MOONWALKER: LEARNING ACCURATE INVERSION STEPS IN DIFFUSION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Standard DDIM inversion attempts to recover the noise underlying an image, but relies on a local approximation of the reverse process whose errors accumulate across steps. As a result, the recovered latent remains entangled with information from the source image. We introduce Moonwalker, a lightweight solution that corrects this approximation by learning "forward" inversion directly from backward generation steps. Specifically, we train a LoRA adapter to take a cleaner latent and predict the exact previous noise from the sampling trajectory, perfectly retracing the path to the correct noisier state. The adapter is activated only during inversion and deactivated during generation, leaving the original diffusion model and sampling process unchanged and incurring no inference overhead. Moonwalker reduces the correlation between recovered latents and their source images, and we evaluate its effect on downstream performance across image content, style, and music editing benchmarks.

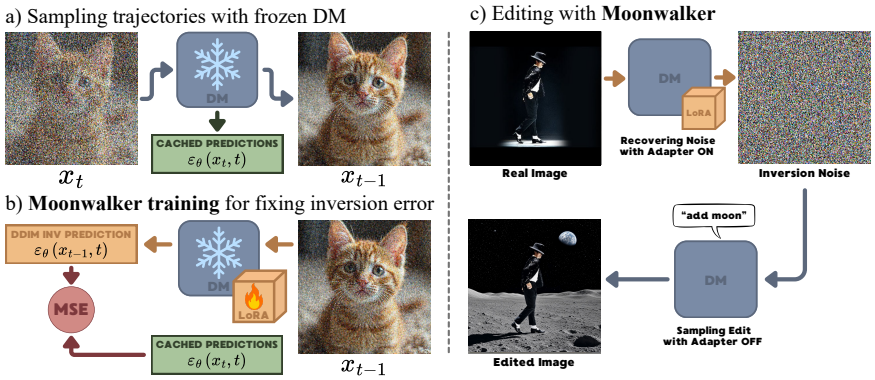


Figure 1: We introduce Moonwalker – a method for fixing DDIM Inversion error through low-rank adaptation.

1 INTRODUCTION

Diffusion models have demonstrated unprecedented capabilities in image generation (Sohl-Dickstein et al., 2015; Dhariwal & Nichol, 2021), but manipulating existing images remains a significant challenge. Because diffusion generation relies on a multi-step iterative process rather than a well-behaved latent space, editing an existing image first requires inversion – mapping the image back to an initial noise vector. A perfectly recovered noise representation is critical, as it serves as the foundation for any subsequent structural or stylistic modifications.

A common approach to inversion is based on Denoising Diffusion Implicit Models (DDIMs) (Song et al., 2021), which define generation as a deterministic trajectory and thus allow the process to be approximately reversed to recover the initial noise. However, the reverse update used by standard DDIM inversion is not an exact inverse of the forward process. Instead, it relies on a local linear approximation that reuses the model's noise prediction across consecutive timesteps. This approximation introduces an error at each inversion step, causing the recovered trajectory to progressively diverge from the one followed during generation.

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.

As a result, the recovered latent does not statistically resemble the underlying Gaussian noise and retains the spatially correlated structure of the input image (Staniszewski et al., 2026). Such deviations compromise the suitability of the inverted latent for subsequent image editing, since image content becomes entangled with the representation intended to encode noise. Existing approaches often address the reconstruction errors through additional stabilization mechanisms (Cao et al., 2023b; Ju et al., 2024a), but the underlying trajectory mismatch remains inherent to the inversion procedure.

In this work, we propose to exactly solve the error introduced by DDIM inversion by directly adapting the model to align with its true trajectory predictions. We call our approach Moonwalker because, as presented in fig. 1, it achieves precise backward movement by mimicking forward steps. First, we sample trajectories using the frozen model, gathering the ground-truth noise predicted at each step as the image becomes cleaner. We then train a LoRA adapter to take these cleaner latents and predict the exact noise gathered by the model in the noisier setup. Because this adapter is applied solely during inversion, the original generation pipeline remains entirely unchanged. Furthermore, Moonwalker is highly efficient: it requires only LoRA finetuning on trajectories from sample generations, relies on no external data, and introduces zero additional inference costs.

By fundamentally correcting the inversion trajectory, our technique successfully removes spatial correlations from the observed latents, resulting in recovered noise that matches the theoretical prior. This foundational correction guarantees superior baseline editability. Through extensive evaluation on existing benchmarks, we demonstrate that delivering mathematically clean latents to existing downstream editors yields state-of-the-art stability and consistent performance gains in tasks such as style transfer and structural editing, proving that better inversion natively unlocks better editing.

Our contribution can be summarized as follows:

- We introduce the Moonwalker, a lightweight, trajectory-supervised LoRA adapter that corrects DDIM linear approximation errors introduced during inversion
- We demonstrate that exact trajectory alignment successfully eliminates spatial correlations from the latents, recovering unstructured noise that matches the theoretical Gaussian prior.
- We show that correcting inverted latents improves the editing performance of existing state-of-the-art techniques for image editing, style transfer, and music editing.

2 Background & related work

Diffusion models. The diffusion framework formulates the generative process as a transition between two distributions: training data $\mathcal{X}$ (e.g., image and audio samples) and Gaussian $\mathcal{N}(0, \mathcal{I})$. Diffusion Models (DMs) are trained to denoise, i.e., to reverse a process of gradually corrupting data with noise, referred to as *forward diffusion*. Given a pair of samples $\big(x_0 \sim \mathcal{X}; \epsilon \sim \mathcal{N}(0, \mathcal{I})\big)$ from both distributions, such a process interpolates between them as $x_t = \sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{1 - \bar{\alpha}_t}\, \epsilon$, given $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$ and $\alpha_t = 1 - \beta_t$, where $\{\beta_t\}_{t=1}^{T}$ is a noise variance schedule.

For Denoising Diffusion Implicit Models (DDIMs, Song et al. (2021)), the sampling starts from random noise $x_T \sim \mathcal{N}(0, \mathcal{I})$ and conditioning c and produces a clean sample x_0 by iteratively, in T steps ($x_T \to x_{T-1} \to \cdots \to x_1 \to x_0$), subtracting the noise predicted by trained model ϵ_θ:

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}} \cdot x_t - \Big(\frac{\sqrt{1 - \bar{\alpha}_t}}{\sqrt{\alpha_t}} - \sqrt{1 - \bar{\alpha}_{t-1}}\Big) \cdot \epsilon_\theta(x_t, t, c). \tag{1}$$

Notably, the process of obtaining a less noisy sample (x_{t-1}) from a noisier one (x_t) is deterministic, providing a fixed mapping between the two distributions. Such determinism makes the generative process reversible in principle: any arbitrary image can be encoded back to its corresponding noise by rearranging terms in eq. (1) and running it backward in time ($x_0 \to x_1 \to \cdots \to x_{T-1} \to x_T$):

$$x_t = \sqrt{\alpha_t}\, x_{t-1} + \Big(\sqrt{1 - \bar{\alpha}_t} - \sqrt{\alpha_t - \bar{\alpha}_t}\Big) \cdot \epsilon_\theta(x_t, t, c). \tag{2}$$

This process, referred to in the literature as *inversion*, lays the foundation for numerous methods for downstream tasks, such as editing, style transfer, or content watermarking. However, directly applying the update from eq. (2) is infeasible in a real scenario, as computing x_t requires the model prediction at x_t itself. Thus, standard DDIM Inversion resolves this circular dependency through local approximation, evaluating the model at the available (cleaner) state instead:

$$\epsilon_\theta(x_t, t, c) \approx \epsilon_\theta(x_{t-1}, t, c). \tag{3}$$

Recent studies have analyzed the consequences of this approximation. Accumulated over a sequence of steps, it causes the recovered noise to deviate statistically from the Gaussian prior and entangle the structure of the input image (Parmar et al., 2023a; Garibi et al., 2024; Lin et al., 2024; Staniszewski et al., 2026; Jiao et al., 2026) or significantly degrade downstream editing, especially when employed in settings with a higher guidance scale (Ho & Salimans, 2021; Mokady et al., 2023). Staniszewski et al. (2026) trace the problem to approximation error at the first inversion steps at low-frequency image areas, whereas Lin et al. (2024) attribute it to the singularity problem in the noise schedule.

