Interpretable Prompts made Edit-Friendly: Token-to-Token Similarity Reduction in dLLMs for Edit-Friendly Hard Prompt Inversion

Supplementary Material

Contents

1. Additional Qualitative Reconstruction Comparisons with Baselines

In Fig. 2 of the main paper, we provided qualitative reconstruction comparisons on SD 3.5 Medium to highlight that our prompt inversion method generates aligned images while baseline methods miss salient attributes and/or introduce unintended artifacts compared to reference images. Fig. 1 provides additional samples under the same evaluation protocol. **Across a broad set of scenes, our dLLM-based inversion produces prompts that remain highly aligned to the reference image, which in turn yields reconstructions that are consistently more similar to the reference than those from baseline prompts.** Baselines often miss salient entities, introduce spurious attributes, or entangle multiple concepts within a single span, leading to reconstruction drift.

These results reinforce that our method generates interpretable, aligned text prompts that are suitable for downstream text-to-image generation.

2. Image Reconstruction Across Different Text-to-Image Pipelines

We extend reconstruction evaluation beyond Stable Diffusion (SD) 3.5 to SD 2.1, SDXL and FLUX, in order to show the gains in alignment of our method across various text-to-image backbones. For each reference image I, we reuse the same inverted hard prompts produced by our method and baselines, and only swap the downstream generator. This isolates prompt quality from generator choice.

Given an inverted prompt T for a reference image I, we generate a reconstruction $\hat{I}$ with each T2I backbone and measure CLIP-Image similarity:

$$\text{CLIP-I}(I, \hat{I}) = \cos\left(f_I(I), f_I(\hat{I})\right), \tag{1}$$

where f_I is a fixed CLIP image encoder used consistently across all architectures.

Figs. 2, 3, and 4 show qualitative reconstructions on SD 2.1, SDXL, and FLUX, respectively comparing our method with two interpretable baselines (CLIP Interrogator 2.1 [5], and VGD [3]). Across all three backbones, ***our prompts preserve global composition and object identity while remaining robust to architectural changes, including different tokenizers and diffusion inductive biases.*** Baseline prompts degrade more noticeably when transferred, particularly on FLUX, where minor phrasing entanglements can cause strong background or style drift as shown in Fig. 4.

Table 1 summarizes CLIP-I for each backbone.

3. Concept Swap Comparison with VGD Across T2I Pipelines

Our objective is to obtain prompts that support reliable token-level edits to consistent, localized image-level edits. To test generalization, we evaluate **concept swap** across SD 2.1, SDXL, and FLUX. As mentioned in our paper and in section Sec. 6, we perform a token-level swap at the prompt level and utilize this new prompt as input to a text-to-image generation

Figure 1. **Additional qualitative reconstruction comparisons on SD 3.5 Medium.** We compare reconstructions obtained from inverted prompts produced by our method versus all baselines (Captioning [11], BLIP-2 [4], CLIP Interrogator 2.1 [5], PEZ [9], and VGD [3]) using the Stable Diffusion 3.5 Medium backbone. For each reference image, we show the reconstructed images generated from the inverted prompts together with the corresponding prompts. Our method yields more descriptive, structured, and aligned prompts, leading to reconstructions that more faithfully preserve object identity, global layout, and fine-grained attributes relative to baseline inverted prompts.

model to generate a new image (unlike editing where the image before editing is utilized) and then compare concept swap success with the reconstructed image.

Setup. For each reference image I, we invert to a hard prompt $T^\star$ using either VGD or our method. We then replace a target concept token x_i with a new concept $\tilde{x}$ to form:

$$T^{\text{swap}} = \text{Swap}\big(T^\star, x_i \to \tilde{x}\big), \tag{2}$$

and regenerate images with each T2I backbone.

Evaluation. Swap success is measured via:

- **TIFA** [2]: QA-based faithfulness for concept swap.
- **GPT-V**: LLM-as-a-judge scoring for swap correctness and background preservation.

The concept-swap portion of Table 2 (upper block) reports per-backbone TIFA and GPT-V scores for VGD and our method. Figs. 5 and 6 show qualitative CS results on SDXL and FLUX, respectively. ***Across backbones, our prompts yield localized swaps (correct subject/object replacement) with stable preservation of other attributes. VGD prompts exhibit higher leakage, where swapping one token perturbs***

Figure 2. **Reconstruction on SD 2.1.** Reconstructions from the same inverted prompts using the SD 2.1 backbone, comparing our method against CLIP Interrogator and VGD.

Table 1. Image reconstruction quality across different T2I backbones. We report CLIP-Image similarity (CLIP-I; ↑) between reference images and reconstructions generated from inverted prompts.

Method	CLIP-Image Similarity ↑		
	SD 2.1	SDXL	FLUX
Captioning [11]	0.46	0.49	0.56
BLIP-2 [4]	0.49	0.51	0.57
CLIP Interrogator [5]	0.49	0.51	0.59
PEZ [9]	0.58	0.59	0.65
VGD [3]	0.60	0.62	0.67
Ours	**0.64**	**0.65**	**0.68**

background, lighting, or secondary objects, and such effects can be consistently observed across SD 2.1, SDXL and FLUX.

Figure 3. **Reconstruction on SDXL.** Reconstructions from identical inverted prompts transferred to SDXL without re-inversion.

4. Concept Append Comparison with VGD Across T2I Pipelines

We next evaluate **concept append** to test whether prompts remain modular under additions, again transferring identical prompts across SD 2.1, SDXL, and FLUX.

Setup. Starting from $T^\star$, we append a phrase P_{append} (style or secondary concept):

$$T^{\text{append}} = \text{Concat}\big(T^\star, P_{\text{append}}\big), \tag{3}$$

and regenerate with each backbone.

Evaluation. Append success is measured by:

- **TIFA**: QA-based faithfulness for concept append.
- **GPT-V**: llm-as-a-judge-based correctness and preservation scoring.

