender, seed 1234, 1000 images. A and C use ResNet-18 with different label sources; B is the FairFace ViT. See Appendix E.2.

method	A	C	B
Vanilla	0.114	0.112	0.196
H-Distribution guidance	0.003	0.035	0.091
Initial-noise optimization	0.001	0.005	0.113

Computation. The reported total runtime is $2.4\times$ vanilla for initial-noise optimization and $1.4\times$ for in-loop distribution guidance on an A100 at the operating batch size. A terminal-time backward path eliminates backpropagation through the rendering trajectory, but these measured ratios do not show a runtime advantage over the in-loop baseline. A matched comparison with output-based input optimization must report both wall time and peak memory. Replay from stored tensors and generating a new independent collection are accounted for separately.

4.4 DIRECTIONAL RESPONSE AND TESTS OF THE JACOBIAN INTERPRETATION

The new analysis concerns the direction induced by $J_H^\top g_H$. Existing sign and prior-only interventions motivate this question but do not estimate local gradient alignment. On CelebA-HQ, the available diagnostic reports class-1 rates of $44.9 \pm 2.2\%$ for Gaussian-score prior-only updates, $49.3 \pm 1.6\%$ for the full method, and 32.4% for a sign-reversed attribute update. These arms have different realization counts and are not a matched decomposition. Appendix E.3 preserves them with their provenance.

Direct alignment and finite differences. On a tractable subset of batches, compute both g_Z and ∇F through the fixed renderer. Record their cosine similarity and $\langle \nabla F, g_Z \rangle / \|g_Z\|^2$ when $g_Z \neq 0$. Alternatively, use paired renders of $Z \pm \eta g_Z$ at several small step lengths. The signed finite difference estimates the same directional derivative. Compare with norm-matched random directions and report each image evaluator separately. Neither a probe AUC nor an aggregate FD improvement substitutes for this measurement.

Table 4: **Tests of the proposed interpretation; measurements pending.** All result fields await matched experiments. The text states the hypotheses and controls, not conclusions inferred from these empty cells.

comparison	primary measurement	CelebA	FFHQ
Surrogate vs output gradient	alignment and signed response	—	—
Input vs direct H-space update	FID at matched output FD	—	—
Predicted vs actual feature change	relative linearization error	—	—
Terminal vs output-based objective	FD at matched time; memory	—	—
Bottleneck vs alternative features	response at matched budget	—	—

Feature sensitivity and recomputation. For fixed selection and decreasing η, compare the measured $H(Z - \eta g_Z) - H(Z)$ with $-\eta J_H g_Z$, obtained by a Jacobian–vector product. This checks the local approximation in (5). At matched output FD, compare paired-image change and non-target attributes between input updates and direct bottleneck edits. The former verifies the Jacobian calculation; the latter tests whether the architectural difference has an image-level benefit. They answer separate questions.

Shared stochastic controls. The finite optimizer also contains prior drift, injected noise, adaptive selection, and stopping. Its mechanism comparison therefore fixes the initial tensors and update budget, shares noise draws, and varies only the attribute direction or update location. Local tests hold selection fixed; complete-procedure comparisons evaluate the specified adaptive rules. Corollary 3 supplies the one-step comparison, and Proposition 4 supplies an exact Gaussian-score prior control. Appendix F specifies the extended experiments, including independent evaluators, joint attributes, and rendering transfer.

REFERENCES

Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Soumyadip Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Universal guidance for diffusion models. In *CVPR Workshops*, 2023.

Heli Ben-Hamu, Omri Puny, Itai Gat, Brian Karrer, Uriel Singer, and Yaron Lipman. D-flow: Differentiating through flows for controlled generation. In *ICML*, 2024.

Changgu Chen, Libing Yang, Xiaoyan Yang, Lianggangxu Chen, Gaoqi He, Changbo Wang, and Yang Li. Find: Fine-tuning initial noise distribution with policy optimization for diffusion models. In *ACM Multimedia*, 2024.