Inversion methods. Despite its approximation error, DDIM Inversion forms the basis for numerous applications, including image morphing (Dhariwal & Nichol, 2021; Samuel et al., 2023; Zheng et al., 2024; Zhang et al., 2024a; Bodin et al., 2025), content watermarking (Wen et al., 2023; Hong et al., 2024; Ci et al., 2024; Zhang et al., 2024b; Gunn et al., 2025; Arabi et al., 2025), style transfer (Hertz et al., 2024; Benaroche et al., 2024; Rout et al., 2025), and, most prominently, instruction-based editing of non-generated content, i.a., in the form of real images (Hertz et al., 2022; Kim et al., 2022; Su et al., 2022; Zhang et al., 2023; Parmar et al., 2023a; Dong et al., 2023; Wallace et al., 2023; Pan et al., 2023; Miyake et al., 2023; Mokady et al., 2023; Garibi et al., 2024; Brack et al., 2024; Huberman-Spiegelglas et al., 2024; Hong et al., 2024; Han et al., 2024; Tang et al., 2024; Wang et al., 2024a; Lin et al., 2024; Ju et al., 2024a; Samuel et al., 2025) or audio tracks (Liang et al., 2024; Jia et al., 2025; Zhang et al., 2024c; Manor & Michaeli, 2024; Niu et al., 2026; Lee et al., 2026; Yang et al., 2026). A large body of work in editing tasks improves inversion for text-to-image models, where classifier-free guidance further amplifies the inversion error. First family of methods focuses on optimizing the conditioning: null embeddings (Mokady et al., 2023) or prompt embeddings (Miyake et al., 2023; Dong et al., 2023; Han et al., 2024)). Other approaches leverage numerical improvements: fixed-point iteration (Pan et al., 2023; Samuel et al., 2025; Garibi et al., 2024), and higher-order solvers (Hong et al., 2024; Wang et al., 2024a). Alternative methods focus on bridging the statistical gap between inversions and Gaussian noise through gradient regularization (Parmar et al., 2023a), modifying the noise schedule for stabilization (Lin et al., 2024), or coupling inversion with forward diffusion (Staniszewski et al., 2026). The other group stores intermediate inversion variables to reduce either reconstruction error in stochastic DMs (Huberman-Spiegelglas et al., 2024; Brack et al., 2024; Manor & Michaeli, 2024) or the difference between inversion and reconstruction trajectories (Ju et al., 2024b). These methods either accept the local approximation (eq. (3)) or pay for exactness by incurring additional model evaluations and storing inversion states. In contrast, we tackle the source of the error directly by fine-tuning the adapter on the frozen model's own trajectories in a parameter-efficient manner, incurring no additional inference cost and leaving the generation process unchanged.

For such fine-tuning, we use Low-Rank Adaptation (LoRA, Hu et al. (2022)), altering a pretrained diffusion model via a low-rank update. Closely to our approach, recent works (Wang, 2026; Zhang et al., 2026; Dai et al., 2026) train a network to improve inversion accuracy and speed. Wang (2026) trains a separate inversion network with cycle-consistency and adversarial losses to map the clean latent of a few-step generator to its noise. Zhang et al. (2026) train a dedicated inversion solver on real images, supervised by the frozen model's prediction at the noisier latent produced by the solver's own step. Dai et al. (2026) fine-tune the generator itself with LoRA to predict average velocities in both the generation and inversion directions, tied together by a bidirectional consistency loss. These methods either evaluate an additional network during inversion or replace the original sampler with a distilled few-step model. In contrast, we keep the standard multi-step DDIM Inversion and only correct the noise prediction at each step (eq. (3)) with an adapter trained on the model's own sampling trajectories.

3 METHOD

In this work, we propose Moonwalker to reduce the approximation error (eq. (3)) of the DDIM Inversion algorithm by replacing, during inversion, the frozen noise predictor ϵ_θ with its LoRA-adapted counterpart $\epsilon_{\theta,\phi}$. Following, we describe how we train our adapter to map an available cleaner latent to the noise prediction required to recover its noisier trajectory state.

3.1 TRAJECTORY-SUPERVISED LORA INVERSION

First, we construct the training dataset from deterministically sampled trajectories $x_T \to x_{T-1} \to \cdots \to x_0$ generated by the frozen diffusion model. According to eq. (1), at each intermediate trajectory step ($x_t \to x_{t-1}$), the model performs noise prediction that is inaccessible during inversion. Thus, we collect it as a teacher signal for fine-tuning and denote it by $\hat{\epsilon}_t := \epsilon_\theta(x_t, t, c)$.

Our adapter learns to alleviate inversion error from eq. (3), i.e., given the less-noisy sample x_{t-1} at diffusion step t, it is trained to predict the teacher's signal $\hat{\epsilon}_t$ produced for the noisier state:

$$\mathcal{L}_{\text{inv}} = \mathbb{E}_{x_t, x_{t-1}, t, c} \left\| \epsilon_{\theta,\phi}(x_{t-1}, t, c) - \epsilon_\theta(x_t, t, c) \right\|_2^2 . \tag{4}$$

We then use such an adapter during inversion, substituting $\epsilon_\theta(x_t, t, c)$ with the prediction from the adapted model $\epsilon_{\theta,\phi}(x_{t-1}, t, c)$, effectively removing the circular dependency from eq. (2):

$$x_t = \sqrt{\alpha_t}\, x_{t-1} + \left(\sqrt{1-\bar{\alpha}_t} - \sqrt{\alpha_t - \bar{\alpha}_t}\right) \cdot \epsilon_{\theta,\phi}(x_{t-1}, t, c). \tag{5}$$

3.2 CYCLE-CONSISTENCY REGULARIZATION

The inversion-correction objective from eq. (4) directly constrains the adapted inversion noise prediction. However, to evaluate its downstream effect, we additionally impose a cycle-consistency objective that requires the inverted state to denoise back to the original (cleaner) one.

Let $I_{t,\theta,\phi}(\cdot; c)$ denote our LoRA-enabled inverse update (eq. (5)) from x_{t-1} to x_t, and let $G_{t,\theta}(\cdot; c)$ denote the frozen diffusion model denoising update (eq. (1)) over the same interval. Starting from x_{t-1}, we calculate $\tilde{x}_t = I_{t,\theta,\phi}(x_{t-1}; c)$, disable the adapter, and produce estimation $\tilde{x}_{t-1}$ of starting point with frozen denoising model as $\tilde{x}_{t-1} = G_{t,\theta}(\tilde{x}_t; c)$. Formally, this objective is defined as:

$$\mathcal{L}_{\text{cycle}} = \mathbb{E}_{x_t, x_{t-1}, t, c} \left\| G_{t,\theta}(I_{t,\theta,\phi}(x_{t-1}; c); c) - x_{t-1} \right\|_2^2 . \tag{6}$$

We define our final, cycle-regularized loss as $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{inv}} + \lambda_{\text{cycle}} \mathcal{L}_{\text{cycle}}$, where $\lambda_{\text{cycle}} \geq 0$ controls the regularization. While θ weights remain frozen during training, we retain gradients with respect to the input $\hat{x}_t$, letting the cycle error propagate through the denoiser to the LoRA parameters ϕ.

3.3 INTEGRATION WITH EDITION METHODS

At inference time, the adapter is enabled only during the inversion process and remains disabled during reconstruction or editing. Thus, Moonwalker can be easily combined with state-of-the-art editors by merging our trained adapters into the diffusion model during inversion. In our experiments, we test it by combining our inversion with Prompt-to-Prompt (P2P)(Hertz et al. (2023), Plug-and-Play (PnP)(Tumanyan et al. (2023)), MasaCtrl(Cao et al. (2023a)), and Pix2Pix-Zero(Parmar et al. (2023b)). Additionally, we show that we can improve existing inversion-specific techniques by combining Moonwalker with Direct Inversion (Ju et al., 2024b) (Moonwalker+DirectInv).

3.4 INVERSION WITH CLASSIFIER-FREE GUIDANCE

In image editing, to apply manipulations that closely align with instructions, a common approach is to use Classifier-Free Guidance (CFG; Ho & Salimans (2021)). CFG, at inference time, combines prompt-conditioned (c) and unconditional (or null-prompt, $\varnothing$) DM outputs as $\epsilon_\theta^w(x, t, c) = (1-w) \cdot \epsilon_\theta(x, t, \varnothing) + w \cdot \epsilon_\theta(x, t, c)$, guiding generation process to follow prompts with steerable strength w.

Here, we propose employing our trajectory-supervised fine-tuning to correct inversion errors in DM predictions operating within CFG. To this end, we collect teacher signal from guided denoising trajectories, caching conditional $\hat{\epsilon}_t^c = \epsilon_\theta(x_t, t, c)$ and unconditional $\hat{\epsilon}_t^\varnothing = \epsilon_\theta(x_t, t, \varnothing)$ predictions, and computing their composition as $\hat{\epsilon}_t^w = (1-w) \cdot \hat{\epsilon}_t^\varnothing + w \cdot \hat{\epsilon}_t^c$. During fine-tuning, the frozen model with trained adapters receives a less-noisy state x_{t-1}, and performs conditional $\epsilon_t^c = \epsilon_{\theta,\phi_c}(x_{t-1}, t, c)$ and unconditional $\epsilon_t^\varnothing = \epsilon_{\theta,\phi_\varnothing}(x_{t-1}, t, \varnothing)$ predictions. We consider two configurations of how adapted weights are distributed for both predictions: *shared* – when a single adapter is used for both ($\phi_\varnothing = \phi_c$), and *paired*, with separate trainable weights ($\phi_\varnothing \neq \phi_c$).

