Guidance for Low-Level Perceptual Editing in Unconditional Diffusion Models

Shreyansh Modi* Akshat Tomar* Aarush Aggarwal*
shreyansh_m@ee.iitr.ac.in akshat_t@mfs.iitr.ac.in aarush_a@ma.iitr.ac.in

Abstract

Unconditional diffusion models offer powerful generative priors, yet steering them toward aesthetically enhanced outputs remains largely unexplored. We show that h-space patching, the dominant paradigm for training-free diffusion editing, systematically fails for global, low-level transformations required for aesthetic and perceptual refinement. We introduce a novel, generalized framework for image-editing in unconditional diffusion models without explicit training. This inference-time mechanism operates on low-level features by extracting degradation concept vectors and combining bottleneck patching with classifier-free guidance to guide sampling away from the degraded manifold, producing consistently improved images without any model retraining.

1. Introduction

Diffusion models have emerged as the state-of-the-art paradigm for image synthesis [10], with deterministic DDIM sampling [20] enabling near-perfect reconstruction.

Guidance mechanisms [2, 9] further steer the reverse diffusion process toward desired distributions, with recent inference-time methods constructing negative baselines by intervening directly in internal representations [1, 5, 11, 12, 19] circumventing the need for conditional training entirely.

Despite these advances, fine-grained control over low-level perceptual features such as sharpness, contrast, saturation remain an open problem in unconditional diffusion models. These concepts are well-studied in generative adversarial networks [7]. Interpretable directions in the GAN latent space have been shown to correspond to photometric transformations including brightness, color balance, and contrast [4, 13, 16, 21].

A promising direction comes from the U-Net bottleneck, or *h-space*, which behaves as a semantically dense latent space amenable to linear manipulation [8, 14, 17]. These properties make *h-space* a natural candidate for encoding low-level concept directions.

Figure 1. Our method yields sharper details and fewer artifacts than the baseline in all three examples.

We introduce a unified editing paradigm to achieve this. We first extract supervised directions in *h-space* corresponding to low-level degradations (blur, low contrast, grayscale)(Figure 3). After inhibiting this direction in the bottleneck of the U-Net, the structurally degraded noise prediction is used to guide the generative trajectory towards aesthetically enhanced images using classifier-free guidance (Figure 2). Since our inference method generalizes [8, 17], it supports both low-level perceptual and semantic concept editing. We validate through FID comparisons and human evaluation studies (Table 1), consistently outperforming activation patching baselines across low-level editing directions.

We show that classical bottleneck patching fails on this problem due to destructive interference in the decoder of the U-Net. We perform ablations showing our method's transferability across datasets. We also experiment with time-dependent guidance schedules during reverse diffusion to decrease computational cost.

2. Methodology

We formalize an inference-time mechanism aimed at guiding the diffusion process away from concepts associated with low-level perceptual degradation. Our approach is: (1) isolating degraded bottleneck feature representations; (2) applying activation patching during inference; and (3) updating the noise predictions through classifier-free guidance.

*Equal contribution.
†Indian Institute of Technology Roorkee.

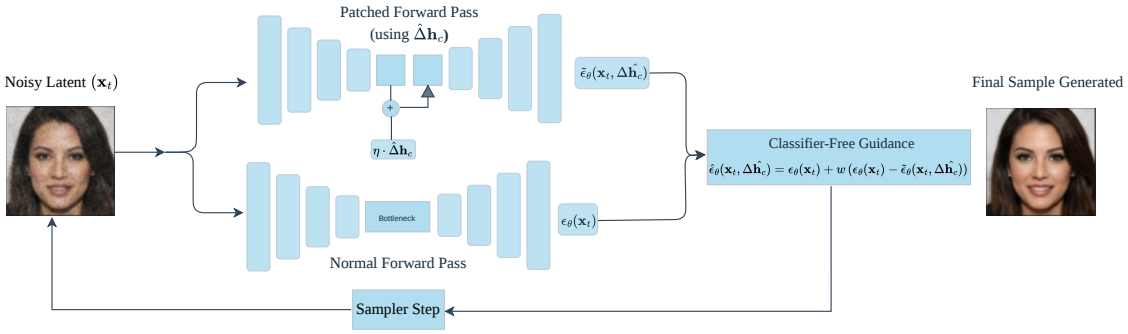


Figure 2. Overview of our inference framework. The process begins after the concept vector, shown in Figure 3, is computed as the mean pairwise difference of the degraded and clean *h-space* vectors at the extraction timestep k. This is followed by (a) Patching where the vector is injected into the bottleneck during inference. Finally, the modified activations are used to compute the (b) Guided Noise Prediction, which is used as the updated prediction of the score function $\nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t)$

2.1. Concept Vector Extraction

Let $\mathcal{T}_c$ denote a transformation[1] causing a degradation c (e.g. blur, grayscale, low-contrast) in a clean image ($\mathbf{x}$). Let $\mathbf{h}_t$ and $\mathbf{h}'_t$ denote the bottleneck activations of $\mathbf{x}$ and $\mathcal{T}_c(\mathbf{x})$ at extraction timestep t. The degraded concept vector $\Delta\mathbf{h}_c$ is then computed as the average pairwise difference:

$$\Delta\mathbf{h}_c = \frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{h}'^{(i)}_t - \mathbf{h}^{(i)}_t\right) \tag{1}$$

This yields a direction $\Delta\mathbf{h}_c$ in the *h-space* that points from the clean toward the degraded concept manifold. Fig 3 displays the approach.