Jooyoung Choi, Jungbeom Lee, Chaehun Shin, Sungwon Kim, Hyunwoo Kim, and Sungroh Yoon. Perception prioritized training of diffusion models. In *CVPR*, 2022.

Kristy Choi, Aditya Grover, Trisha Singh, Rui Shu, and Stefano Ermon. Fair generative modeling via weak supervision. In *ICML*, 2020.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. In *NeurIPS*, 2021.

Xiefan Guo, Jinlin Liu, Miaomiao Cui, Jiankai Li, Hongyu Yang, and Di Huang. Initno: Boosting text-to-image diffusion models via initial noise optimization. In *CVPR*, 2024.

René Haas, Inbar Huberman-Spiegelglas, Rotem Mulayoff, and Tomer Michaeli. Discovering interpretable directions in the semantic latent space of diffusion models. In *arXiv:2303.11073*, 2023.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *NeurIPS*, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. URL https://arxiv.org/abs/2006.11239.

Korrawe Karunratanakul, Konpat Preechakul, Emre Aksan, Thabo Beeler, Supasorn Suwajanakorn, and Siyu Tang. Optimizing diffusion noise can serve as universal motion priors. In *CVPR*, 2024.

Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. Diffusion models already have a semantic latent space. In *ICLR*, 2023.

Carolina Lopez Olmos, Alexandros Neophytou, Sunando Sengupta, and Dim P. Papadopoulos. Latent directions: A simple pathway to bias mitigation in generative ai. In *CVPR Workshop on Responsible Generative AI*, 2024.

Harvey Mannering, Zhiwu Huang, and Adam Prugel-Bennett. Noise-level diffusion guidance: Well begun is half done. *arXiv preprint arXiv:2509.13936*, 2025. URL https://arxiv.org/abs/2509.13936.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *ICML*, pp. 8162–8171, 2021. URL https://proceedings.mlr.press/v139/nichol21a.html.

Rishubh Parihar, Abhijnya Bhat, Abhipsa Basu, Saswat Mallick, Jogendra Nath Kundu, and R Venkatesh Babu. Balancing act: Distribution-guided debiasing in diffusion models. In *CVPR*, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021a.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations (ICLR)*, 2021b.

Soobin Um, Beomsu Kim, and Jong Chul Ye. Boost-and-skip: A simple guidance-free diffusion for minority generation. In *ICML*, 2025.

Bram Wallace, Akash Gokul, Stefano Ermon, and Nikhil Naik. End-to-end diffusion latent optimization improves classifier guidance. In *ICCV*, 2023.

A Theoretical Analysis and Proofs

The analysis concerns a local attribute update, its effect on the output objective, and a matched stochastic comparison. The feature map H stacks the batch bottlenecks and all norms below are Euclidean, with the induced operator norm for Jacobians. A local surrogate $\ell(H)$ includes the probe, probability or margin loss, and fixed selection. Majority indices, membership, and the objective mode are held fixed in each local statement. These are applications of the chain rule, quadratic minimization, and the smooth descent inequality to the proposed computation. Their role is to identify the relevant quantities and conditions for empirical measurement.

A.1 Input update and induced feature geometry

Proposition 1 (Local input update induced by a feature objective)**.** *Let $H : \mathbb{R}^n \to \mathbb{R}^m$ be differentiable at Z, let $\ell : \mathbb{R}^m \to \mathbb{R}$ be differentiable at $H(Z)$, and write $J = DH(Z)$, $g_H = \nabla\ell(H(Z))$, and $g = J^\top g_H$. For $\eta > 0$, the unique solution of*

$$\min_u \; g_H^\top Ju + \frac{1}{2\eta}\|u\|_2^2 \tag{7}$$

is $u^\star = -\eta g$. If DH is locally M_H-Lipschitz on the segment from Z to $Z - \eta g$, then

$$H(Z - \eta g) - H(Z) = -\eta JJ^\top g_H + e_H, \qquad \|e_H\| \le \frac{M_H\eta^2}{2}\|g\|^2. \tag{8}$$

If $\ell \circ H$ has L_s-Lipschitz gradient on this segment, then

$$\ell(H(Z - \eta g)) \le \ell(H(Z)) - \eta(1 - L_s\eta/2)\|g\|^2. \tag{9}$$

For $g \neq 0$, the final inequality gives strict surrogate descent when $0 < \eta < 2/L_s$, with $L_s > 0$.