Table 2 (lower block) reports per-backbone TIFA and GPT-V scores for concept append. Figs. 7 and 8 show qualitative Concept Append results on SDXL and FLUX. Our

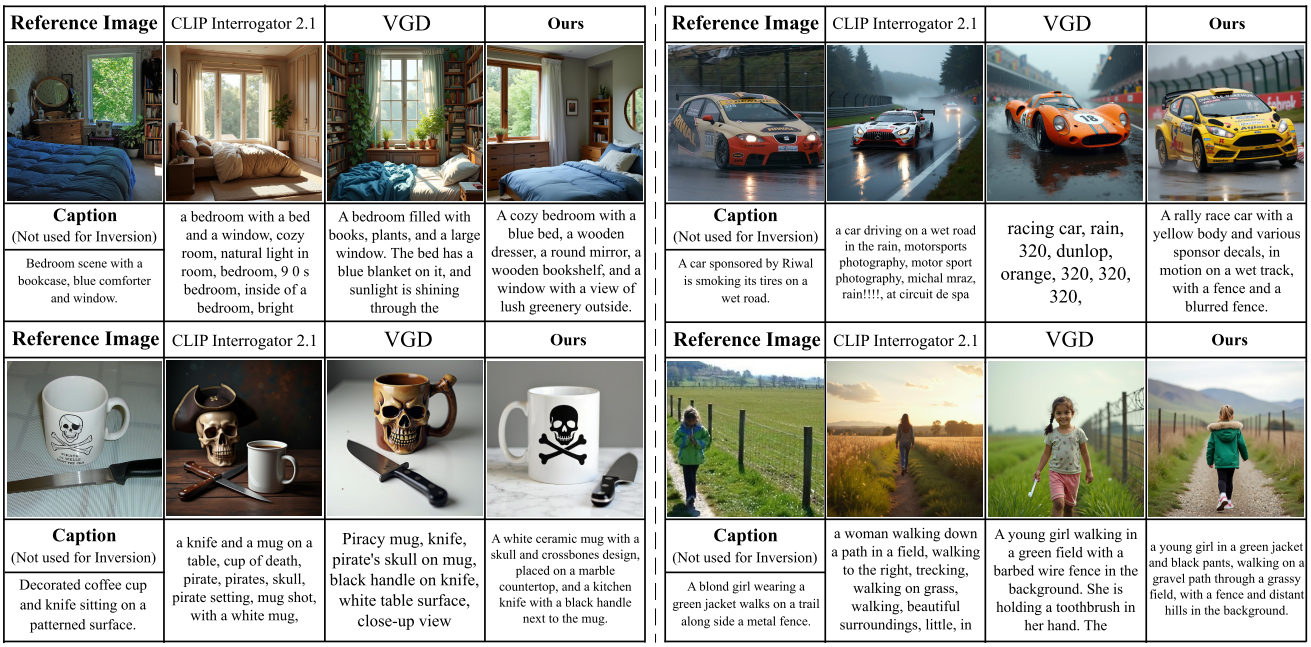

Figure 4. **Reconstruction on FLUX.** Reconstructions using the FLUX backbone from the same inverted prompts, demonstrating strong cross-architecture robustness of our prompts.

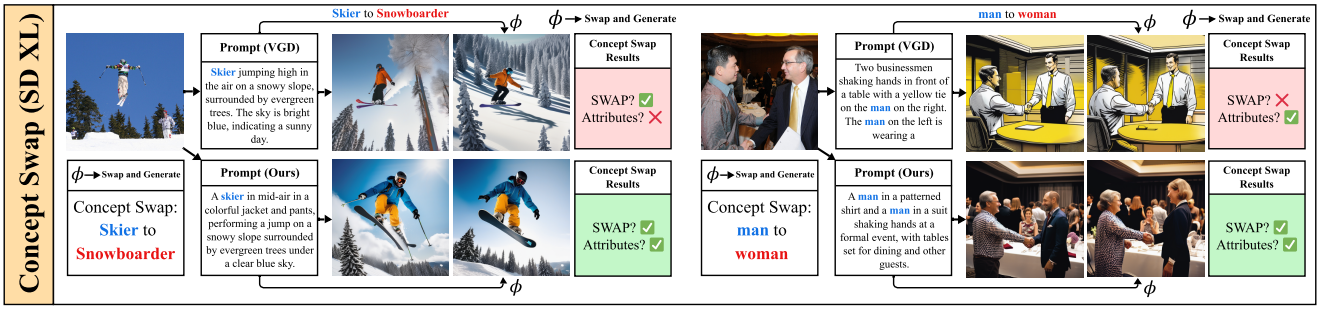

Figure 5. **Concept swap on SDXL.** Swapped generations from VGD versus our edit-friendly prompts. Our method yields localized swaps while preserving background and non-affected attributes.

prompts reliably incorporate new content (objects shown, styles evaluated in Tab. 3) while preserving the base scene. VGD frequently under-applies the append or causes broader drift, reflecting higher token entanglement.

5. User Study Interface and Results

We conducted four controlled user studies to validate interpretability, reconstruction quality, and editability under concept swap and append, comparing to the strongest interpretable hard-prompt baseline, VGD [3]. Each study uses **200 samples** and **25 users**, with randomized anonymized A/B trials.

All studies share a common web-based interface: a landing page explaining the task with an example trial, followed

Table 2. Concept swap (CS) and concept append (CA) evaluation across architectures. We report TIFA and GPT-V scores (↑) per backbone for VGD and our method.