Finally, we propose two objective alternatives: reconstructing the teacher's CFG-combined signal ($\mathcal{L}_{\text{cfg}}$) or estimating both teacher predictions separately ($\mathcal{L}_{\text{branch}}$):

$$\mathcal{L}_{\text{cfg}} = \mathbb{E}_{x_t, x_{t-1}, t, c} \left\| (1-w) \cdot \epsilon_t^{\varnothing} + w \cdot \epsilon_t^{c} - \hat{\epsilon}_t^{w} \right\|_2^2, \tag{7}$$

$$\mathcal{L}_{\text{branch}} = \mathbb{E}_{x_t, x_{t-1}, t, c} \left(\left\| \epsilon_{\theta, \phi_{\varnothing}}(x_{t-1}, t, \varnothing) - \hat{\epsilon}_t^{\varnothing} \right\|_2^2 + \left\| \epsilon_{\theta, \phi_c}(x_{t-1}, t, c) - \hat{\epsilon}_t^{c} \right\|_2^2 \right). \tag{8}$$

3.5 Extension to Flow Matching

Finally, we apply our fine-tuning method to flow matching models (FMs; Lipman et al., 2023; Liu et al., 2023b), for which inversion is also commonly used in image editing (Rout et al., 2025; Avrahami et al., 2025). In FMs, the frozen network v_θ parametrizes the velocity field of the ordinary differential equation (ODE) defined over the time interval $t \in [0, 1]$ as $\mathrm{d}x_t = v_\theta(x_t, t, c)\,\mathrm{d}t$. To construct trajectory dataset, sampling starts from initial condition $x_1 \sim \mathcal{N}(0, \mathcal{I})$ and integrates the ODE over time (from $t = 1$ to $t = 0$) with the Euler method:

$$x_{t-\Delta t} = x_t - \Delta t \cdot v_\theta(x_t, t, c), \tag{9}$$

where $\Delta t > 0$ denotes step size at time t set by noise schedule. Each transition in eq. (9) produces a teacher velocity $\hat{v}_t = v_\theta(x_t, t)$ that we cache for training.

During inversion, the velocity cannot be evaluated at the unknown noisier state x_t, so the standard ODE Inversion step, as used in Avrahami et al. (2025), predicts at the cleaner, available state instead: $v_\theta(x_t, t) \approx v_\theta(x_{t-1}, t)$. To reduce this local approximation error, as in section 3.1 for DMs, we fine-tune the LoRA-adapted predictor $v_{\theta,\phi}$ so that its prediction matches teacher's output:

$$\mathcal{L}_{\text{vel}} = \mathbb{E}_{x_{t-\Delta t}, t, c} \left\| v_{\theta,\phi}(x_{t-\Delta t}, t, c) - \hat{v}_t \right\|_2^2. \tag{10}$$

After fine-tuning, the adapted predictor constructs the inversion updates as

$$x_t = x_{t-\Delta t} + \Delta t \cdot v_{\theta,\phi}(x_{t-\Delta t}, t, c). \tag{11}$$

4 Experiments

4.1 Experimental setup

We evaluate Moonwalker across noise recovery and downstream tasks, including image reconstruction, text-guided image editing, style transfer, and, outside the visual domain, music editing. Throughout these experiments, the adapter is enabled only during inversion; reconstruction and editing use the corresponding frozen base model and unchanged downstream method. The image adapters are trained on deterministic 50-step trajectories synthesized by the frozen backbones from ReCap-COCO (Li et al., 2025) prompts, while the teacher signal for audio models is collected via 200-step (AudioLDM2, Liu et al., 2023a) or 100-step (Stable Audio Open, Evans et al., 2024) sampling from the MusicCaps (Agostinelli et al., 2023) prompts. We train separate adapters for U-Net-based image generators – Stable Diffusion 1.5 (Rombach et al., 2022)) and SDXL (Podell et al., 2024) – and flow-based FLUX-1 Labs (2024), evaluating Lora ranks $r \in \{8, 16, 32\}$. The following subsections describe the task-specific protocols, while Appendix A provides complete details.

4.2 Inversion quality

To measure the inversion quality, we use a held-out set of 1,522 ReCap-COCO prompts. For each prompt, we generate a reference trajectory and final image using a deterministic DDIM sampler and classifier-free guidance (CFG) scale 1. Because the initial noise is known, we can evaluate both the recovery of that noise and the reconstruction of the generated reference image. We compare standard DDIM inversion, Moonwalker, and the forward-diffusion correction of Staniszewski et al. (2026) under the same prompts, seeds, scheduler, and conditioning. For the latter, Forward 2% and Forward 4% replace the first inversion steps with closed-form forward diffusion. Reconstruction always uses the frozen base model without the adapter.

Table 1: Comparison of DDIM inversion, the forward-diffusion correction (Staniszewski et al., 2026), and Moonwalker. Correlation and KL divergence assess the normality of the recovered noise, while MAE, LPIPS, PSNR, and SSIM assess reconstruction. Bold denotes the best result.

Method	Corr. ↓	KL Div. $\times 10^2$ ↓	MAE ↓	LPIPS ↓	PSNR ↑	SSIM ↑
DDIM Inv.	0.17	0.71	0.03	0.10	28.22	0.81
Forward 2%	0.15	0.54	0.03	0.10	27.69	0.79
Forward 4%	0.11	0.24	0.04	0.14	25.84	0.70
Moonwalker $r = 8$	0.10	0.08	0.02	0.06	30.06	0.87
Moonwalker $r = 16$	0.10	0.07	0.02	0.06	30.22	0.87
Moonwalker $r = 32$	**0.10**	**0.05**	**0.02**	**0.06**	**30.40**	**0.87**

Table 1 evaluates both the initial noise recovery and reconstruction of the generated reference image. Correlation is the mean top-k Pearson correlation within 8×8 latent patches, and KL is the Gaussian divergence between the recovered and original noise distributions. Moonwalker improves both sides of the normality–fidelity trade-off: increasing rank monotonically lowers correlation and KL while also lowering MAE and LPIPS and increasing PSNR and SSIM. In particular, $r = 32$ improves PSNR from the DDIM regime to 30.40 and reduces KL by more than an order of magnitude.

The forward-diffusion correction reduces spatial correlation by injecting independent noise, but stronger replacement also sacrifices reconstruction fidelity. In contrast, Moonwalker removes structured residuals while preserving information needed by the unchanged reverse process. Appendix C reports rank, active-step, regional, text-editing, and CFG ablations. Qualitative reconstructions illustrating these improvements are shown in Appendix Figures 5

4.3 PIEBENCH EDITING BENCHMARK

For the downstream evaluation, we compare standard DDIM, Direct Inversion, Moonwalker, and Moonwalker+DirectInv on all 700 PIE-Bench (Ju et al., 2024a) examples using Stable Diffusion 1.5 and four editors: Prompt-to-Prompt (P2P, Hertz et al., 2023), Plug-and-Play (PnP, Tumanyan et al., 2023), MasaCtrl (Cao et al., 2023a)), and Pix2Pix-Zero (Parmar et al., 2023b). Structure measures global consistency; PSNR, LPIPS, MSE, and SSIM measure preservation in the unedited region; and the two CLIP scores measure target-prompt alignment to the fully edited image and its edited region. Complete inference settings and configuration-selection details are provided in App. A.

Table 2 compares the reproduced controls with selected configurations from each Moonwalker training variants. For P2P, cycle-consistent Moonwalker+DirectInv substantially improves every preservation metric over DirectInv, while standalone Moonwalker gives the strongest CLIP scores. PnP and MasaCtrl exhibit a trade-off: DirectInv retains the strongest preservation, whereas Moonwalker-based configurations improve text alignment. Pix2Pix-Zero shows the same split. Final-tail $r = 8$ gives the highest CLIP target score, while cycle-consistent $r = 32$ gives the highest CLIP edit score.

Moonwalker configurations obtain lower preservation scores than DirectInv in Table 2. However, these metrics are computed on the final edited images, so they reflect both reconstruction quality and changes introduced by the editor, even within regions intended to remain unchanged. App. B.1 evaluates reconstruction separately by omitting the target edit. In this setting (Table 5), Moonwalker+DirectInv substantially improves P2P reconstruction, gives similar results to DirectInv for PnP and Pix2Pix-Zero, and causes only a small degradation for MasaCtrl. Therefore, lower preservation scores after editing do not necessarily imply worse reconstruction from the inverted latent.

4.4 STYLE TRANSFER

We evaluate Moonwalker on StyleBench (Gao et al., 2024), which comprises 490 style-reference images spanning 73 styles and 20 official content prompts, using two existing style-transfer pipelines: StyleAligned (Hertz et al., 2024) and Inversion-InstantStyle (Benaroche et al., 2024). Each experiment compares DDIM with two rank-16 adapters used only during inversion. *Moonwalker-ReCap* and *Moonwalker-BLIP* follow the same 50-step synthetic-trajectory training protocol at CFG 1, using ReCap-COCO captions and StyleBench-derived BLIP-2 captions (Li et al., 2023), respectively.

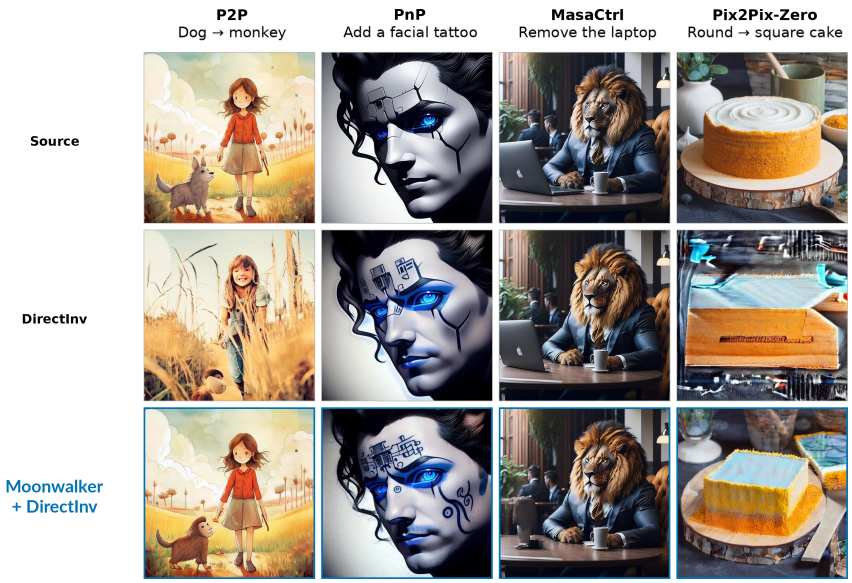


Figure 2: Qualitative PIE-Bench comparison of SD1.5 DirectInv and Moonwalker+DirectInv. Columns show the source, DirectInv, and Moonwalker+DirectInv; rows show P2P (dog to monkey), PnP (facial tattoo), MasaCtrl (remove laptop), and Pix2Pix-Zero (round to square cake).