Proof. The gradient of the objective in (7) with respect to u is $J^\top g_H + u/\eta$, and its Hessian is $I/\eta \succ 0$. Its unique stationary point is therefore the stated minimizer. For $u = -\eta g$, the integral remainder satisfies

$$H(Z + u) - H(Z) - Ju = \int_0^1 [DH(Z + tu) - DH(Z)]u\, dt.$$

Its norm is at most $\int_0^1 M_H t\|u\|^2 dt = M_H\|u\|^2/2$, which proves (8). The chain rule gives $\nabla(\ell \circ H)(Z) = g$. Applying the smooth descent inequality to the step $-\eta g$ proves (9). □

Spectral interpretation and scope. For a singular value decomposition $J = U\Sigma V^\top$, $JJ^\top = U\Sigma\Sigma^\top U^\top$. Each feature-gradient component along a left singular vector is weighted by the corresponding squared singular value. This weighting is generally not an orthogonal projection: its eigenvalues need not be zero or one. If J has full row rank, every small feature direction is realizable to first order. The argument does not assume a lower-dimensional feature manifold. In a nonsmooth network, the Taylor statement applies within a region satisfying the stated regularity; changes of activation pattern or selection require their own local assessment. The stochastic optimizer additionally moves along a prior drift and random perturbation, so its complete step is not the minimizer of (7).

Bottleneck and skip features. Writing $U(Z) = (H(Z), V(Z))$ for the bottleneck and stacked skips, an input update produces $(H(Z + u), V(Z + u))$. Direct bottleneck editing can instead evaluate a decoder at $(H(Z) + v, V(Z))$. The former is computed from one input, but this fact alone

does not imply higher perceptual quality or that the latter pair is unattainable. The matched-FD comparison in Appendix F tests the practical consequence of this distinction.

A.2 Output descent from gradient alignment

Theorem 2 (Conditional descent of the image-distribution objective)**.** *Let F be differentiable with L_F-Lipschitz gradient on a convex neighborhood containing Z and $Z - \eta g$, with $L_F > 0$ and a fixed direction g. Then*

$$F(Z - \eta g) \le F(Z) - \eta \langle \nabla F(Z), g \rangle + \frac{L_F \eta^2}{2} \|g\|^2. \tag{10}$$

If $g \neq 0$ and $\langle \nabla F(Z), g \rangle \ge c\|g\|^2$ for some $c > 0$, then $F(Z - \eta g) < F(Z)$ whenever $0 < \eta < 2c/L_F$ and the segment remains in that neighborhood.

Proof. For $u = -\eta g$, the fundamental theorem of calculus gives

$$F(Z + u) - F(Z) = \langle \nabla F(Z), u \rangle + \int_0^1 \langle \nabla F(Z + tu) - \nabla F(Z), u \rangle \, dt.$$

Cauchy–Schwarz and the Lipschitz condition bound the integral by $L_F\|u\|^2/2$. Substitution yields (10). Under the alignment condition, the change is at most $-\eta(c - L_F\eta/2)\|g\|^2$, which is strictly negative in the stated step-size interval. □

The theorem applies to the smooth probability-based objective in (1) when the sampler, preprocessing, and evaluator satisfy the local assumptions. It does not apply directly to hard-label FD. For a single attribute, decreasing $F = \mathrm{FD}^2/2$ also decreases soft FD. For a sum of attribute objectives it guarantees descent of the sum, not simultaneous descent of each marginal. Probe AUC, calibration, and positive gradient alignment are different properties; none of the first two establishes the third. The constant c is a condition to measure or bound, not a value supplied by the existing experiments.