Method	SD 2.1		SDXL		FLUX	
	TIFA ↑	GPT-V ↑	TIFA ↑	GPT-V ↑	TIFA ↑	GPT-V ↑
CS (Concept Swap)						
VGD	0.63	0.69	0.65	0.72	0.71	0.74
Ours	**0.83**	**0.83**	**0.89**	**0.85**	**0.90**	**0.91**
CA (Concept Append)						
VGD	0.61	0.70	0.66	0.71	0.69	0.73
Ours	**0.83**	**0.82**	**0.87**	**0.88**	**0.89**	**0.90**

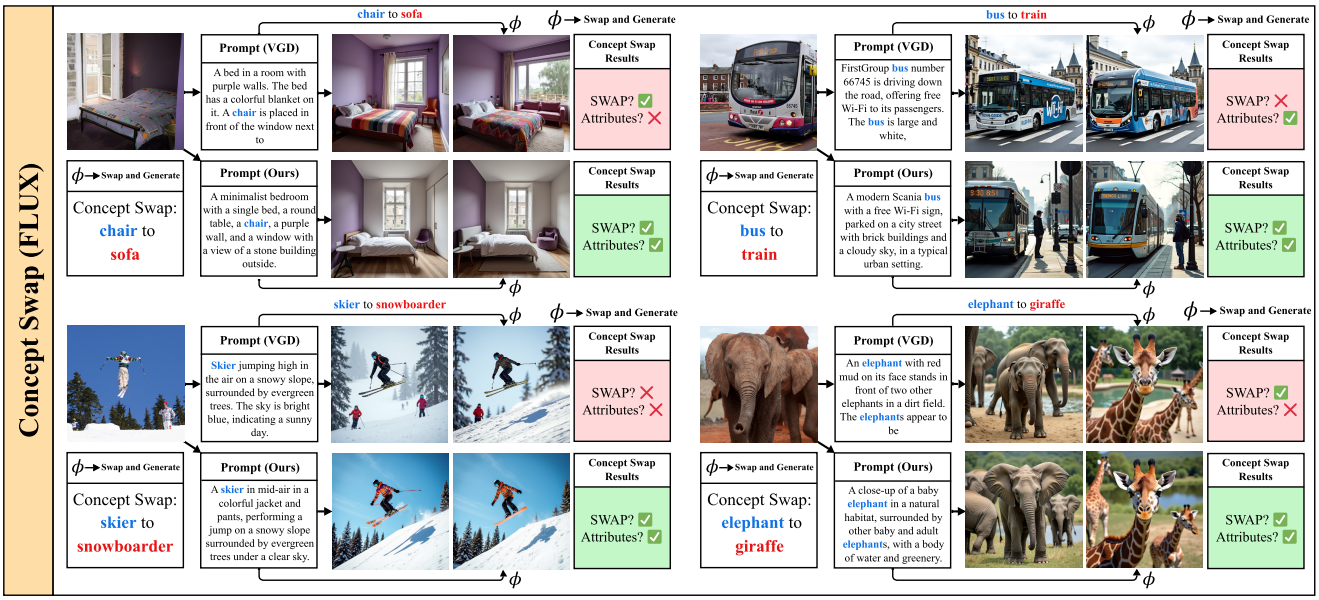

Figure 6. **Concept swap on FLUX.** The same swapped prompts applied on FLUX. Our prompts generalize across architectures, producing faithful swaps with minimal leakage.

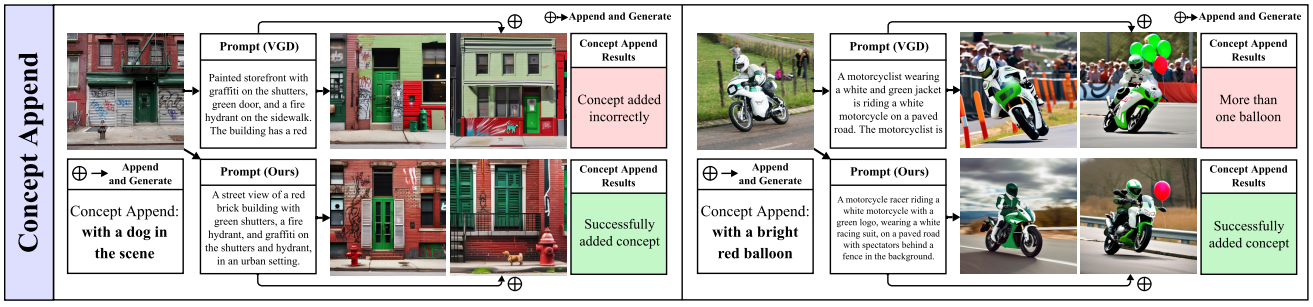

Figure 7. **Concept append on SDXL.** Appending new objects or styles to inverted prompts and regenerating on SDXL. Our prompts support faithful additions while preserving original scene structure.

by randomized A/B trials where participants choose between two anonymized options (A/B) or indicate no preference.

5.1. Text Quality User Study

This study evaluates prompt interpretability and semantic alignment. Participants are shown two inverted prompts corresponding to the same reference image (VGD vs. Ours, identity hidden) and answer: *"Which of the two prompts provides a coherent, interpretable description of the objects, attributes, background, and style in the input image so that the original image could be reconstructed?"* Users consistently prefer our prompts, citing better structure, clearer entity descriptions, and reduced ambiguity. We provide our results from our user study in Fig. 4 in our main paper.

5.2. Image Reconstruction User Study

Participants see a reference image and two reconstructions (A/B) generated from inverted prompts of different methods, and answer: *"Which image better matches the reference image?"* Preferences align with our CLIP-I trends (Table 1), favoring reconstructions from our prompts.

5.3. Concept Swap User Study

For each trial, participants see the reference image and a swap instruction (e.g., *"swap horse to zebra"*) with two swapped generations (A/B). They answer: *"Which edited image better satisfies the requested concept change while preserving the rest of the scene?"* Users prefer our swaps due to improved localization and lower background drift, consistent with the Concept Swap block of Table 2.