Table 2: PIE-Bench editing with SD1.5. PSNR, LPIPS, MSE, and SSIM measure preservation inside the unedited-region mask of the edited output. Bold denotes the best result within each editor.

Editor	Inversion	Training	Structure ($\times 10^3$) ↓	PSNR ↑	LPIPS ($\times 10^3$) ↓	MSE ($\times 10^4$) ↓	SSIM ($\times 10^2$) ↑	CLIP target ↑	CLIP edit ↑
P2P	DDIM	–	70.50	17.84	210.15	220.28	71.56	25.19	22.39
P2P	DirectInv	–	31.35	22.57	125.51	91.27	78.54	25.07	22.22
P2P	Moonwalker	uniform $r = 8$	70.45	17.81	211.10	221.32	71.37	**25.38**	**22.57**
P2P	Moonwalker+DirectInv	uniform $r = 8$	14.12	26.02	73.25	44.40	83.50	25.33	22.31
P2P	Moonwalker+DirectInv	final-tail $r = 8$	14.27	25.95	72.48	44.31	83.55	25.35	22.37
P2P	Moonwalker+DirectInv	cycle $r = 16, \lambda = 0.5$	**14.00**	**26.05**	**71.76**	**43.57**	**83.60**	25.28	22.31
PnP	DDIM	–	27.32	22.31	112.76	82.98	79.58	25.59	22.57
PnP	DirectInv	–	**23.33**	**22.46**	**105.51**	**79.97**	**80.22**	25.63	22.67
PnP	Moonwalker	uniform $r = 32$	27.08	22.33	113.47	81.70	79.54	25.61	22.81
PnP	Moonwalker+DirectInv	uniform $r = 8$	24.27	22.37	110.17	81.82	79.90	**25.83**	22.85
PnP	Moonwalker+DirectInv	final-tail $r = 32$	24.43	22.36	110.82	81.83	79.88	25.82	**22.89**
PnP	Moonwalker+DirectInv	cycle $r = 8, \lambda = 1$	24.19	22.38	110.03	81.64	79.92	25.82	**22.89**
MasaCtrl	DDIM	–	27.68	22.32	105.51	84.69	80.41	24.16	21.20
MasaCtrl	DirectInv	–	**23.57**	**22.82**	**86.83**	**79.19**	**82.02**	24.51	21.47
MasaCtrl	Moonwalker	uniform $r = 16$	26.49	22.29	105.30	87.57	80.37	24.22	21.18
MasaCtrl	Moonwalker+DirectInv	uniform $r = 8$	28.58	21.90	99.01	99.13	80.49	24.61	**21.50**
MasaCtrl	Moonwalker+DirectInv	final-tail $r = 8$	28.38	21.93	98.72	98.66	80.54	**24.64**	21.46
MasaCtrl	Moonwalker+DirectInv	cycle $r = 16, \lambda = 4$	28.51	21.90	98.89	98.91	80.50	24.62	**21.50**
Pix2Pix-Zero	DDIM	–	64.97	20.31	177.80	150.71	74.91	22.97	20.69
Pix2Pix-Zero	DirectInv	–	**49.99**	**21.47**	**142.46**	**133.49**	**77.49**	23.53	21.09
Pix2Pix-Zero	Moonwalker	uniform $r = 8$	51.59	21.33	145.87	140.59	76.98	23.46	20.99
Pix2Pix-Zero	Moonwalker+DirectInv	uniform $r = 16$	51.63	21.25	147.07	137.65	76.94	23.39	20.95
Pix2Pix-Zero	Moonwalker+DirectInv	final-tail $r = 8$	51.67	21.33	144.87	135.70	77.15	**23.56**	21.01
Pix2Pix-Zero	Moonwalker+DirectInv	cycle $r = 32, \lambda = 1$	53.15	21.24	146.02	138.96	76.95	23.45	**21.11**

Dataset sizes and splits are detailed in Appendix A. Generation always uses the frozen base model. CLIP-T and ImageReward measure target-prompt alignment, CLIP-I and DINOv2 measure reference similarity, and CSD measures style similarity. Within each setting, all methods share references, inversion prompts, target prompts, and seeds. Both pipelines are evaluated on StyleBench with BLIP-2 inversion captions, 50 inversion and generation steps, and inversion/generation CFG $1/8$. Inversion-InstantStyle is evaluated at $s \in \{0, 1\}$.

StyleAligned replays the inverted trajectory and shares attention features with the generations. Table 3 shows small increases in CLIP-I and CSD for both adapters over DDIM. Moonwalker-BLIP offers the highest DINOv2. CLIP-T is nearly unchanged, and ImageReward decreases. The adapters, therefore, shift the output toward reference- and style-similarity without improving all metrics.

In Inversion-InstantStyle, the inverted endpoint initializes InstantStyle (Wang et al., 2024b), and s controls how much noise is mixed into it. With BLIP-2 inversion captions (table 4), DDIM gives the highest CLIP-T and ImageReward at $s = 0$, while Moonwalker-BLIP gives the highest CSD, CLIP-I, and DINOv2. At $s = 1$, Moonwalker-ReCap gives the highest CLIP-T, CSD, and ImageReward, whereas Moonwalker-BLIP retains the highest CLIP-I and DINOv2. No method is consistently best across target, reference, and style metrics. We present a qualitative comparison in fig. 6

Table 3: StyleAligned on StyleBench. Entries are reference-weighted means, where higher is better. Adapters have rank 16. Bold denotes the best mean in each column.

Inversion	CLIP-T ↑	CLIP-I ↑	DINOv2 ↑	CSD ↑	ImageReward ↑
DDIM	19.698	69.704	32.950	0.5501	**-1.218**
Moonwalker-ReCap	**19.702**	70.018	32.871	0.5556	-1.243
Moonwalker-BLIP	19.684	**70.057**	**32.981**	**0.5597**	-1.244

Table 4: Inversion-InstantStyle on StyleBench with BLIP-2 inversion captions. Both adapters have rank 16. Higher is better; bold denotes the best mean at each noise level.

s	Inversion	CLIP-T ↑	CLIP-I ↑	DINOv2 ↑	CSD ↑	ImageR. ↑
0	DDIM	**18.234**	74.719	46.267	0.6436	**-1.518**
	Moonwalker-ReCap	18.173	75.105	46.236	0.6486	-1.539
	Moonwalker-BLIP	17.960	**75.226**	**46.931**	**0.6498**	-1.552
1	DDIM	20.307	71.151	36.884	0.6105	-0.922
	Moonwalker-ReCap	**20.370**	71.152	36.714	**0.6135**	**-0.908**
	Moonwalker-BLIP	20.178	**71.369**	**37.216**	0.6122	-0.942

4.5 Music Editing

Finally, we test whether Moonwalker transfers beyond images to music editing. We use two diffusion models: AudioLDM2 (Liu et al., 2023a), a U-Net sampled with DDIM, and Stable Audio Open (Evans et al., 2024), a diffusion transformer that we sample with a deterministic first-order ODE solver. For each model, we construct the trajectory dataset via 200-step (AudioLDM2) or 100-step (Stable Audio Open) sampling with the frozen model using MusicCaps (Agostinelli et al., 2023) captions. We adapt models' attention layers with rank $r = 8$.

Following Manor & Michaeli (2024), we edit 34 real MedleyDB (Bittner et al., 2014) tracks with the source and target captions of MedleyMDPrompts (Manor & Michaeli, 2024), yielding 696 edits, and compare with DDPM (Huberman-Spiegelglas et al., 2024; Manor & Michaeli, 2024) and DDIM Inversions, and SDEdit (Meng et al., 2022). We measure how well the edited audio preserves the input with LPAPS (Iashin & Rahtu, 2021), and how it follows the target caption with CLAP-T (Wu et al., 2022) and MuQ-T (Zhu et al., 2025). All evaluated methods trade preservation for adherence via the inversion depth t' (i.e., the fraction of the sampling trajectory that is inverted) and the guidance scale w during generation. We therefore sweep both parameters ($t' \in \{100\%, 75\%, 50\%, 33\%, 25\%\}$, $w \in \{3.5, 7, 10.5, 14\}$) and compare the resulting Pareto fronts at a matched function evaluations budget (NFEs): 400 for AudioLDM2 and 100 for Stable Audio Open.