A.3 Comparisons with shared stochastic perturbations

Corollary 3 (One-step guided versus prior-only response)**.** *Fix Z, $g = g_Z(Z)$, and a shared drift/noise increment b. Let $Y = Z + b$ be the prior-only update and $Y_g = Y - \eta g$ the guided update. If F has L_F-Lipschitz gradient on their segment, then*

$$F(Y_g) - F(Y) \le -\eta \langle \nabla F(Y), g \rangle + \frac{L_F \eta^2}{2} \|g\|^2. \tag{11}$$

For random b, suppose the quantities are integrable, a common L_F bounds the relevant segments almost surely, and $\mathbb{E}[\langle \nabla F(Y), g \rangle \mid Z] \ge c\|g\|^2$ with $c > 0$. Then the conditional expected difference is negative for $g \neq 0$ and $0 < \eta < 2c/L_F$.

Proof. Apply (10) at Y with the fixed direction g, then take the conditional expectation. Given Z, g is fixed, and the assumed expected alignment gives the same strict upper bound. □

For (6), set $b = \epsilon s_T(Z)/2 + \sqrt{\epsilon}\,\Xi$ and $\eta = \epsilon\lambda/2$. The alignment in the corollary is at Y, while the surrogate gradient is computed at Z. Alignment at Z alone does not supply this condition. After multiple steps, trajectories and selection may differ even under common noise; consequently one-step comparisons and complete-trajectory ablations are reported separately.

A.4 Exact Gaussian-score prior control

Proposition 4 (Conditional law of a finite Gaussian-score chain)**.** *Let $0 < \epsilon < 4$, $a = 1 - \epsilon/2$, and let $S \ge 1$ be fixed. For independent standard Gaussian $\xi^{(s)}$, independent of $z^{(0)}$, consider $z^{(s+1)} = az^{(s)} + \sqrt{\epsilon}\,\xi^{(s)}$. Then, conditionally on $z^{(0)}$,*

$$z^{(S)} \overset{d}{=} \rho z^{(0)} + \sqrt{v_S}\xi, \quad \rho = a^S, \quad v_S = \frac{1 - \rho^2}{1 - \epsilon/4}, \quad \xi \sim \mathcal{N}(0, I_d). \tag{12}$$

If $z^{(0)} \sim \mathcal{N}(0, I_d)$, its final covariance is $[1 + \epsilon(1 - a^{2S})/(4 - \epsilon)]I_d$.

Proof. Unrolling the chain gives $z^{(S)} = a^S z^{(0)} + \sqrt{\epsilon}\sum_{j=0}^{S-1} a^{S-1-j}\xi^{(j)}$. The sum is independent Gaussian with covariance $\epsilon\sum_{j=0}^{S-1} a^{2j} I_d = (1-a^{2S})/(1-\epsilon/4)I_d$. Adding the initial covariance proves the marginal statement. □

The exact one-shot control uses v_S, not $1-\rho^2$. At $\epsilon = 0.01, S = 50$, the marginal variance multiplier is approximately 1.000988 and $\rho \approx 0.778313$. These are analytic values, not experimental estimates. Fixed S, the Gaussian score, and the absence of an attribute term are essential to this control. A data-dependent stopping time or a denoiser score defines a different transition. Equal conditional latent laws yield equal output laws through the same deterministic sampler.

B Implementation and Objective Details

Selection and stopping. For each binary uniform-target attribute, set $m = \arg\max_j(\bar{p}_{a,j} - \pi_{a,j})$, $\mathcal{M} = \{i : p_{i,a,m} > 1/2\}$, and $e = |\mathcal{M}| - B/2$. The selected count is $k = \max(0, \mathrm{round}(e/2))$; choose the k members of $\mathcal{M}$ with the smallest majority probability. The supplied single-attribute implementation stops the entire chain when $e < \sqrt{B/4}$ or selection is empty, and otherwise uses at most $S = 50$ steps. This is an operating tolerance, not a calibrated hypothesis test. The exact half-integer rounding and tie rules require confirmation from run metadata. Membership and majority are detached; the selected probabilities within the loss remain differentiable.