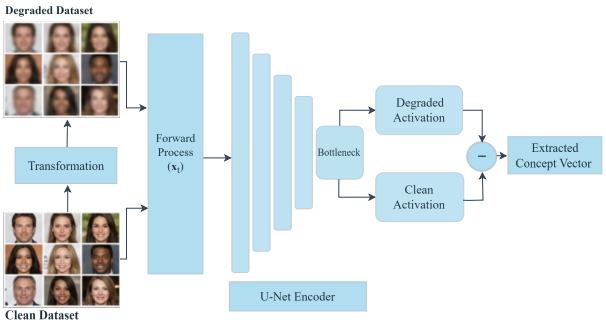


Figure 3. Paired-data concept-vector extraction $\Delta\mathbf{h}_c$.

The choice of timestep t governs the frequency content of the activations used for extraction.[2]

2.2. Inference Activation Patching

During the reverse diffusion process, at each timestep t, we intercept the bottleneck activation $\mathbf{h}_t$ produced by the U-Net encoder and apply a directional patch in the direction of $\Delta\hat{\mathbf{h}}_c$:

$$\tilde{\mathbf{h}}_t = \mathbf{h}_t + \eta \cdot \Delta\hat{\mathbf{h}}_c \tag{2}$$

where $\Delta\hat{\mathbf{h}}_c = \Delta\mathbf{h}_c / \|\Delta\mathbf{h}_c\|_2$ is the unit-normalized concept vector, $\eta \in \mathbb{R}^+$ controls the magnitude of the patching. Unlike prior activation patching methods [6, 14], we patch in a direction *towards* the degradation of the image. We justify this empirically in Section 3.3.

2.3. Negative Classifier-Free Guidance

We use the noise predicted by the U-Net decoder as a *conditional proxy* for the degraded concept. We then apply classifier-free guidance [9] to guide away from it by reversing the sign of the guidance scale. At each timestep t, the modified noise prediction is:

$$\hat{\epsilon}_\theta(\mathbf{x}_t, \Delta\hat{\mathbf{h}}_c) = \epsilon_\theta(\mathbf{x}_t) + w\,(\epsilon_\theta(\mathbf{x}_t) - \tilde{\epsilon}_\theta(\mathbf{x}_t, \Delta\hat{\mathbf{h}}_c)) \tag{3}$$

where $\epsilon_\theta(\mathbf{x}_t)$ is the standard unconditional noise prediction, $\tilde{\epsilon}_\theta(\mathbf{x}_t, \Delta\hat{\mathbf{h}}_c)$ is the noise prediction after intercepting with the patched bottleneck $\tilde{\mathbf{h}}_t$ using Equation (2), and $w \in \mathbb{R}^+$ is the guidance scale.

Figure 2 summarizes the overall workflow, detailing the corresponding procedure. We also conduct a thorough inspection of this formulation in Section 3.3.

3. Experiments

3.1. Implementation Details and Metrics

We instantiate our method on a frozen pretrained unconditional DDPM backbone[3], applying all edits at inference time without updating any model parameters. All experiments use 256×256 resolution with 30 DDIM steps on CelebA-HQ. For each degradation concept, vectors are extracted from paired degraded/clean subsets of $N \leq 100$ images.

[1] Mathematical formulas provided in Appendix A
[2] Further analysis is provided in Appendix C
[3] HuggingFace checkpoint: https://huggingface.co/google/ddpm-celebahq-256.

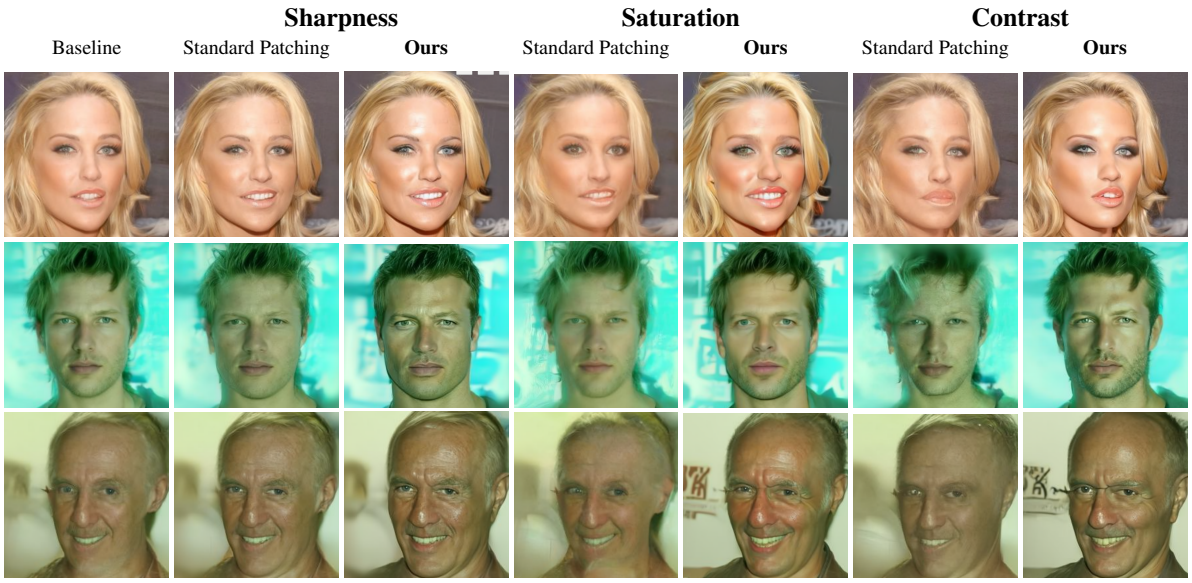


Figure 4. Comparison of Baseline, Standard Patching and our method across sharpness, saturation, and contrast. For Guidance, a CFG-scale of +1.5 was used for all the samples across all the directions.

We consider three low-level degradations *blur*, *low contrast*, and *grayscale* and report their positive edit directions as *sharpness*, *contrast*, and *saturation*, respectively.