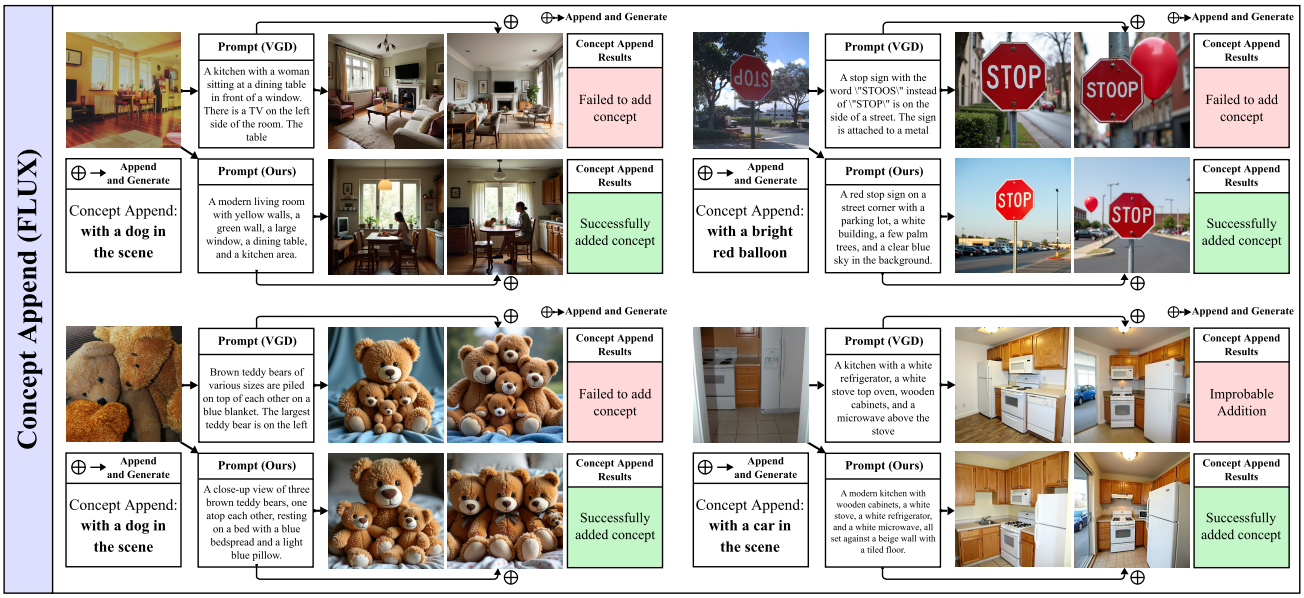

Figure 8. **Concept append on FLUX.** The same appended prompts applied on FLUX. Our method remains robust, enabling successful additions with minimal disruption to non-edited content.

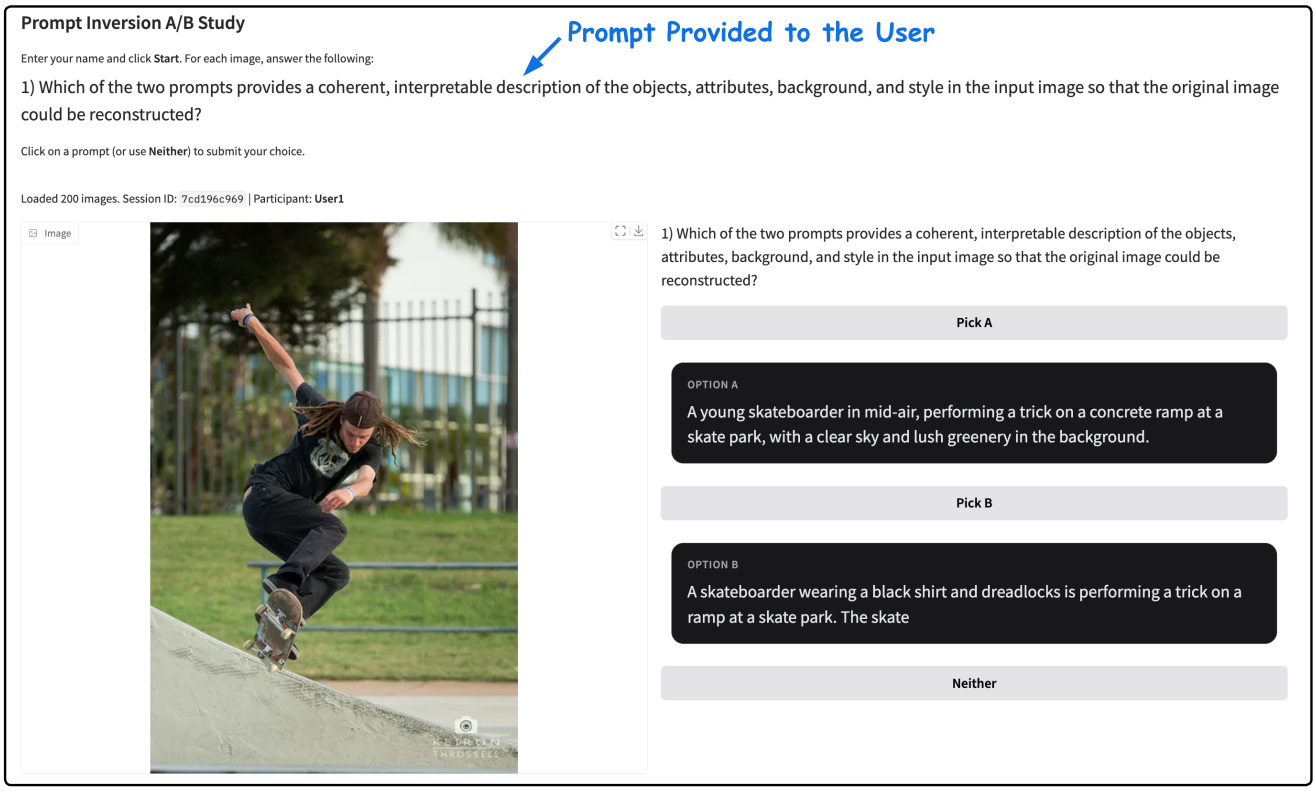

Figure 9. **UI for text quality user study.** Participants see two inverted prompts (A/B) for the same reference image with method identities hidden, and select the prompt that is more interpretable and better aligned with the image.