Margin objective. With logits $\ell_i = C_a(h_T(z_i))$ and $d_i = \ell_{i,m} - \ell_{i,1-m}$, the alternative is

$$\mathcal{L}^{\mathrm{margin}} = \frac{1}{k}\sum_{i\in\mathcal{S}} [d_i]_+^2. \tag{13}$$

The reported mode rule uses the margin objective if the probability loss at the first active step is within 10^{-3} of κ; otherwise it uses (3). The mode remains fixed for the batch. For fixed membership, writing $p_i = (1+e^{-d_i/\tau})^{-1}$ and $q = k^{-1}\sum_{i\in\mathcal{S}} p_i$ gives

$$\nabla_{z_i}\mathcal{L}^{\mathrm{prob}} = \frac{8\kappa(q-1/2)}{k\tau} p_i(1-p_i)\nabla_{z_i} d_i, \qquad \nabla_{z_i}\mathcal{L}^{\mathrm{margin}} = \frac{2[d_i]_+}{k}\nabla_{z_i} d_i.$$

These follow from $\mathcal{L}^{\mathrm{prob}} = 4\kappa(q-1/2)^2$ and the derivative of the squared positive part. Comparisons between the objectives therefore match gradient norm or input displacement as well as computational budget.

Multiple attributes and population statistics. Use $\mathcal{L} = \sum_a w_a \mathcal{L}_a$ with fixed $w_a \geq 0$; each attribute has its own selection and overlapping gradients are summed. The exact joint stopping rule and weights must accompany the joint run. The selected loss is not the full-batch discrepancy. For one attribute and $k < B$, $\bar{p} = (k/B)\bar{p}^{\mathcal{S}} + (1-k/B)\bar{p}^{\mathcal{S}^c}$. If $\delta = \|r - \bar{p}\|_2$, the triangle inequality gives

$$\|r-\pi\|_2 \leq \delta + \frac{k}{B}\sqrt{\frac{\mathcal{L}^{\mathrm{prob}}}{2\kappa}} + \left(1-\frac{k}{B}\right)\|\bar{p}^{\mathcal{S}^c} - \pi\|_2.$$

For $k = B$ omit the complement term. This separates current-batch proxy error, selected fit, and the remaining samples. It is not a convergence guarantee. Prior and noise updates act on all samples during an active step.

Reference distribution and score. For a fixed energy, the prior-regularized reference is $q_\lambda(Z) \propto p_0(Z)\exp[-\lambda\mathcal{L}(Z)]$. The exact score of p_0 is $-Z$; the denoiser score estimates the terminal noisy-data distribution. Initialization is sampled from the Gaussian law regardless of which score regularizes the subsequent update. The finite, adaptively selected procedure is not claimed to sample exactly from this reference distribution. Stored outputs must be the optimized tensors themselves; an integer RNG seed alone does not encode them.

C Evaluation Protocol and Reproduction

Models and probes. The unconditional experiments use P2 models on CelebA-HQ and FFHQ at 256^2. The reported UNet has width 128 and multipliers $(1, 1, 2, 2, 4, 4)$, with a linear $\beta \in [10^{-4}, 0.02]$ schedule over $T = 1000$ training timesteps. Rendering uses 50 DDIM steps and zero stochasticity. The probe is a linear map from 32768 features to two logits, trained on DDIM-inverted CelebA-HQ images with CLIP-ranked labels (Radford et al., 2021); top and bottom 1000 examples define the reported training subsets. Terminal-time weights are frozen for optimization.