For evaluation, we report Fréchet Inception Distance (FID) and human pairwise A/B comparisons between our method and [8] under identical initial noise and sampling settings. 60 human evaluators selected the image better representing the target edit, with the baseline sample as anchor, and a *neither* option available.

3.2. Results

Qualitatively, our method produces edits that are stronger and cleaner than standard patching alone (Fig 4). For sharpness, patch-only editing introduces indiscriminate smoothing or structural artifacts, whereas our method yields the intended enhancement while preserving global composition and identity. Similar trends hold for contrast and saturation, where patch-only editing produces washed-out highlights or uneven desaturation compared to our more uniform transformations.

Quantitatively, Table 1 confirms these observations. Across all concepts, our method achieves lower FID and higher direction-specific metrics: Laplacian variance for sharpness, mean S-channel for saturation, and RMS contrast for contrast. Human evaluations corroborate this: annotators prefer our method in **76.0%** of pairwise comparisons for sharpness, with consistent advantages across remaining concepts. Together, these results demonstrate that combining bottleneck patching with noise-space extrapolation outperforms either component in isolation.[4]

Table 1. Percentage change relative to baseline FID, where negative values indicate improvement. Direction-specific metrics report absolute values, higher is better.

Direction	Metric	Baseline	Standard Patching	Ours
Sharpness	FID (%Δ)	25.43	+3.54%	**-6.07%**
	Laplacian variance	143.77	212.72	**386.45**
Saturation	FID (%Δ)	25.43	+7.03%	**-7.76%**
	Mean S-channel	0.43	0.44	**0.47**
Contrast	FID (%Δ)	25.43	+4.85%	**-13.90%**
	RMS contrast	0.18	0.17	**0.21**

Further ablations validating generalization, semantic concept directions, and our partial guidance variant are provided in Sec. D.

3.3. Empirical Analysis of Patching and Guidance

Empirically validating Sec 2, we demonstrate that the *h-space* linearity assumed in prior work [8, 14, 17] fails for low-level perceptual degradation. Unlike high-level features, negatively patching our concept vector into the U-Net bottleneck causes destructive interference and collapses image quality. However, our method remains stable, concentrating the sampling distribution toward viable, upgraded outputs.

[4]Baseline FID is comparatively high due to limited compute, which constrained the number of generated samples to 10k per concept direction and inference steps to 30 (DDIM) used for evaluation.

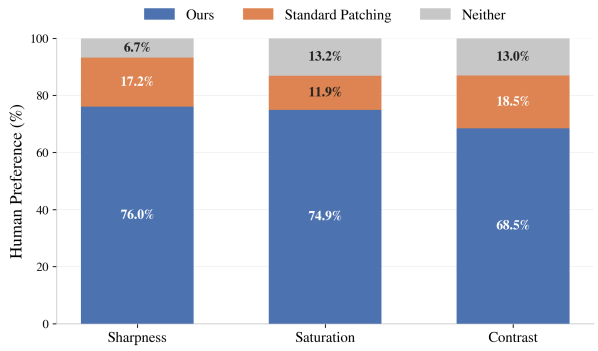


Figure 5. Human preference rates across sharpness, saturation, and contrast. Our method is preferred in all three settings.

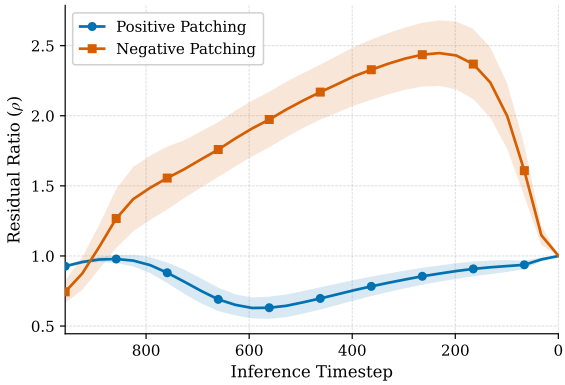


Figure 6. Relative residual ratio ρ across denoising timesteps for the blur direction.

Non-linear disturbance in the decoder: We express the noise prediction $\hat{\epsilon}_\theta$ (3) in terms of the U-Net decoder $\mathcal{D}_\theta$, which is written as a function of the bottleneck activation $\mathbf{h}_t$, skip-connections $\{s_i\}_{i=1}^{L}$, timestep embedding (τ) and concept vector $\Delta\hat{\mathbf{h}}_c$.

The skip-connections $\{s_i\}$ and timestep embedding τ are suppressed from the notation for clarity.

$$\hat{\epsilon}_\theta\left(\mathbf{x}_t, \Delta\hat{\mathbf{h}}_c\right) = \mathcal{D}_\theta(\mathbf{h}_t) + w \cdot \left[\mathcal{D}_\theta(\mathbf{h}_t) - \mathcal{D}_\theta(\mathbf{h}_t + \eta\Delta\hat{\mathbf{h}}_c)\right]$$

Since $\mathcal{D}_\theta$ is composed of differentiable operations,[5] we expand $\mathcal{D}_\theta(\mathbf{h}_t + \eta\,\Delta\hat{\mathbf{h}}_c)$ in a Taylor series about $\mathbf{h}_t$:

$$\mathcal{D}_\theta\left(\mathbf{h}_t + \eta\,\Delta\hat{\mathbf{h}}_c\right) = \mathcal{D}_\theta(\mathbf{h}_t) + \eta\,\mathbf{J}_\mathcal{D}(\mathbf{h}_t)\,\Delta\hat{\mathbf{h}}_c + \mathbf{R}_{\geq 2} \quad (4)$$

where $\mathbf{J}_\mathcal{D}(\mathbf{h}_t)\,\Delta\hat{\mathbf{h}}_c$ is the Jacobian-vector product, and $\mathbf{R}_{\geq 2}$ collects all higher-order terms.