5.4. Concept Append User Study

Participants see an append instruction (style or object addition) and two appended generations (A/B), and answer:

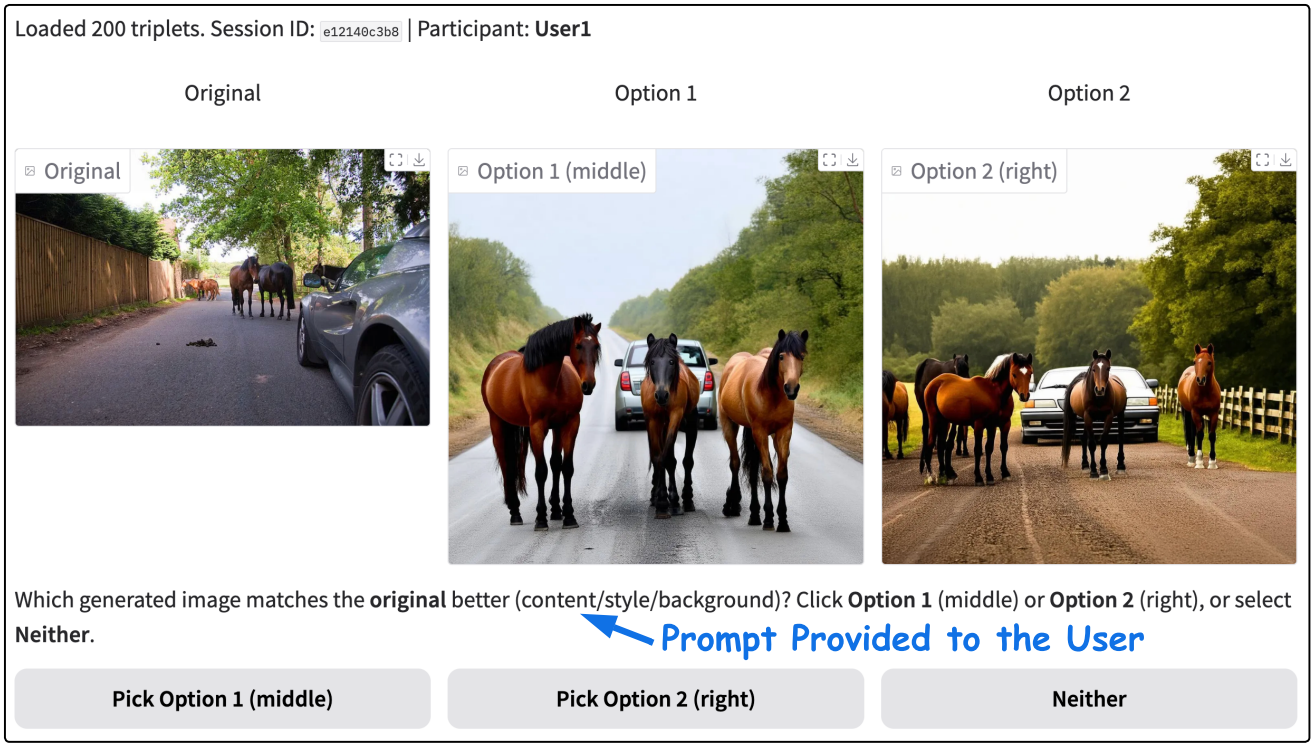

Figure 10. **UI for reconstruction user study.** Participants see the reference image and two reconstructions (A/B) generated from different inverted prompts, and choose the reconstruction that better matches the reference image.

"Which image better satisfies the requested concept/style addition while preserving the original scene?" Our method is preferred in the majority of trials, supporting the Concept Append block of Table 2 and the style-consistency trends in Section 8.

6. Additional Details on Concept Swap and Concept Append Pipelines

We clarify the prompt-level edit pipeline used throughout the paper. Swap/append are treated as *text edits only*; the generator never receives an edited image.

Pipeline description. Given a reference image I, a T2I model $\mathcal{G}$, and inversion method $\mathcal{F}$ (ours or VGD):

1. **Prompt inversion.**

$$T^\star = \mathcal{F}(I). \tag{4}$$

2. **Base reconstruction.**

$$\hat{I} = \mathcal{G}(T^\star). \tag{5}$$

$\hat{I}$ is used only for evaluation and visualization.

3. **Concept swap.** Select a concept token x_i and a new concept $\tilde{x}$, and form

$$T^{\text{swap}} = \text{Swap}(T^\star, x_i \to \tilde{x}), \qquad \hat{I}^{\text{swap}} = \mathcal{G}(T^{\text{swap}}). \tag{6}$$

4. **Concept append.** Choose an append phrase P_{append} (style or object addition) and form

$$T^{\text{append}} = \text{Concat}(T^\star, P_{\text{append}}), \qquad \hat{I}^{\text{append}} = \mathcal{G}(T^{\text{append}}). \tag{7}$$

Difference from image editing. Unlike Prompt-to-Prompt [1], or other image editing techniques, we do not condition on I or $\hat{I}$ during image generation. Edit success is therefore a direct test of modularity of $T^\star$: localized text edits should induce localized, predictable changes in the generated image. We evaluate against $\hat{I}$ to isolate (i) edit correctness and (ii) preservation of non-edited content. Our token-similarity reduction objective in dLLM is tailored to encourage precisely this behavior.

7. Implementation Details of the LLaDA dLLM

We provide additional implementation details for the discrete diffusion language model (dLLM) used in our inversion procedure, referred to as **LLaDA**.