Evaluators and quality. The primary image evaluator is a ResNet-18 with a $512 \to 256 \to 2$ head on 224^2 crops. FD uses its mean probability vector; hard-label proportions are separate observations and are not reconstructed from FD. FID uses 2048-dimensional pytorch-fid features and a balanced CelebA-HQ reference. Generated/reference sample counts and preprocessing must match within a comparison. Runtime includes optimization and rendering at the operating batch size on an A100. No verified peak-memory measurement is available for the new matched-gradient comparison.

Seed provenance and aggregation. The supplied protocol lists tuning seed 1234 and four evaluation seeds $\{5678, 9012, 3456, 7890\}$. Its comparison caption instead says "five-seed", while its statistical notes use three degrees of freedom. The reported means and standard deviations are preserved, but the exact contributing runs must be confirmed from manifests before asserting their sample count or recomputing significance. Run-level tables remain separate from these aggregate summaries. Samples optimized jointly are dependent; future uncertainty estimates use independent batches or paired seed-level differences rather than independent-image Bernoulli intervals.

Table 5: **Reproduction fields awaiting artifact confirmation.**

field	verified value
Model, probe, and evaluator checkpoint hashes	—
Exact 50-step inference timestep sequence	—
Seed manifests underlying aggregate tables	—
Half-integer rounding and tie conventions	—
Joint-attribute weights and stopping configuration	—
Binary race-label mapping	—
FID reference identity and sample count	—
Direct-H-space arm budget and guidance schedule	—

Baseline configuration. The main H-Distribution comparison uses the tuned $\gamma = 50$, while $\gamma = 5$ is retained as a separate reported setting below. Direct H-space steering in Table 2 is a different component arm. Rejection sampling selects images from a generated pool; its selecting classifier and final evaluator must be distinguished, and discarded generations count toward runtime. Rollout optimization propagates through $K = 5$ denoising steps before reading the probe; it is not a final-image objective and cannot substitute for the proposed output-gradient baseline. Annealed Langevin and adaptive steering are variants developed in this study, not additional published methods.

D Additional Reported Distribution and Component Measurements

The tables in this section retain relevant unconditional measurements from the supplied manuscript. They describe existing operating points, not new matched experiments establishing the Jacobian interpretation.

Independent rejection evaluation. The supplied rejection comparison reports independent-evaluator FD 0.066 versus 0.065 for our method, and race drift 0.013 versus 0.043. These are separate runs from Table 3. The source gives rejection time ratios $1.17\times$ and $1.21\times$ in different passages; a single matched-runtime claim awaits the run identifiers. This comparison is retained because rejection also produces stored inputs for an unmodified renderer.

Table 6: **Run-level attribute coverage.** Reported FD, $n \geq 1000$. The last column retains the same procedure with its stopping check disabled.

backbone	attribute	vanilla	H-Dist.	ours	no stopping
CelebA-HQ	gender	0.114	0.003	0.000	0.001
CelebA-HQ	race	0.428	0.410	0.268	0.267
FFHQ	gender	0.048	0.047	0.037	0.121
FFHQ	race	0.410	0.298	0.348	0.309
CelebA-HQ	gender + race	0.120/0.428	0.104/0.409	0.006/0.271	0.048/0.236

Table 7: **Additional reported evaluation-seed summaries.** Same aggregation source as Table 1. Rejection FD uses the selecting evaluator here, so its low value is not independent validation.

method	FD ↓
H-Distribution, $\gamma = 5$	0.123 ± 0.026
H-Sample guidance	0.370 ± 0.030
Rejection sampling, selecting evaluator	0.006 ± 0.004
Annealed Langevin	0.007 ± 0.005
Adaptive steering	0.015 ± 0.009

D.1 Sensitivity and selection

The following one-axis sweeps use CelebA-HQ gender, seed 1234, and 1000 images from the reported operating point. They do not match output FD or stochastic displacement across settings.

parameter	tested values	corresponding FD
ϵ	$0.003, 0.01, 0.03$	$0.071, 0.001, 0.383$
S	$10, 25, 50, 100$	$0.047, 0.018, 0.001, 0.074$
λ	$0.3, 1, 3$	$0.023, 0.001, 0.057$

The separate selection sweep expands from half the excess ($\alpha = 0$) to the whole batch ($\alpha = 1$). Its endpoint is a different configuration from the unselected-batch ablation and must not be substituted for it.