We obtain the decomposed update equation as:

$$\hat{\epsilon}_\theta\left(\mathbf{x}_t, \Delta\hat{\mathbf{h}}_c\right) = \mathcal{D}_\theta(\mathbf{h}_t) - w\cdot\eta\;\mathbf{J}_\mathcal{D}(\mathbf{h}_t)\Delta\hat{\mathbf{h}}_c - w\cdot\mathbf{R}_{\geq 2} \quad (5)$$

To highlight the non-linearity present in the decoder, which acts as the cause of failure in negative patching, we track the *relative residual ratio*:

$$\rho \;=\; \frac{\|w\cdot\mathbf{R}_{\geq 2}\|_2}{\|\hat{\epsilon}_\theta\left(\mathbf{x}_t, \Delta\hat{\mathbf{h}}_c\right) - \mathcal{D}_\theta(\mathbf{h}_t)\|_2} \quad (6)$$

Relative residual ratio (ρ) quantifies how well the linear term $(\hat{\epsilon}_\theta^1)$ aligns with the nonlinear components in (5).[6]

When patching positively $(\eta > 0)$, $\rho < 1$ throughout the trajectory, confirming that higher-order terms remain aligned with the guidance direction, preserving image quality. Conversely, negative patching $(\eta < 0)$ causes ρ to overshoot unity mid-trajectory, inducing destructive interference with the first-order term, where the decoder's non-linear response opposes the intended guidance direction and degrades output quality.

Additionally, we observed that direct patching in the negative direction blurs the decoder's attention maps. This phenomenon is analyzed in detail in the Sec. B.3.

4. Future Work

Several directions remain open. More principled approaches to reducing computational cost, such as adaptive guidance scheduling driven by real-time signals like attention entropy, could make the method practical at larger scale. Extending the framework to flow-based generative models and text-conditioned diffusion models would significantly broaden its applicability. Replacing manual hyperparameter selection with self-tuning mechanisms that infer appropriate patching and guidance strengths directly from image content would improve usability. Finally, a sophisticated, theoretical study on failure of traditional patching with low-level transformations constitutes an interesting direction of work.

5. Conclusion

We presented a training-free, inference-time framework for inducing low-level perceptual transformations in unconditional diffusion models, successfully producing improvements without any model retraining. Our pipeline combined two separately effective techniques in the field of diffusion models: *h-space* manipulation and classifier-free guidance. We further provided a mechanistic analysis of why prior methods fail for low-level edits, tracing the breakdown to unstable non-linear residual in the decoder of the U-Net. Ablations across datasets and concept directions confirm the generality of our approach.

[5] Architecture details: `google/ddpm-celebahq-256`.

[6] A complete derivation is provided in Appendix B.1.

References

[1] Donghoon Ahn, Hyoungwon Cho, Jaewon Min, Wooseok Jang, Jungwoo Kim, SeonHwa Kim, Hyun Hee Park, Kyong Hwan Jin, and Seungryong Kim. Self-rectifying diffusion sampling with perturbed-attention guidance. In *ECCV*, 2024. 1

[2] Prafulla Dhariwal and Alex Nichol. Diffusion models beat GANs on image synthesis. In *NeurIPS*, 2021. 1

[3] Ronald A Fisher. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2):179–188, 1936. 8

[4] Ingo Fründ, J. Patel, and E. D. Stalker. Contrast invariant tuning in human perception of image content. *bioRxiv preprint 10.1101/711804*, 2019. 1

[5] Fengyi Fu, Mengqi Huang, Lei Zhang, and Zhendong Mao. Layeredit: Disentangled multi-object editing via conflict-aware multi-layer learning. *arXiv preprint arXiv:2511.08251*, 2025. 1

[6] Rohit Gandikota, Joanna Materzynska, Tingrui Zhou, Antonio Torralba, and David Bau. Concept sliders: LoRA adaptors for precise control in diffusion models. In *ECCV*, 2024. 2

[7] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014. 1

[8] Rene Haas, Inbar Huberman-Spiegelglas, Rotem Mulayoff, Stella Grasshof, Sami Sebastian Brandt, and Tomer Michaeli. Discovering interpretable directions in the semantic latent space of diffusion models. In *IEEE International Conference on Automatic Face and Gesture Recognition*, 2024. 1, 3, 9

[9] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 1, 2

[10] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 1

[11] Susung Hong. Smoothed energy guidance: Guiding diffusion models with reduced energy curvature of attention. *arXiv preprint arXiv:2408.00760*, 2024. 1

[12] Susung Hong, Gyuseong Lee, Wooseok Jang, and Seungryong Kim. Improving sample quality of diffusion models using self-attention guidance. In *ICCV*, pages 7428–7437, 2023. 1

[13] Ali Jahanian, Lucy Chai, and Phillip Isola. On the "Steerability" of generative adversarial networks. In *ICLR*, 2020. 1

[14] Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. Diffusion models already have a semantic latent space. In *ICLR*, 2023. 1, 2, 3

[15] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, pages 3730–3738, 2015. 9

[16] Amar Mušić, Anne-Sofie Maertens, and Johan Wagemans. Beautification of images by generative adversarial networks. *J. Vis.*, 23(10):14, 2023. 1

[17] Yong-Hyun Park, Mingi Kwon, Junghyo Jo, and Youngjung Uh. Unsupervised discovery of semantic latent directions in diffusion models. *arXiv preprint arXiv:2302.12469*, 2023. 1, 3

[18] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763, 2021. 9

[19] Dazhong Shen, Guanglu Song, Zeyue Xue, Fu-Yun Wang, and Yu Liu. Rethinking the spatial inconsistency in classifier-free diffusion guidance. In *CVPR*, 2024. 1

[20] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. 1

[21] Nurit Spingarn, Ron Banner, and Tomer Michaeli. GAN "steerability" without optimization. In *ICLR*, 2021. 1

Appendix

A. Transformations

For each contrastive pair, the degraded image is obtained by applying the transformation to the clean RGB image prior to resizing and normalization. The three transformations are defined as follows.