We present the results in Figure 3. When editing music with Stable Audio Open, Moonwalker improves upon DDIM Inversion in preservation, lowering LPAPS in all 20 swept settings by 0.04 on average, while changing CLAP-T and MuQ-T by less than 0.01. Furthermore, Moonwalker outperforms SDEdit wherever their ranges overlap and performs on par with DDPM Inversion, as neither method consistently preserves the source better at matched adherence. However, Moonwalker reaches stronger maximum edits (CLAP-T 0.35 vs. 0.34 and MuQ-T 0.31 vs. 0.27). On AudioLDM2, Pareto fronts of the DDIM Inversion and Moonwalker coincide within the standard with the DDPM Inversion over most of the range.

5 Additional experiments

5.1 Inversion with CFG

We compare two guidance regimes: equal CFG during inversion and generation, and lower CFG inversion with higher CFG generation. The matched evaluation covers $w \in \{2, 3, 4, 5, 7.5, 10\}$, while the unequal setup covers generation $w \in \{5, 7.5, 10\}$ with inversion CFG fixed at 2.

Table 14 evaluates DDIM and four mixed-CFG-trained Moonwalker configurations at matching inversion and generation guidance: $w_{\text{inv}} = w_{\text{gen}} \in \{2, 3, 4, 5, 7.5, 10\}$. Each guidance setting contains 166–167 examples. Reconstruction compares the source image with its reconstruction, whereas CLIP evaluates the output of the synthetic prompt-shift editing probe.

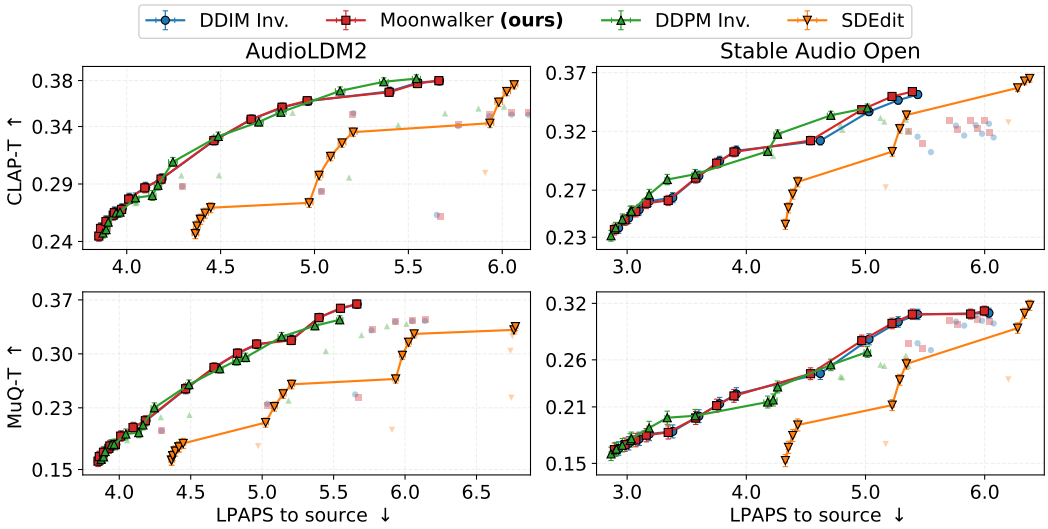


Figure 3: **Comparison of inversion-based music editing methods.** We report audio preservation vs. edit prompt alignment Pareto fronts. On AudioLDM2, the DDIM Inversion, DDPM Inversion, and Moonwalker fronts coincide, whereas on Stable Audio, our adapter shifts the DDIM Inversion front toward better preservation at matched adherence.

Table 16 compares DDIM using matched inversion and generation CFG with DDIM using inversion CFG 2. Source generation, reconstruction, and editing retain CFG 5, 7.5, or 10, with 50 steps in each direction. The two inversion settings share the same source images, prompts, and seeds: 167 examples at CFG 5 and 166 at each other's setting. Editing prompts are shifted by 1.

5.2 Flow Matching

We evaluate FLUX.1-dev with the base Euler inversion, Moonwalker, and latent nudging introduced by Avrahami et al. (2025), which multiplies the clean input latent before inversion. All methods use 50 inversion steps and 50 steps for each generation pass. Reconstruction uses the source prompt and edits the target prompt.

Table 17 evaluates reconstruction and a controlled editing probe on 1,000 source images at 1024×1024, generated beforehand by FLUX.1-dev from the prompts from Recap-COCO. They are synthetic FLUX outputs rather than images taken from the COCO dataset. Guidance $g = 1$ and $g = 3.5$ use separately trained adapters and separately generated source trajectories. Within each block, all methods share source latents, prompts, and seeds. The synthetic editing probe shifts target prompts by one position.

We additionally evaluate all 700 PIE-Bench examples at 512×512. Embedded guidance is $g = 1$ for inversion, reconstruction, and editing. We score the source-prompt reconstruction over the entire image, background preservation on the edited output outside the mask, and target-prompt CLIP agreement on the whole edited image and inside the mask.

6 Conclusion

We present Moonwalker, a trajectory-supervised LoRA adapter designed to correct the local approximation errors inherent in standard DDIM inversion. By learning to perfectly retrace sampling trajectories from a frozen model, Moonwalker recovers mathematically clean noise that rigorously matches the theoretical Gaussian prior while preserving high reconstruction fidelity. This parameter-efficient approach requires zero additional inference overhead, leaves the pretrained generative mapping unchanged, and extends seamlessly to classifier-free guidance, flow-matching, and audio diffusion models. Ultimately, by successfully transforming deterministic inversion error into a learnable component, Moonwalker unlocks enhanced stability and performance gains across diverse downstream editing pipelines.

AI USE STATEMENT

In this work, we used generative AI to implement author-designed methods, including code for dataset preparation, experiments, and evaluation; provide feedback on the research methodology and experiments; and support qualitative analysis and the interpretation of results. We did not use generative AI to originate the research question or core idea, develop the conceptual or mathematical formulation, or propose or refine research hypotheses. Formal proofs and translation were not applicable to this work. Additionally, we used generative AI tools to create and edit software and visualizations, prepare research artifacts, improve the clarity and readability of the manuscript, refine its structure and title, and assist with the presentation of results. We reviewed all AI-assisted work: we manually assessed methodological feedback and interpretations, inspected and tested code changes, checked reported values against the corresponding experimental artifacts, and reviewed and edited AI-assisted text and visualizations. We made all final methodological and interpretive decisions and take responsibility for the final content of this work, including text, claims, code, artifacts, and visualizations produced with the aid of generative AI.

ETHICS STATEMENT

This work does not involve a human-subject study or the collection of new personal or sensitive data. It uses existing research datasets and pretrained generative models. Image and audio editing can support legitimate creative applications, but it can also be misused to create deceptive media, imitate copyrighted material or artistic styles, or alter content without the consent of affected parties. Moonwalker does not provide a content-safety mechanism, and its outputs inherit biases and failure modes from the underlying models and datasets. Users of the method should respect dataset and model licenses, intellectual-property and privacy rights, and applicable consent and provenance requirements.

REPRODUCIBILITY STATEMENT

Sections 3.1–3.5 specify the training objectives and inference procedures. Appendix A documents the core image-model training and inference protocol; the experiment-specific sections and remaining appendices provide data-split and metric definitions, additional settings, and ablation results. The accompanying code release provides data-generation, training, evaluation, and analysis implementations, together with experiment configuration files, dependency specifications, and manifests recording dataset and split definitions, prompt identifiers, seeds, sampling schedules, checkpoint-selection rules, and metric aggregation procedures. Inputs are matched across methods for comparisons reported as paired. Large generated trajectories and benchmark outputs are not bundled with the paper. The release instead provides scripts and manifests for rerunning the corresponding training and evaluation pipelines; the cited datasets, pretrained model weights, and third-party editing implementations must be obtained separately under their original terms.

REFERENCES

Andrea Agostinelli, Timo I. Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, Matt Sharifi, Neil Zeghidour, and Christian Frank. Musiclm: Generating music from text. *arXiv preprint arXiv: 2301.11325*, 2023.

Kasra Arabi, Benjamin Feuer, R Teal Witter, Chinmay Hegde, and Niv Cohen. Hidden in the noise: Two-stage robust watermarking for images. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=ll2nz6qwRG.

Omri Avrahami, Or Patashnik, Ohad Fried, Egor Nemchinov, Kfir Aberman, Dani Lischinski, and Daniel Cohen-Or. Stable flow: Vital layers for training-free image editing. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 7877–7888, 2025.

Eyal Benaroche, Clément Chadebec, Onur Tasar, and Benjamin Aubin. Style-Rank: Benchmarking stylization for diffusion models. Project website and open-source software, 2024. URL https://gojasper.github.io/style-rank-project/. Code: https://github.com/gojasper/style-rank.

Rachel M. Bittner, Justin Salamon, Mike Tierney, Matthias Mauch, Chris Cannam, and Juan Pablo Bello. MedleyDB: A multitrack dataset for annotation-intensive MIR research. In *International Society for Music Information Retrieval Conference*, 2014.

Erik Bodin, Alexandru I Stere, Dragos D Margineantu, Carl Henrik Ek, and Henry Moss. Linear combinations of latents in generative models: subspaces and beyond. In *The Thirteenth International Conference on Learning Representations*, 2025.

Manuel Brack, Felix Friedrich, Katharina Kornmeier, Linoy Tsaban, Patrick Schramowski, Kristian Kersting, and Apolinário Passos. Ledits++: Limitless image editing using text-to-image models. *CVPR*, 2024.

Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22560–22570, October 2023a.

Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 22560–22570, 2023b.