α	0	0.15	0.35	0.60	0.85	1.00
class-1 (%)	49.8	53.5	57.5	60.4	59.1	55.9
FD	0.001	0.051	0.107	0.142	0.127	0.090
FID	46.5	47.4	48.6	49.6	49.6	48.5

D.2 Rendering stochasticity

The main method uses deterministic rendering. A supplied sampler sweep records the following FD values on CelebA-HQ gender, seed 1234, $n = 1000$. The symbol η_{DDIM} here denotes sampler stochasticity, not the attribute-step length in the analysis.

η_{DDIM}	0	0.25	0.50	1.00
Vanilla FD	0.114	0.108	0.128	0.232
Initial-noise optimization FD	0.001	0.034	0.125	0.225

These measurements delimit the evaluated rendering configuration. They do not establish transfer to arbitrary samplers or checkpoint changes. The exact deterministic configuration must accompany stored optimized tensors.

Table 8: **Additional component controls.** CelebA-HQ gender, seed 1234, 1000 images. Together with Table 2, these retain the reported ablation; stopping-disabled FFHQ appears in Table 6.

configuration	class-1 (%)	FD	FID
Gaussian prior score $-z$	50.6	0.010	46.5
No injected Langevin noise	98.4	0.348	340.1

E Probe, Evaluator, and Directional Diagnostics

E.1 Predictive diagnostics

The supplied diagnostic scores probe predictions against final image labels on 300 vanilla samples across the rendering schedule:

t	999	978	795	489	305	203	101	0
AUC	0.557	0.553	0.507	0.533	0.532	0.520	0.509	0.528

FFHQ terminal-time AUC is 0.537. The reported per-image Pearson correlation between probe and image probabilities is 0.094 on CelebA-HQ and 0.051 on FFHQ. The source also reports batch-mean correlations 0.47 and 0.00, respectively, without sufficient batch-level metadata to reproduce them here. Neither the ranking of AUC across timesteps nor these correlations proves that the terminal-time gradient is optimal or positively aligned with the output gradient. The directional experiments below measure the property required by Theorem 2.

A separate trace in the supplied draft reports selected majority probability $0.549 \to 0.511$ and full-batch probability $0.672 \to 0.666$. These are proxy statistics of different populations. They motivate logging the selected loss, full-batch mean, and image-level discrepancy separately; they are not substituted for a direct gradient-alignment measurement.

E.2 Evaluator definitions and additional rows

Evaluator A is the CelebA-HQ ResNet-18 with CLIP-ranked labels; C shares the architecture and training images but uses FairFace-supplied labels; B is the FairFace ViT. In addition to Table 3, annealed Langevin has FD $0.023/0.013/0.056$ under A/C/B and adaptive steering has $0.016/0.027/0.169$. Per-image agreement between A and B is 0.909 on vanilla and 0.881 on our outputs. Changing both architecture and training distribution in B does not isolate either factor causally. The current evaluator evidence motivates reporting each evaluator individually.

The source flags eyeglasses and smile evaluators with an incorrect image-label filename mapping. Those attributes are excluded from this manuscript's measurements. Any future non-target-attribute evaluation must use verified image-label pairs, including a real-data base-rate check.