Blur. The blurred image $\tilde{x}$ is obtained by convolving the clean image x with a 2-D Gaussian kernel K:

$$\tilde{x}[i,j] = \sum_{m=-r}^{r} \sum_{n=-r}^{r} K[m,n] \cdot x[i+m,\, j+n] \tag{7}$$

where the kernel is constructed as the outer product of a 1-D Gaussian:

$$K = \mathbf{k}\mathbf{k}^{\top}, \qquad k_i = \frac{1}{Z} \exp\!\left(-\frac{i^2}{2\sigma^2}\right) \tag{8}$$

with kernel size 21×21, standard deviation $\sigma = 3.0$, half-width $r = 10$, and Z being the normalisation constant. Border pixels are handled via reflect padding.

Grayscale. The image is converted to luminance using the ITU-R BT.601 luma coefficients and then replicated across all three channels:

$$\tilde{x} = 0.299\, R + 0.587\, G + 0.114\, B \tag{9}$$

yielding a three-channel tensor $(\tilde{x},\, \tilde{x},\, \tilde{x})$ with no chromatic information.

Low Contrast. Contrast is reduced by linearly interpolating each pixel toward the mean luminance μ of the image:

$$\tilde{x} = \mu + \alpha \cdot (x - \mu) \tag{10}$$

where $\alpha = 0.6$ is the contrast factor and μ denotes the per-image mean luminance. Since $\alpha < 1$, the pixel range is compressed toward the mean, attenuating contrast.

B. Empirical Analysis of Standard Patching and Guidance

B.1. Residual-Ratio Derivation

Let $\hat{\boldsymbol{\epsilon}}_1 = -w \cdot \eta\, \mathbf{J}_{\mathcal{D}}(\mathbf{h}_t)\Delta\hat{\mathbf{h}}_c$ denote the first-order guidance term and $\widehat{\mathbf{R}}_{\geq 2} = -w \cdot \mathbf{R}_{\geq 2}$ denote the scaled higher-order residual from Eq. (5), so that:

$$\hat{\epsilon}_\theta\Big(\mathbf{x}_t, \Delta\hat{\mathbf{h}}_c\Big) - \mathcal{D}_\theta(\mathbf{h}_t) = \hat{\boldsymbol{\epsilon}}_1 + \widehat{\mathbf{R}}_{\geq 2} \tag{11}$$

The relative residual ratio is then:

$$\rho = \frac{\|\widehat{\mathbf{R}}_{\geq 2}\|_2}{\|\hat{\epsilon}_\theta\Big(\mathbf{x}_t,\, \Delta\hat{\mathbf{h}}_c\Big) - \mathcal{D}_\theta(\mathbf{h}_t)\|_2} = \frac{\|\widehat{\mathbf{R}}_{\geq 2}\|_2}{\|\hat{\boldsymbol{\epsilon}}_1 + \widehat{\mathbf{R}}_{\geq 2}\|_2} \tag{12}$$

Whenever $\rho > 1$, the numerator exceeds the denominator:

$$\|\widehat{\mathbf{R}}_{\geq 2}\|_2 > \|\hat{\boldsymbol{\epsilon}}_1 + \widehat{\mathbf{R}}_{\geq 2}\|_2 \tag{13}$$

For this inequality to hold, the two terms $\hat{\boldsymbol{\epsilon}}_1$ and $\widehat{\mathbf{R}}_{\geq 2}$ must be *mutually opposing* i.e. they must have a negative inner product and $\widehat{\mathbf{R}}_{\geq 2}$ must be the larger of the two in norm. To see this formally, square both sides of Eq. (13) and expand the right-hand side:

$$\begin{aligned} \|\widehat{\mathbf{R}}_{\geq 2}\|^2 &> \|\hat{\boldsymbol{\epsilon}}_1 + \widehat{\mathbf{R}}_{\geq 2}\|^2 \\ &= \|\hat{\boldsymbol{\epsilon}}_1\|^2 + \|\widehat{\mathbf{R}}_{\geq 2}\|^2 + 2\langle \hat{\boldsymbol{\epsilon}}_1, \widehat{\mathbf{R}}_{\geq 2} \rangle \end{aligned} \tag{14}$$

Cancelling $\|\widehat{\mathbf{R}}_{\geq 2}\|^2$ from both sides gives:

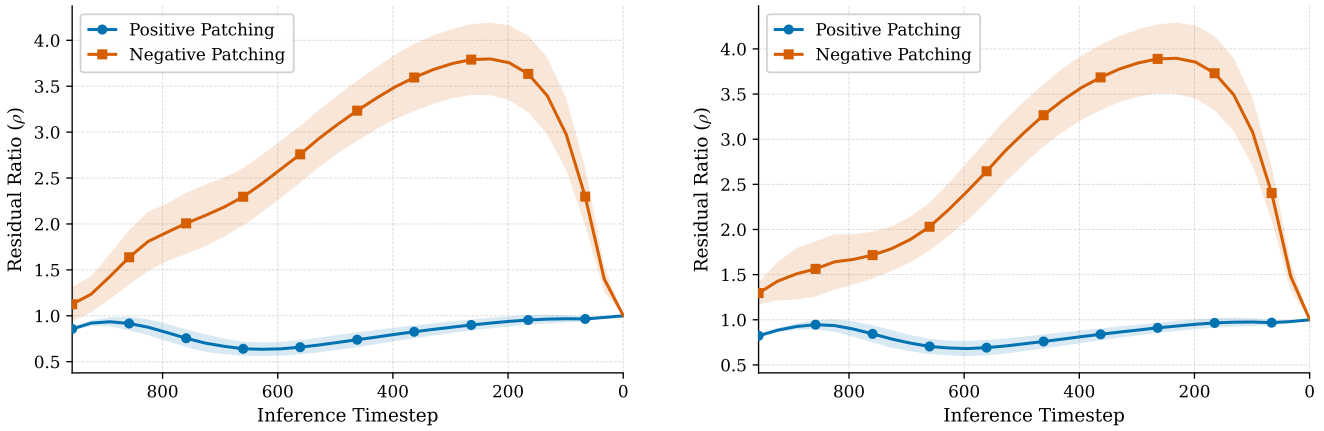
$$0 > \|\hat{\boldsymbol{\epsilon}}_1\|^2 + 2\langle \hat{\boldsymbol{\epsilon}}_1, \widehat{\mathbf{R}}_{\geq 2} \rangle \tag{15}$$