Hai Ci, Pei Yang, Yiren Song, and Mike Zheng Shou. Ringid: Rethinking tree-ring watermarking for enhanced multi-key identification. In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XXVIII*, volume 15086 of *Lecture Notes in Computer Science*, pp. 338–354. Springer, 2024. doi: 10.1007/978-3-031-73390-1_20. URL https://doi.org/10.1007/978-3-031-73390-1_20.

Yasong Dai, Zeeshan Hayder, David Ahmedt-Aristizabal, and Hongdong Li. Bifm: Bidirectional flow matching for few-step image editing and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23325–23334, June 2026.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. *Advances in Neural Information Processing Systems*, 34, 2021.

Wenkai Dong, Song Xue, Xiaoyue Duan, and Shumin Han. Prompt tuning inversion for text-driven image editing using diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7430–7440, 2023.

Zach Evans, Julian D. Parker, CJ Carr, Zack Zukowski, Josiah Taylor, and Jordi Pons. Stable audio open. *arXiv preprint arXiv:2407.14358*, 2024.

Junyao Gao, Yanchen Liu, Yanan Sun, Yinhao Tang, Yanhong Zeng, Kai Chen, and Cairong Zhao. StyleShot: A snapshot on any style. *arXiv preprint arXiv:2407.01414*, 2024. URL https://arxiv.org/abs/2407.01414.

Daniel Garibi, Or Patashnik, Andrey Voynov, Hadar Averbuch-Elor, and Daniel Cohen-Or. Renoise: Real image inversion through iterative noising. In *European Conference on Computer Vision*, pp. 395–413. Springer, 2024.

Sam Gunn, Xuandong Zhao, and Dawn Song. An undetectable watermark for generative image models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=jlhBFm7T2J.

Ligong Han, Song Wen, Qi Chen, Zhixing Zhang, Kunpeng Song, Mengwei Ren, Ruijiang Gao, Anastasis Stathopoulos, Xiaoxiao He, Yuxiao Chen, et al. Proxedit: Improving tuning-free real image editing with proximal guidance. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 4291–4301, 2024.

Amir Hertz, Ron Mokady, J. Tenenbaum, Kfir Aberman, Y. Pritch, and D. Cohen-Or. Prompt-to-prompt image editing with cross attention control. *International Conference on Learning Representations*, 2022. doi: 10.48550/arXiv.2208.01626.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-Prompt Image Editing with Cross-Attention Control. In *International Conference on Learning Representations*, 2023. URL https://mlanthology.org/iclr/2023/hertz2023iclr-prompttoprompt/.

Amir Hertz, Andrey Voynov, Shlomi Fruchter, and Daniel Cohen-Or. Style aligned image generation via shared attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4775–4785, 2024.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. URL https://openreview.net/forum?id=qw8AKxfYbI.

Seongmin Hong, Kyeonghyun Lee, Suh Yoon Jeon, Hyewon Bae, and Se Young Chun. On exact inversion of dpm-solvers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7069–7078, 2024.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Inbar Huberman-Spiegelglas, Vladimir Kulikov, and Tomer Michaeli. An edit friendly ddpm noise space: Inversion and manipulations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12469–12478, 2024.

Vladimir E. Iashin and Esa Rahtu. Taming visually guided sound generation. *British Machine Vision Conference*, 2021. doi: 10.5244/c.35.336.

Yuhang Jia, Yang Chen, Jinghua Zhao, Shiwan Zhao, Wenjia Zeng, Yong Chen, and Yong Qin. Audioeditor: A training-free diffusion-based audio editing framework. *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2025. doi: 10.1109/ICASSP49660.2025.10887821.

Guanlong Jiao, Biqing Huang, Kuan-Chieh Jackson Wang, and Renjie Liao. Uniedit-flow: Unleashing inversion and editing in the era of flow models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=ArU2CeB7Tm.

Xuan Ju, Ailing Zeng, Yuxuan Bian, Shaoteng Liu, and Qiang Xu. PnP Inversion: Boosting Diffusion-Based Editing with 3 Lines of Code. In *International Conference on Learning Representations*, 2024a. URL https://mlanthology.org/iclr/2024/ju2024iclr-pnp/.

Xuan Ju, Ailing Zeng, Yuxuan Bian, Shaoteng Liu, and Qiang Xu. Pnp inversion: Boosting diffusion-based editing with 3 lines of code. In *The Twelfth International Conference on Learning Representations*, 2024b. URL https://openreview.net/forum?id=FoMZ4ljhVw.

Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2426–2435, June 2022.

Black Forest Labs. Flux.1, 2024. URL https://blackforestlabs.ai/announcing-black-forest-labs/.

Ching Ho Lee, Javier Nistal, Stefan Lattner, Marco Pasini, and George Fazekas. Diffusion timbre transfer via mutual information guided inpainting. *arXiv preprint arXiv:2601.01294*, 2026.

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. URL https://arxiv.org/abs/2301.12597.

Xianhang Li, Haoqin Tu, Mude Hui, Zeyu Wang, Bingchen Zhao, Junfei Xiao, Sucheng Ren, Jieru Mei, Qing Liu, Huangjie Zheng, Yuyin Zhou, and Cihang Xie. What if we recaption billions of web images with LLaMA-3? In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 35957–35976. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/li25ch.html.

Jinhua Liang, Yi Yuan, Dongya Jia, Xiaobin Zhuang, Zhengxi Liu, Yuanzhe Chen, Zhuo Chen, Yuping Wang, and Yuxuan Wang. Audiomorphix: Training-free audio editing with diffusion probabilistic models, 2024. URL https://openreview.net/forum?id=a8dQutiF9E.

Haonan Lin, Mengmeng Wang, Jiahao Wang, Wenbin An, Yan Chen, Yong Liu, Feng Tian, Guang Dai, Jingdong Wang, and Qianying Wang. Schedule your edit: A simple yet effective diffusion noise schedule for image editing. *Neural Information Processing Systems*, 2024. doi: 10.48550/arXiv.2410.18756.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Haohe Liu, Yi Yuan, Xubo Liu, Xinhao Mei, Qiuqiang Kong, Qiao Tian, Yuping Wang, Wenwu Wang, Yuxuan Wang, and Mark D. Plumbley. AudioLDM 2: Learning holistic audio generation with self-supervised pretraining. *arXiv preprint arXiv:2308.05734*, 2023a.

Xingchao Liu, Chengyue Gong, and qiang liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations*, 2023b. URL https://openreview.net/forum?id=XVjTT1nw5z.

Hila Manor and Tomer Michaeli. Zero-shot unsupervised and text-based audio editing using DDPM inversion. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 34603–34629. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/manor24a.html.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=aBsCjcPu_tE.

Daiki Miyake, Akihiro Iohara, Yu Saito, and Toshiyuki Tanaka. Negative-prompt inversion: Fast image inversion for editing with text-guided diffusion models. *arXiv preprint arXiv: 2305.16807*, 2023.

Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6038–6047, June 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Mokady_NULL-Text_Inversion_for_Editing_Real_Images_Using_Guided_Diffusion_Models_CVPR_2023_paper.html.

Xinlei Niu, Kin Wai Cheuk, Jing Zhang, Naoki Murata, Chieh-Hsin Lai, Michele Mancusi, Woosung Choi, Giorgio Fabbro, Wei-Hsiang Liao, Charles Patrick Martin, and Yuki Mitsufuji. Steermusic: enhanced musical consistency for zero-shot text-guided and personalized music editing. In *Proceedings of the Fortieth AAAI Conference on Artificial Intelligence and Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence and Sixteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'26/IAAI'26/EAAI'26. AAAI Press, 2026. ISBN 978-1-57735-906-7. doi: 10.1609/aaai.v40i3.37181. URL https://doi.org/10.1609/aaai.v40i3.37181.

Zhihong Pan, Riccardo Gherardi, Xiufeng Xie, and Stephen Huang. Effective real image editing with accelerated iterative diffusion inversion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15912–15921, October 2023.

Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. In *ACM SIGGRAPH 2023 Conference Proceedings*, pp. 1–11, 2023a.

Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation, 2023b. URL https://arxiv.org/abs/2302.03027.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=di52zR8xgf.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, June 2022.

Litu Rout, Yujia Chen, Nataniel Ruiz, Constantine Caramanis, Sanjay Shakkottai, and Wen-Sheng Chu. Semantic image inversion and editing using rectified stochastic differential equations. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Hu0FSOSEyS.

Dvir Samuel, Rami Ben-Ari, Nir Darshan, Haggai Maron, and Gal Chechik. Norm-guided latent space exploration for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:57863–57875, 2023.

Dvir Samuel, Barak Meiri, Haggai Maron, Yoad Tewel, Nir Darshan, Shai Avidan, Gal Chechik, and Rami Ben-Ari. Lightning-fast image inversion and editing for text-to-image diffusion models. In *The Thirteenth International Conference on Learning Representations*, 2025.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pp. 2256–2265. PMLR, 2015.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=St1giarCHLP.

Łukasz Staniszewski, Łukasz Kuciński, and Kamil Deja. There and back again: On the relation between noise and image inversions in diffusion models. In *International Conference on Learning Representations*, volume 2026, pp. 20792–20838, 2026.

Xu Su, Jiaming Song, Chenlin Meng, and Stefano Ermon. Dual diffusion implicit bridges for image-to-image translation. *International Conference on Learning Representations*, 2022. doi: 10.48550/arXiv.2203.08382.