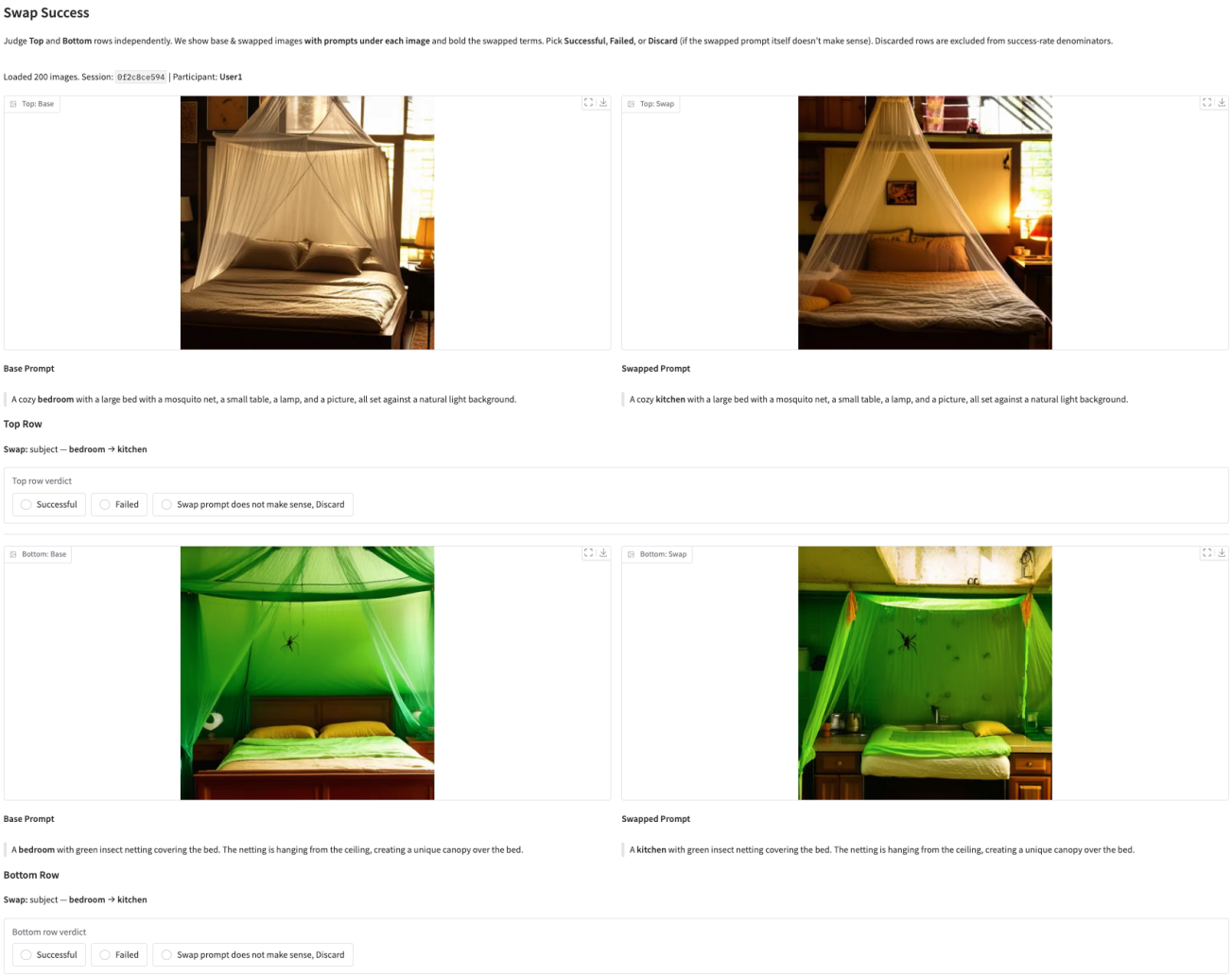

Figure 11. **UI for concept swap user study.** Participants see the reference image, the swap instruction, and two swapped generations (A/B), and choose the image that best satisfies the requested swap while preserving the rest of the scene.

7.1. Generation Schedule

We use a mask-and-refill generative process with the following hyperparameters:

- **Number of steps:** $T = 128$ reverse diffusion steps.
- **Prompt length:** `gen_length` $= 32$ tokens (main experiments), adjusted elsewhere to match budgets (16, 32, 64, 77).
- **Block length:** `block_length` $= 32$ (parallel refinement of all positions).
- **Temperature:** deterministic sampling with `temperature` $= 0.0$.
- **Mask token:** dedicated mask ID `mask_id = 126336`.
- **Stopping criterion:** stop on $<|$`eot_id`$|>$ after at least `min_out_tokens_before_stop` $= 6$ tokens.

7.2. Online CLIP Steering and Token-Token Similarity Reduction

We apply FK steering [7] using CLIP alignment and token-similarity regularization:

- **CLIP configuration:** ViT-H/14.
- **Guidance frequency:** every step (`guidance_interval = 1`), with `guidance_positions_per_step = 12` and top-k candidates `guidance_topk = 4`.
- **CLIP tilt:** `guidance_lambda = 40.0`.
- **Token-similarity penalty:** `overlap_penalty_weight = 25.0`, `global_decorrelate = true`, `global_decorrelate_lambda = 4.0`, `lookback_k = 48`, `decorrelation_step_eta = 0.8`.