and therefore:

$$\langle \hat{\boldsymbol{\epsilon}}_1, \widehat{\mathbf{R}}_{\geq 2} \rangle < -\tfrac{1}{2}\|\hat{\boldsymbol{\epsilon}}_1\|^2 < 0 \tag{16}$$

Equation (16) establishes two simultaneous conditions. First, the inner product is strictly negative, confirming that $\widehat{\mathbf{R}}_{\geq 2}$ is directionally antagonistic to $\hat{\boldsymbol{\epsilon}}_1$: the nonlinear decoder response actively opposes the intended first-order guidance direction. Second, the magnitude of this opposition is bounded below by $\frac{1}{2}\|\hat{\boldsymbol{\epsilon}}_1\|^2$, which grows with the guidance scale w and patching magnitude η the stronger the intended guidance, the more aggressively the nonlinear residual resists it. Therefore, whenever $\rho > 1$, the decoder's higher-order response is not merely large in magnitude; it is geometrically antagonistic to the linear guidance signal, actively inverting the intended update direction in noise space and destabilising the denoising trajectory.

B.2. Additional Diagnostic Plots



(a) Contrast direction.

(b) Color direction.

Figure 7. Residual ratios for the additional learned directions. As in the blur case, positive patching yields lower entropy than negative patching across most of the denoising trajectory, indicating more coherent skip bottleneck alignment.

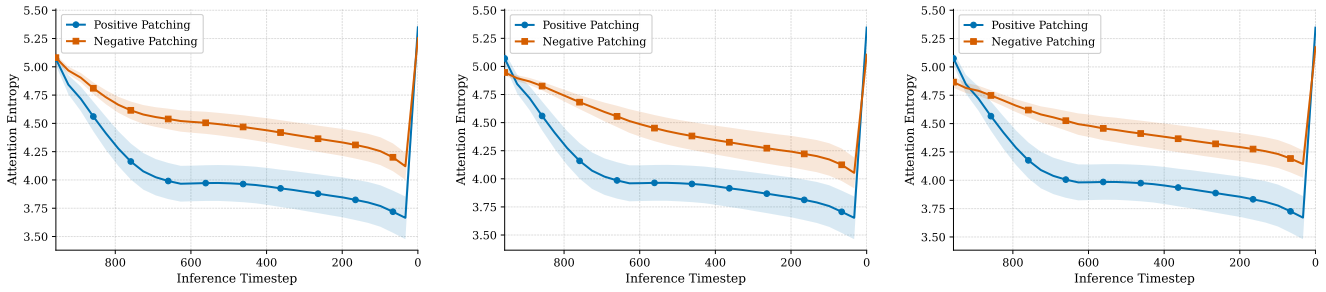
B.3. Attention Entropy Analysis

After bottleneck patching, denoted by $\tilde{\mathbf{h}}_t = \mathbf{h}_t + \eta\,\hat{\Delta}\mathbf{h}_c$, the U-Net decoder upsamples and concatenates this representation with the corresponding encoder skip connections $\{s_i\}$. Since these skip connections are computed from the unpatched encoder, they preserve the features of the original mid-trajectory latent $\mathbf{x}_t$. If the patched bottleneck is poorly aligned with these skip features, the decoder must reconcile inconsistent information.

To quantify the resulting structural mismatch, we measure the attention entropy across denoising time inside the decoder block that contains self-attention in our architecture. Lower entropy indicates a more focused and coherent attention pattern, while higher entropy reflects greater uncertainty.

Positive patching ($\eta > 0$). Patching in the degradation direction yields consistently lower attention entropy throughout the denoising trajectory, indicating better alignment between bottleneck and skip features.

Negative patching ($\eta < 0$). Figure 8 shows persistently higher attention entropy. Injecting the opposite direction creates a mismatch with the skip connections and increases decoder uncertainty.



(a) Sharpness direction. (b) Contrast direction. (c) Saturation direction.

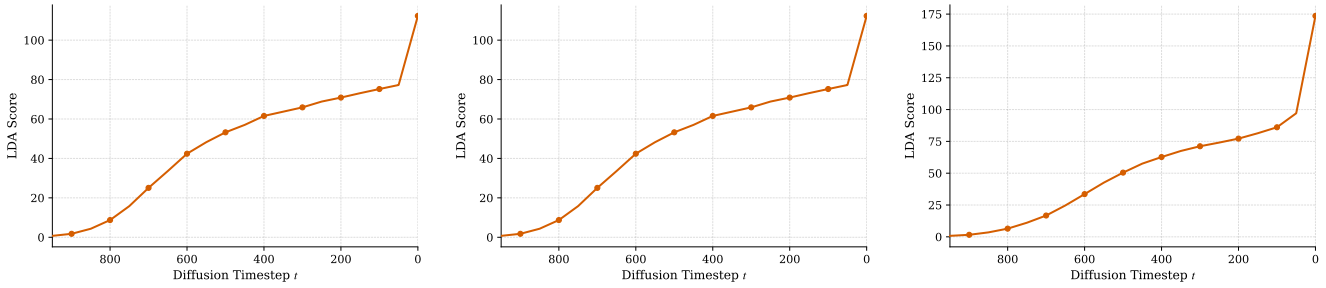
Figure 8. Attention entropy across denoising timesteps for positive and negative bottleneck patching under the sharpness, contrast, and color directions. Across all three directions, negative patching produces higher entropy, indicating less focused attention and poorer skip–bottleneck alignment.