Chuanming Tang, Kai Wang, and Joost van de Weijer. IterInv: Iterative Inversion for Pixel-Level T2I Models . In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, Los Alamitos, CA, USA, 2024. IEEE Computer Society. doi: 10.1109/ICME57554.2024.10687547. URL https://doi.ieeecomputersociety.org/10.1109/ICME57554.2024.10687547.

Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1921–1930, June 2023.

Bram Wallace, Akash Gokul, and Nikhil Naik. Edict: Exact diffusion inversion via coupled transformations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22532–22541, 2023.

Fangyikang Wang, Hubery Yin, Yuejiang Dong, Huminhao Zhu, Chao Zhang, Han Zhao, Hui Qian, and Chen Li. Belm: Bidirectional explicit linear multi-step sampler for exact inversion in diffusion models. *Neural Information Processing Systems*, 2024a. doi: 10.48550/arXiv.2410.07273.

Guangzhi Wang. Flashin: Fast and accurate image inversion for real-time image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 30425–30434, 2026.

Haofan Wang, Matteo Spinelli, Qixun Wang, Xu Bai, Zekui Qin, and Anthony Chen. InstantStyle: Free lunch towards style-preserving in text-to-image generation. *arXiv preprint arXiv:2404.02733*, 2024b. URL https://arxiv.org/abs/2404.02733.

Yuxin Wen, John Kirchenbauer, Jonas Geiping, and Tom Goldstein. Tree-rings watermarks: Invisible fingerprints for diffusion images. *Advances in Neural Information Processing Systems*, 36: 58047–58063, 2023.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Marianna Nezhurina, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. *arXiv preprint arXiv:2211.06687*, 2022.

Yi Yang, Haowen Li, Tianxiang Li, Boyu Cao, Xiaohan Zhang, Liqun Chen, and Qi Liu. Melodia: Training-free music editing guided by attention probing in diffusion models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(3):2209–2217, Mar. 2026. doi: 10.1609/aaai.v40i3.37204. URL https://ojs.aaai.org/index.php/AAAI/article/view/37204.

Kaiwen Zhang, Yifan Zhou, Xudong Xu, Bo Dai, and Xingang Pan. Diffmorpher: Unleashing the capability of diffusion models for image morphing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7912–7921, 2024a.

Lijun Zhang, Xiao Liu, Antoni Martin, Cindy Bearfield, Yuriy Brun, and Hui Guan. Attack-resilient image watermarking using stable diffusion. *Advances in Neural Information Processing Systems*, 37:38480–38507, 2024b.

Yixiao Zhang, Yukara Ikemiya, Gus Xia, Naoki Murata, Marco A. Martínez-Ramírez, Wei-Hsiang Liao, Yuki Mitsufuji, and Simon Dixon. Musicmagus: Zero-shot text-to-music editing via diffusion models. In Kate Larson (ed.), *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pp. 7805–7813. International Joint Conferences on Artificial Intelligence Organization, 8 2024c. doi: 10.24963/ijcai.2024/864. URL https://doi.org/10.24963/ijcai.2024/864.

Yuechen Zhang, Jinbo Xing, Eric Lo, and Jiaya Jia. Real-world image variation by aligning diffusion inversion chain. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 30641–30661. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/61960fdfda4d4e95fa1c1f6e64bfe8bc-Paper-Conference.pdf.

Ziyue Zhang, Luxi Lin, Xiaolin Hu, Chao Chang, HuaiXi Wang, Yiyi Zhou, and Rongrong Ji. Deepinv: A novel self-supervised learning approach for fast and accurate diffusion inversion. *arXiv preprint arXiv:2601.01487*, 2026.

PengFei Zheng, Yonggang Zhang, Zhen Fang, Tongliang Liu, Defu Lian, and Bo Han. Noisediffusion: Correcting noise for image interpolation with diffusion models beyond spherical linear interpolation. In *The Twelfth International Conference on Learning Representations*, 2024.

Haina Zhu, Yizhi Zhou, Hangting Chen, Jianwei Yu, Ziyang Ma, Rongzhi Gu, Yi Luo, Wei Tan, and Xie Chen. MuQ: Self-supervised music representation learning with mel residual vector quantization. *arXiv preprint arXiv:2501.01108*, 2025.

A TRAINING AND IMPLEMENTATION DETAILS

Trajectory data. ReCap-COCO is used as a source of text prompts; the associated COCO images are not used for training. For SDXL, we generate one trajectory for each of 27,396 training prompts and reserve 1,522 disjoint prompts for inversion-quality evaluation. For Moonwalker-BLIP, BLIP-2 captions of the 490 StyleBench references are split into 400 training and 90 validation captions. Synthetic trajectories. Each SDXL trajectory is sampled at 1024×1024 with 50 deterministic DDIM steps, CFG 1, and an empty negative prompt. For Stable Diffusion 1.5, we use all 30,441 prompts because evaluation is performed on the independent PIE-Bench dataset. We generate one 512×512 trajectory per prompt with the same number of steps and guidance setting. The SD1.5 trajectories, conditioning tensors, and teacher targets are stored in float32. Each saved transition contains the adjacent latent states, its timestep and text conditioning, and the frozen model prediction used as the regression target.

LoRA configuration and optimization. All backbone, VAE, and text-encoder parameters remain frozen. LoRA is applied to the query, key, value, and output projections of the denoising network, with zero dropout. We use ranks $r \in \{8, 16, 32\}$ with scaling constants $\gamma \in \{4, 8, 16\}$, respectively. Trainable adapter parameters are maintained in float32 and optimized with AdamW using learning rate 5×10^{-5}, zero weight decay, gradient clipping at 1, and a cosine learning-rate schedule.

Training variants. The uniform objective samples all 50 trajectory transitions with equal probability. The final-tail variant keeps the same number of draws but assigns half of them to the final 10% of each saved generation trajectory, nearest the clean endpoint. Cycle-regularized adapters use uniform transition sampling and the objective in Section 3.2. We evaluate cycle weights $\lambda_{\text{cycle}} \in \{0.5, 1, 2, 4, 8, 16\}$. No adapter is active during reconstruction or editing.

PIE-Bench inference. All Stable Diffusion 1.5 baselines and Moonwalker variants use float16 inference, 50 inversion and generation steps, inversion CFG 1, and generation CFG 7.5. Standalone Moonwalker supplies its trajectory directly to the editor. In Moonwalker+DirectInv, that trajectory replaces only the initial DDIM inversion in Direct Inversion; its correction stage and each editor's reverse process are otherwise unchanged.

B EVALUATION PROTOCOL AND METRIC SEMANTICS

The SDXL inversion experiment and PIE-Bench answer different questions. In the SDXL experiment, each source image is generated from a known initial noise, inverted, and reconstructed without an edit. MAE, LPIPS, PSNR, and SSIM therefore compare the reference generation directly with a pure reconstruction. In PIE-Bench, the candidate is the final edited image. Its preservation metrics are computed only inside the benchmark's unedited-region mask, while structure is computed on the full image. Accordingly, the PIE-Bench columns quantify collateral change during editing and should not be interpreted as direct inversion-reconstruction error.

B.1 PIE-BENCH RECONSTRUCTION CONTROL

To evaluate reconstruction directly, independently of editor-specific modifications, we invert each PIE-Bench source image with every selected DirectInv and Moonwalker+DirectInv configuration and then reconstruct the source without applying the target edit. The reported metrics compare this reconstruction directly with the original source image. Table 5 reports the results on the same PIE-Bench inputs. The large P2P gap disappears under Moonwalker+DirectInv, while PnP and Pix2Pix-Zero are nearly unchanged and MasaCtrl shows a small degradation.

Table 5: PIE-Bench source reconstruction without applying the target edit. Values are mean ± standard deviation. Structure, LPIPS, and MSE are minimized; PSNR and SSIM are maximized. Bold denotes the best result within each editor, determined before rounding.