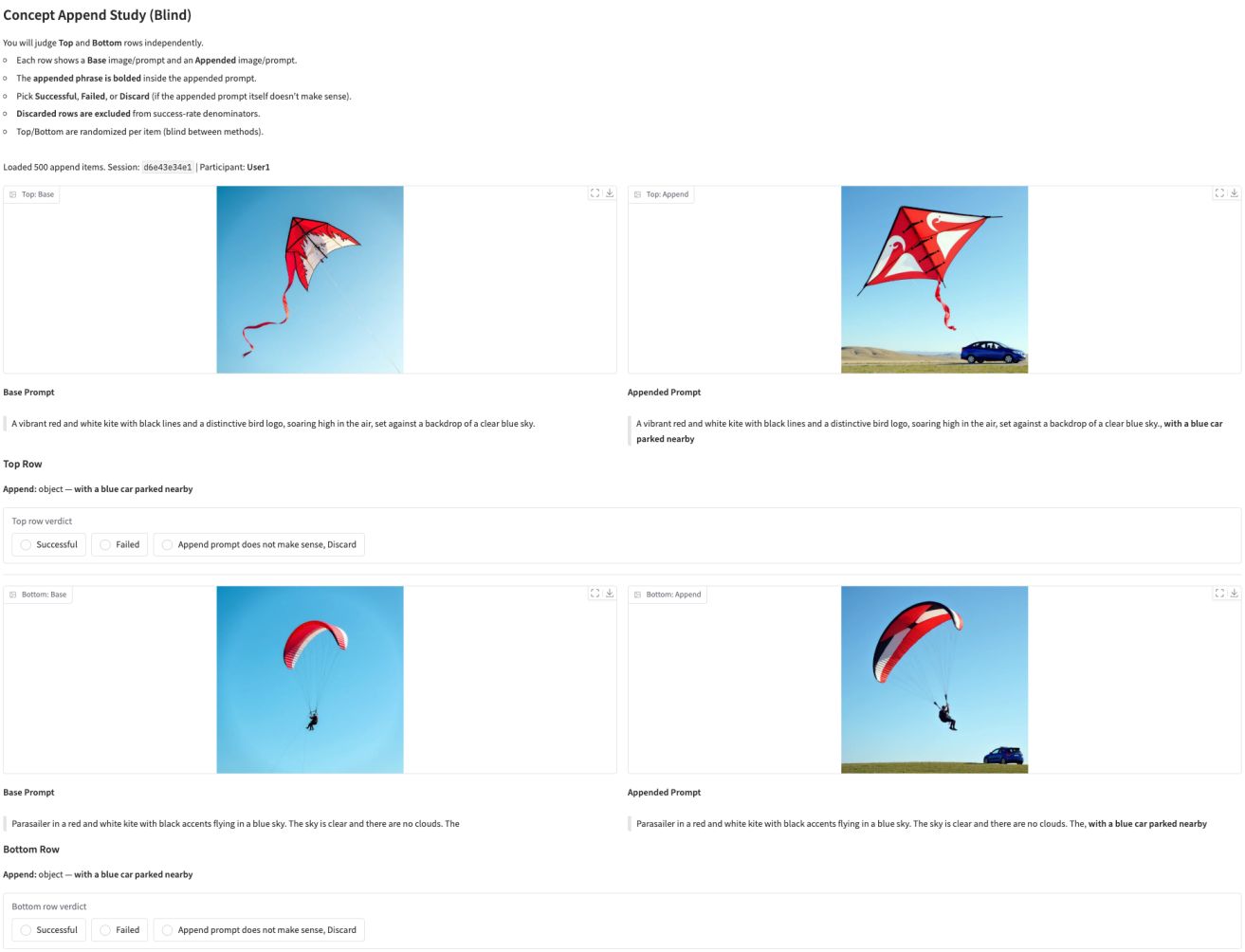

Figure 12. **UI for concept append user study.** Participants see the reference image, the append instruction, and two appended generations (A/B), and select the image that best adds the requested concept/style while keeping non-edited content intact.

- **Local temperature bump:** `local_temp_bump = 0.15`.

This configuration implements the joint objective (Eq. 12 in the main paper), tilting the final prompt distribution toward both CLIP alignment and edit-friendliness while preserving fluency through the dLLM prior.

8. Style Append Success Rate

We evaluate *style append* success using a style-consistency score **CSD** [8]. In our setting, Higher CSD indicates better style realization without disrupting base content.

For each method:

1. Invert I to obtain $T^\star$.
2. Append a style phrase to form

$$T^{\text{append}} = \text{Concat}(T^\star, P_{\text{append}}). \tag{8}$$

3. Generate $\hat{I}^{\text{append}}$ from T^{append}.

Table 3. Style append success rate via CSD [8] (style-consistency score; ↑). Higher values indicate better realization of the requested style while preserving scene content.

Method	CSD % ↑
CLIP Interrogator	0.74
VGD	0.73
Ours	**0.82**

4. Compute CSD between the requested style and $\hat{I}^{\text{append}}$.

We threshold CSD and calculate the % of successful style addition and report our results in Tab. 3.

Qualitatively, our prompts support style append operations that overlay the requested style (color palette, texture, lighting) on top of the original composition, whereas baseline prompts either under-encode the style or over-encode it

Table 4. Multi-step regeneration robustness. CLIP-Image similarity (CLIP-I; ↑) between $I^{(0)}$ and the image after k rounds of inversion+regeneration. Stable Diffusion 3.5 Medium Model used for text-to-image generation.

Method	$k=0$	$k=1$	$k=2$	$k=3$
VGD	0.67	0.63	0.58	0.52
Ours	**0.71**	**0.68**	**0.63**	**0.57**

in a way that disrupts object identity or background structure.

9. Multi-Step Regeneration Analysis

We evaluate robustness under repeated inversion and regeneration. Starting from $I^{(0)}$, we repeat four cycles:

$$T^{(k)} = \mathcal{F}(I^{(k)}),$$
$$I^{(k+1)} = \mathcal{G}(T^{(k)}), \quad k = 0, 1, 2, 3,$$

where $\mathcal{F}$ is ours or VGD and $\mathcal{G}$ is fixed (SD 2.1 in this experiment).

After each cycle k, we measure similarity to $I^{(0)}$:

$$\text{CLIP-I}^{(k)} = \cos\left(f_I(I^{(0)}), f_I(I^{(k)})\right). \tag{9}$$

Table 4 reports CLIP-I over four rounds. Our method remains substantially more stable across cycles, while VGD accumulates drift.

This multi-step regeneration experiment reinforces our central claim: by explicitly reducing token-to-token similarity during inversion, we obtain prompts that are not only more interpretable and edit-friendly, but also more robust under repeated use in iterative workflows where prompts and images are alternately inverted and regenerated.

10. Comparison to Additional Prompt Optimization Methods

Two recent works, PromptSculptor [10] and STEPS [6], also study prompt optimization for text-to-image models, but they address a different problem setting and optimize a different set of objectives than our method. PromptSculptor proposes a multi-agent LLM framework that takes a short user *text* prompt and iteratively expands it into a longer, more detailed description, using a self-evaluation and feedback loop to improve CLIP/PickScore alignment and aesthetics of the generated images; it does not invert from an *image* to a prompt, does not target faithful reconstruction of a specific reference image, and does not explicitly encourage token-level disentanglement or editability of individual concepts. STEPS [6] formulates hard prompt search as a sequential probability tensor estimation problem on discrete tokens, optimizing CLIP similarity to a target representation via a low-rank tensor (TT) parameterization; while effective for discovering high-scoring prompts for a fixed encoder-generator pair, STEPS also does not seek interpretable, edit-friendly prompts, and does not evaluate localized concept swap/append behavior, cross-architecture edit transfer, or human preference.