C. Concept Separability Across Timesteps

To identify the denoising timestep at which a given concept is most linearly encoded in *h-space*, we measure the linear separability between clean and degraded bottleneck activations using Linear Discriminant Analysis (LDA) [3]. Concretely, we compute the Fisher criterion $\mathcal{F}(t)$ at each timestep t:

$$\mathcal{F}(t) = \frac{(w^{*\top}(\mu^+ - \mu^-))^2}{w^{*\top} S_W \, w^*} \tag{17}$$

where $w^* = S_W^{-1}(\mu^+ - \mu^-)$ is the optimal discriminant direction, μ^+, μ^- are the class means, and S_W is the within-class



(a) Sharpness direction. (b) Contrast direction. (c) Saturation direction.

Figure 9. LDA scores across extraction timesteps for the sharpness, contrast, and saturation directions.

scatter. A higher $\mathcal{F}(t)$ indicates that the two classes are more discriminable along a linear direction in *h-space* at timestep t. We use $\mathcal{F}(t)$ to identify t^*, the timestep at which concept structure is most cleanly isolated, motivating both our choice of extraction timestep (Sec. 2.1) and the selective guidance window. (Sec. D.3).

D. Ablations

We validate three core aspects of our method: generalization beyond the training domain, applicability to semantic concept directions, and a compute-efficient variant that reduces the overhead of continuous guidance. Unless stated otherwise, all ablations use the same hyperparameters as the main experiments.

D.1. Robustness to Dataset Variation

Our main experiments use Celeba-HQ, a face-specific dataset. To verify that our method is not tailored to face imagery, we test on LSUN Church using the same methodology (Section 2). Figure 10 shows qualitative comparisons between baseline, Standard Patching, and our method with guidance towards direction of degradation on church images.

Figure 10. Qualitative comparison between the Baseline, Standard Patching and our Method on three representative examples on LSUN church with Blur direction.

Our method consistently produces sharper, more detailed images than the baseline, while standard patching with negative direction at matched $|\eta|$ introduces the same structural artifacts observed on CelebA-HQ confirming that the failure mode of negative patching and the benefit of CFG guidance are not artifacts of face-specific *h-space* geometry but reflect a general property of the decoder's directional sensitivity.

D.2. Semantic Concept Directions.

While our primary contribution targets perceptual-quality attributes, our framework is agnostic to the choice of concept direction $\Delta\hat{\mathbf{h}}_c$.

Semantic Concept Directions. While our primary contribution targets perceptual-quality attributes, our framework is agnostic to the choice of concept direction $\Delta\hat{\mathbf{h}}_c$. To validate this, we extract semantic directions from CelebA-HQ attribute labels [15] using a stratified difference-of-means procedure,[7] focused on the SMILING and MALE attributes. Since synthetic degradation cannot be applied to semantic concepts, we follow the paired extraction protocol of [8], computing $\Delta\mathbf{h}_c$ as the mean activation difference between attribute-positive and attribute-negative subsets of CelebA-HQ.

Unlike our perceptual directions where patching toward the degraded concept is necessary to avoid destructive interference semantic directions encode high-level structural attributes that are consistent with the mid-trajectory latent distribution. Negative patching ($\eta < 0$) is therefore stable here, and we apply our method with positive guidance to steer toward the target attribute.

We report CLIP scores [18] with the prompts *"a person with a big smile"* and *"a man"*, respectively, to measure how well the steered generations align with the target concepts.

[7] We maintain equal proportions of positive and negative examples in each subset.