E.3 Reported direction and prior-only interventions

Table 9: **Existing directional diagnostic, not a matched decomposition.** Realization counts follow the supplied appendix; its narrative gives a conflicting count of four for the Gaussian prior-only arm.

arm	realizations	class-1 (%)
Vanilla input	1	41.8
Gaussian-preserving one-shot control, four different ρ values	4	41.4 ± 1.2
Gaussian-score prior-only chain	5	44.9 ± 2.2
Denoiser-score prior-only chain	1	47.0
Full method, positive attribute weight	4	49.3 ± 1.6
Sign-reversed attribute update	1	32.4

The one-shot arm uses $\rho \in \{0.95, 0.78, 0.5, 0\}$ and $z' = \rho z + \sqrt{1-\rho^2}\xi$. It is not the exact law of the finite unadjusted chain: Proposition 4 gives the correct variance. Pooling different ρ

values also does not estimate repeated draws at a fixed setting. The observations are retained, but their difference cannot establish a path-dependent effect under identical laws, and differences across these arms are not additive causal contributions. The source's FFHQ sign sweep places all three weights $-1, 0, +1$ within one percentage point of 58.7%; matched local tests are required there as well. New comparisons share initial batches, score, fixed iteration count, and Gaussian draws before attributing changes to the feature gradient.

F Matched Analysis Protocols

These experiments test the new interpretation; their results remain unmeasured in Table 4. Outcomes are reported for all preselected configurations and evaluators, including those that do not support the stated hypotheses.

Gradient alignment. At saved initial or intermediate optimization tensors, fix membership and loss mode. Compute $g = J_H^\top g_H$ and, on batches that fit the available memory, ∇F through the deterministic renderer and differentiable image evaluator. Record cosine similarity and $c(Z) = \langle \nabla F, g \rangle / \|g\|^2$ for $g \neq 0$, together with the proportion of zero gradients. Report distributions across batches, not just a pooled mean. Each evaluator defines a separate F.

Paired finite differences. Render $Z - \eta g$ and $Z + \eta g$ with common configuration and several decreasing η. The central difference $[F(Z + \eta g) - F(Z - \eta g)]/(2\eta)$ converges to $\langle \nabla F, g \rangle$ under differentiability. Compare the descent direction, its sign reversal, and norm-matched random directions. Random directions receive independent repetitions. Local deterministic tests isolate the attribute direction; stochastic tests share the drift and noise and evaluate the condition at Y in Corollary 3.

Jacobian prediction and update location. Use a Jacobian–vector product to compute $J_H g$ and compare it with the observed feature difference after an input step. Report absolute error and relative error with a stated nonzero denominator threshold across step sizes. This verifies local linearization; it does not by itself validate image semantics. Direct-H-space and input-update arms share the probe, target, starting batch, and declared selection rule. Sweep each arm's strength and budget, then compare FID, diversity, and non-target drift at overlapping output-FD levels. Report unattained target levels explicitly.

Alternative representations and output objectives. Compare the bottleneck with other denoiser layers or attention-based representations using controlled training data and recorded probe capacity. For output-based optimization, backpropagate the same differentiable population objective F through the renderer. Report FD versus total time and peak memory under the same hardware, precision, batch size, checkpointing, and iteration accounting. The existing $K = 5$ rollout variant is recorded separately because it reads an intermediate probe.

Joint attributes, prior controls, and rendering transfer. For joint control report the distribution over attribute combinations as well as each marginal. Compare the Gaussian-score prior-only chain with the exact fixed-S law in (12); evaluate the denoiser score separately. For stored optimized tensors first verify deterministic replay, then vary rendering step count and report FD, FID, and diversity for each configuration. New independent collections receive new optimization and their full generation cost is included.

G Scope and Conclusion

The completed measurements concern perceived binary attributes on two P2 unconditional face-generation backbones. The mathematical analysis describes local gradients and explicitly conditional output descent, not convergence of the finite adaptive optimizer. Recomputing internal features from an updated input is a structural property; perceptual preservation and a favorable quality–distribution tradeoff require the matched measurements.

H-space-guided initial-noise optimization connects a population objective on internal features to a concrete sampler input. Its Jacobian determines the induced feature change and its alignment with

the output objective determines local control effectiveness. The reported distribution results support this route to initial-input control in the evaluated settings. The new analysis specifies how to test its mechanism and compare its quality and computational tradeoffs against direct feature and output-based optimization.