Table 5. Mean off-diagonal similarity (O.D.; ↓) and edit success (TIFA; ↑), averaged across datasets at 32 tokens.

Method	O.D. ↓	TIFA ↑
VGD	0.79	0.66
Ours	**0.36**	**0.87**

In contrast, our approach is explicitly designed as a *gradient-free interpretable, and edit-friendly hard prompt inversion* method from image to text, built on a discrete diffusion language model (LLaDA) with online CLIP steering and token-similarity regularization. This design jointly optimizes (i) image reconstruction fidelity, (ii) prompt interpretability, and (iii) token-level disentanglement that enables localized concept swap and concept append operations. Moreover, we demonstrate architecture-agnostic transfer of the same inverted prompts across SD 2.1, SDXL, FLUX, and SD 3.5 Medium, and provide both automatic (CLIP-I, TIFA, GPT-V, CSD) and controlled user-study evaluations on text quality, reconstruction, and editing behavior. Empirically, our method consistently outperforms the strongest interpretable hard-prompt baselines (including VGD) on these metrics, while operating in a more challenging inversion-and-edit setting that is not addressed by PromptSculptor [10] or STEPS [6].

10.1. Additional Quantitative Evidence for Token Disentanglement

To complement the token-token similarity heatmaps in the main paper, we report the mean off-diagonal similarity of the token similarity matrix as a scalar summary of token entanglement. Lower off-diagonal similarity indicates that token positions play more modular roles during decoding, which is desirable for localized downstream edits. As shown in Table 5, our method achieves substantially lower off-diagonal similarity than VGD while also improving edit success measured by TIFA. This supports the claim that reducing token overlap during inversion yields prompts that are both more disentangled and more edit-friendly.

10.2. Scaling with Evaluation Size and Token Budget

We use 200 images per dataset in the main evaluation, which is larger than the 100-image protocol used in prior prompt inversion baselines such as VGD and PEZ. This larger evaluation provides a more reliable estimate of prompt quality while remaining practical because our method substantially

Table 6. Scaling with dataset size and token budget. We report CLIP-Text similarity (↑) and perplexity (PPL; ↓).

Images, Tokens	CLIP-Text ↑	PPL ↓
600, 77	0.61	73.91
1200, 77	0.60	74.04
1200, 150	0.58	82.29

reduces inversion time. We also examine behavior beyond the standard 77-token setting to test whether the method remains stable at longer prompt lengths. Table 6 shows that text quality remains stable as the evaluation set is increased, and that longer-token runs remain usable without collapse, although very long prompts trade some fluency for additional descriptive capacity.

Choice of editability proxy. A natural alternative to our approach would be to keep a full text-to-image diffusion model inside the inversion loop and optimize editability using image-space signals such as cross-attention separation or edit consistency after regeneration. While such objectives can be informative, they are computationally expensive, tightly coupled to a specific generator, and require repeated access to high-dimensional intermediate states during optimization. We instead use a lightweight, training-free, text-side proxy derived from the dLLM predictive distributions. This design preserves model agnosticism, adds little computational overhead, and integrates naturally into our FK-steered decoding framework while still improving downstream edit behavior, alignment, and fluency.

Use of CLIP-based alignment metrics. We use CLIP-based similarity because it is the standard image-text alignment metric adopted throughout prior prompt inversion work and provides a common basis for comparison across captioning, hard inversion, and hybrid baselines. In our evaluation, CLIP is used together with stronger downstream metrics such as TIFA, GPT-V-based judgments, and controlled user studies, so our conclusions do not rely on CLIP alone. We therefore treat CLIP similarity as one component of a broader evaluation suite that jointly measures reconstruction fidelity, prompt interpretability, and edit success.

Effect of token budget on reconstruction detail. The degree of attribute preservation depends in part on the available token budget. Short prompts often capture the dominant objects and scene layout, while longer prompts can encode additional fine-grained details such as secondary attributes, interactions, and small objects. In practice, we observe that increasing the token budget yields more descriptive prompts and more faithful reconstructions, which helps preserve subtle attributes that may be under-specified at shorter lengths.

References

[1] Amir Hertz, Ron Mokady, Neta Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross-attention control. *arXiv preprint arXiv:2208.01626*, 2022. 7

[2] Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. *arXiv preprint arXiv:2303.11897*, 2023. 2

[3] Donghoon Kim, Minji Bae, Kyuhong Shim, and Byonghyo Shim. Visually guided decoding: Gradient-free hard prompt inversion with language models. In *International Conference on Learning Representations (ICLR)*, 2025. 1, 2, 3, 4

[4] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023. 2, 3

[5] pharmapsychotic. Clip interrogator 2.1, 2022. Image-to-prompt tool combining BLIP and CLIP; includes v2.1 release/Colab links. 1, 2, 3

[6] Yuning Qiu, Andong Wang, Chao Li, Haonan Huang, Guoxu Zhou, and Qibin Zhao. Steps: Sequential probability tensor estimation for text-to-image hard prompt search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 28640–28650, 2025. 10

[7] Raghav Singhal, Zachary Horvitz, Ryan Teehan, Mengye Ren, Zhou Yu, Kathleen McKeown, and Rajesh Ranganath. A general framework for inference-time scaling and steering of diffusion models. *arXiv:2501.06848*, 2025. Feynman–Kac (FK) Steering. 8

[8] Gowthami Somepalli, Anubhav Gupta, Kamal Gupta, Shramay Palta, Micah Goldblum, Jonas Geiping, Abhinav Shrivastava, and Tom Goldstein. Measuring style similarity in diffusion models, 2024. 9

[9] Yuxin Wen, Yiran Li, Sifei Liu, Xiaolong Wang, et al. Hard prompts made easy: Gradient-based discrete optimization for prompt tuning and inversion. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 2, 3

[10] Dawei Xiang, Wenyan Xu, Kexin Chu, Tianqi Ding, Zixu Shen, Yiming Zeng, Jianchang Su, and Wei Zhang. Promptsculptor: Multi-agent based text-to-image prompt optimization, 2025. 10

[11] Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. Florence-2: Advancing a unified representation for a variety of vision tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2, 3