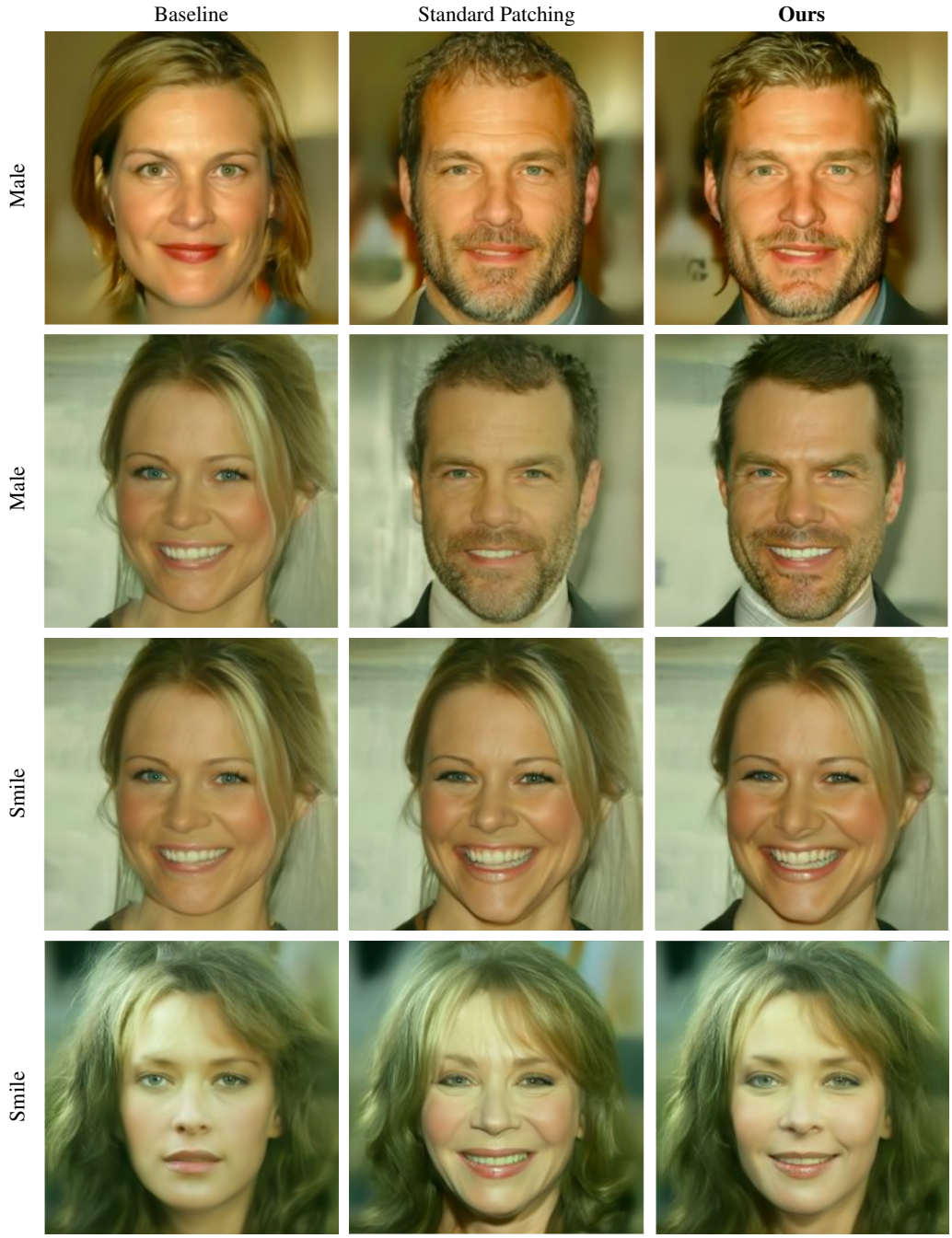


Figure 11. Qualitative comparison between the baseline, Standard Patching and our Method on representative examples with semantically steered directions. This figure shows that our method performs at par, even subjectively better than the patching and it is better at preserving high level features of the image.

Table 2. Semantic steering on CelebA-HQ evaluated via CLIP score. Higher is better.

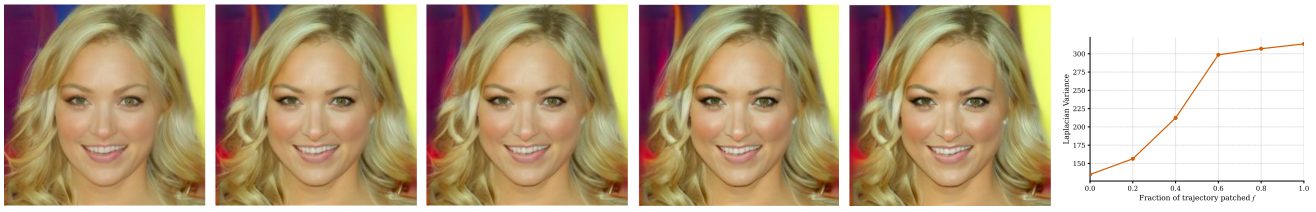
Method	Smiling	Male
Baseline	0.43	0.51
Standard Patching	**0.76**	0.74
Ours	0.73	**0.75**

Our method achieves CLIP alignment comparable to Standard Patching (Table 2) while more consistently preserving high-level image features (Figure 11).

D.3. Timestep-Selective Guidance.

Our method applies patched forward pass at every denoising step, requiring double the compute of standard reverse process. We propose a *partial guidance* variant: for a trajectory of T steps, we run the unpatched baseline for the first $(1-f)$ fraction of steps and switch to guidance only for the final f fraction, at a total cost of $(1+f)$ times a baseline run. This design is motivated by the temporal analysis of linear separability of concepts in *h-space* which were quantised with LDA scores in Section C, peaking at later denoising timestep, indicating that the h-space representation becomes linearly separable with respect to perceptual concepts only as the trajectory approaches the data manifold.

Guidance applied before this regime acts on activations where the concept direction carries little discriminative signal, contributing noise rather than meaningful steering. Figure 12 plots the Laplacian variance obtained for $f \in \{0.2, 0.4, 0.6, 0.8, 1.0\}$, along with the corresponding qualitative examples.



(a) f=0.2 (b) f=0.4 (c) f=0.6 (d) f=0.8 (e) f=1 (f) Lap. var. vs. f

Figure 12. Effect of guidance fraction f. Partial guidance (f=0.6) recovers $\sim$90% of full-guidance quality (Laplacian variance) at only 1.6$\times$ compute vs. 2.0$\times$ for f=1.

These finding confirms that perceptual steering requires intervention only in the late denoising stpes, making our method practical under constrained inference budgets.

E. Hyperparameters

We publicly release our code for reproducibility here. Hyperparameters were tuned manually on a subset of the main dataset. Additionally, details of all hyperparameters have been detailed in the table below.

Hyperparameter	**Value**
Model Configuration	
Model Name	`google/ddpm-celebahq-256`
Main Dataset	CelebA-HQ 256 $\times$ 256
DDIM Inference Steps	30
FID Samples	10,000
Vector Extraction	
Extraction Timestep (k)	50
Number of Samples (N)	100
Guidance Scale (w)	
Our Method	+2.00
Standard Patching	-1.00
Patching Scale (η)	
Our Method	75.00
Standard Patching	-75.00