Editor	Configuration	Adapter	Structure ($\times 10^3$) ↓	PSNR (dB) ↑	LPIPS ($\times 10^3$) ↓	MSE ($\times 10^4$) ↓	SSIM ($\times 10^2$) ↑
P2P	DirectInv	–	38.79 ± 26.95	19.65 ± 3.98	158.91 ± 89.79	154.93 ± 128.85	65.52 ± 14.98
P2P	Best uniform	Moonwalker+DirectInv $r = 8$	$\mathbf{3.31 \pm 1.79}$	$\mathbf{27.28 \pm 3.93}$	37.17 ± 14.27	$\mathbf{28.19 \pm 30.29}$	79.99 ± 12.45
P2P	Best final-tail	Moonwalker+DirectInv $r = 8$	3.31 ± 1.81	27.28 ± 3.93	37.17 ± 14.27	28.19 ± 30.29	79.99 ± 12.45
P2P	Cycle loss	Moonwalker+DirectInv $r = 8$	3.31 ± 1.82	27.28 ± 3.93	$\mathbf{37.15 \pm 14.25}$	28.19 ± 30.28	$\mathbf{79.99 \pm 12.45}$
PnP	DirectInv	–	3.38 ± 1.85	$\mathbf{27.26 \pm 3.92}$	37.15 ± 14.23	$\mathbf{28.24 \pm 30.31}$	79.97 ± 12.44
PnP	Best uniform	Moonwalker+DirectInv $r = 8$	3.37 ± 1.84	27.26 ± 3.92	$\mathbf{37.13 \pm 14.21}$	28.24 ± 30.31	$\mathbf{79.97 \pm 12.44}$
PnP	Best final-tail	Moonwalker+DirectInv $r = 32$	3.37 ± 1.82	27.26 ± 3.92	37.15 ± 14.23	28.24 ± 30.31	79.97 ± 12.44
PnP	Cycle loss	Moonwalker+DirectInv $r = 8$	$\mathbf{3.37 \pm 1.83}$	27.26 ± 3.92	37.15 ± 14.25	28.24 ± 30.31	79.97 ± 12.44
MasaCtrl	DirectInv	–	$\mathbf{3.31 \pm 1.84}$	$\mathbf{27.28 \pm 3.93}$	$\mathbf{37.15 \pm 14.26}$	$\mathbf{28.19 \pm 30.28}$	$\mathbf{79.99 \pm 12.45}$
MasaCtrl	Best uniform	Moonwalker+DirectInv $r = 16$	3.69 ± 3.72	27.22 ± 3.92	37.84 ± 15.70	28.50 ± 30.49	79.91 ± 12.46
MasaCtrl	Best final-tail	Moonwalker+DirectInv $r = 32$	3.75 ± 3.87	27.21 ± 3.90	37.95 ± 15.79	28.54 ± 30.52	79.89 ± 12.48
MasaCtrl	Cycle loss	Moonwalker+DirectInv $r = 8$	3.63 ± 3.03	27.22 ± 3.91	37.84 ± 15.29	28.41 ± 30.29	79.91 ± 12.44
Pix2Pix-Zero	DirectInv	–	$\mathbf{3.35 \pm 1.83}$	$\mathbf{27.36 \pm 4.00}$	37.11 ± 14.36	$\mathbf{27.96 \pm 30.29}$	$\mathbf{80.05 \pm 12.46}$
Pix2Pix-Zero	Best uniform	Moonwalker+DirectInv $r = 16$	3.37 ± 1.89	27.36 ± 4.00	$\mathbf{37.11 \pm 14.36}$	27.96 ± 30.29	80.05 ± 12.46
Pix2Pix-Zero	Best final-tail	Moonwalker+DirectInv $r = 8$	3.36 ± 1.87	27.36 ± 4.00	37.11 ± 14.36	27.96 ± 30.28	80.05 ± 12.45
Pix2Pix-Zero	Cycle loss	Moonwalker+DirectInv $r = 8$	3.36 ± 1.89	27.36 ± 4.00	$\mathbf{37.10 \pm 14.35}$	27.96 ± 30.29	80.05 ± 12.46

Table 6: Full-trajectory Moonwalker across adapter ranks. Bold denotes the best result in each column.

Method	Corr. ↓	KL $\times 10^2$ ↓	MAE ↓	LPIPS ↓	PSNR ↑	SSIM ↑
DDIM Inv.	0.1601	0.8566	0.02663	0.09617	28.32	0.8112
Moonwalker $r = 8, \alpha = 4$	0.0999	0.0767	0.02063	0.05982	30.06	0.8654
Moonwalker $r = 16, \alpha = 8$	0.0989	0.0671	0.02028	0.05774	30.22	0.8699
Moonwalker $r = 32, \alpha = 16$	**0.0972**	**0.0504**	**0.01970**	**0.05538**	**30.40**	**0.8743**

C Extended inversion ablations

C.1 Moonwalker rank

C.2 Where the correction acts

Figure 4 and Table 7 activate Moonwalker for only the first k inverse steps. The large discontinuity from one to two steps is consistent across ranks. Five steps recover, on average, 88.8% of the full model's correlation improvement and 81.8% of its KL improvement, but only about one third of the LPIPS gain. Early steps therefore dominate prior matching, whereas later corrections remain necessary for high-fidelity reconstruction.

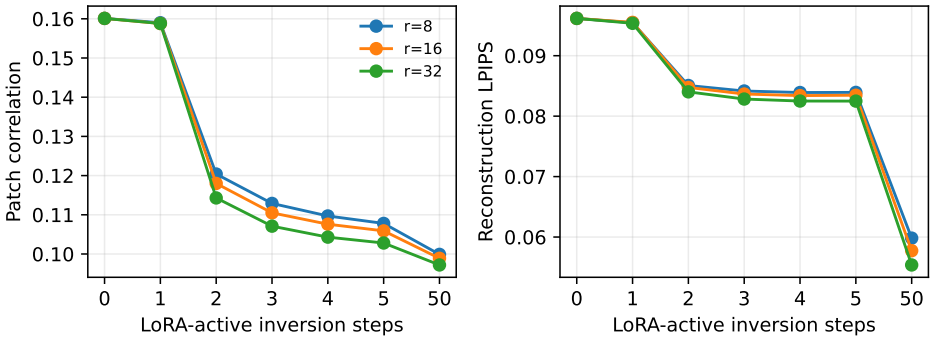


Figure 4: Recovered-noise correlation and reconstruction LPIPS as the number of Moonwalker-active inversion steps increases. The final point uses all 50 steps.

Table 7: Inversion quality when Moonwalker is enabled for only the first k inversion steps. All entries are means.

Adapter	Steps	Corr. ↓	KL $\times 10^2$ ↓	MAE ↓	LPIPS ↓	PSNR ↑	SSIM ↑
None	0	0.1601	0.8566	0.02663	0.09617	28.32	0.8112
$r = 8$	1	0.1590	0.8499	0.02654	0.09546	28.34	0.8123
$r = 8$	2	0.1204	0.3891	0.02508	0.08507	28.65	0.8304
$r = 8$	3	0.1129	0.3007	0.02489	0.08416	28.70	0.8337
$r = 8$	4	0.1097	0.2582	0.02482	0.08391	28.71	0.8349
$r = 8$	5	0.1078	0.2327	0.02481	0.08393	28.71	0.8353
$r = 8$	50	0.0999	0.0767	0.02063	0.05982	30.06	0.8654
$r = 16$	1	0.1588	0.8499	0.02654	0.09551	28.34	0.8122
$r = 16$	2	0.1180	0.3584	0.02508	0.08475	28.64	0.8312
$r = 16$	3	0.1105	0.2687	0.02486	0.08366	28.69	0.8347
$r = 16$	4	0.1076	0.2302	0.02479	0.08340	28.71	0.8359
$r = 16$	5	0.1059	0.2075	0.02478	0.08346	28.70	0.8363
$r = 16$	50	0.0989	0.0671	0.02028	0.05774	30.22	0.8699
$r = 32$	1	0.1588	0.8486	0.02653	0.09536	28.34	0.8124
$r = 32$	2	0.1143	0.3397	0.02497	0.08402	28.67	0.8325
$r = 32$	3	0.1071	0.2502	0.02474	0.08283	28.72	0.8363
$r = 32$	4	0.1043	0.2096	0.02466	0.08250	28.74	0.8377
$r = 32$	5	0.1028	0.1861	0.02463	0.08249	28.74	0.8382
$r = 32$	50	**0.0972**	**0.0504**	**0.01970**	**0.05538**	**30.40**	**0.8743**

Table 8: Fraction of the full-trajectory improvement obtained with five active steps.

Metric	$r = 8$	$r = 16$	$r = 32$	Mean
Corr.	86.9%	88.4%	91.1%	88.8%
KL $\times 10^2$	80.0%	82.2%	83.2%	81.8%
MAE	30.4%	29.2%	28.8%	29.5%
LPIPS	33.7%	33.1%	33.6%	33.4%
CLIP source	69.5%	70.7%	73.2%	71.1%
CLIP target	81.7%	81.2%	82.7%	81.9%
Directional CLIP	88.4%	88.0%	91.9%	89.5%

C.3 Plain and structured regions

C.4 Synthetic text-editing probe

For this controlled probe, each recovered latent is regenerated with a target prompt obtained by cyclically shifting the held-out source-prompt list. Lower CLIP source together with higher CLIP target and directional CLIP indicates greater responsiveness to the target condition. Table 10 shows that most of the gain again appears in the first two to five corrected steps.

D Guidance-scale robustness

Applying the CFG-1 adapter at larger inversion guidance preserves latent normality initially but sharply damages reconstruction (Table 11). This motivates training against the actual conditional/unconditional branches used at the target CFG rather than treating a CFG-1 adapter as guidance-agnostic.

Table 9: Recovered-noise error and latent standard deviation in visually plain and non-plain regions. Values closer to 1 indicate better agreement with the unit-Gaussian diffusion prior.

Adapter	Steps	Plain error ↓	Non-plain error ↓	Plain noise std. → 1	Non-plain noise std. → 1
None	0	0.1956	0.1682	0.8770	0.9573
$r = 8$	1	0.1937	0.1667	0.8773	0.9576
$r = 8$	2	0.1305	0.1204	0.9082	0.9750
$r = 8$	3	0.1196	0.1120	0.9162	0.9801
$r = 8$	4	0.1148	0.1083	0.9203	0.9830
$r = 8$	5	0.1121	0.1061	0.9230	0.9849
$r = 8$	50	0.1015	0.0970	0.9433	1.0022
$r = 16$	1	0.1937	0.1667	0.8773	0.9576
$r = 16$	2	0.1267	0.1175	0.9104	0.9764
$r = 16$	3	0.1155	0.1088	0.9188	0.9819
$r = 16$	4	0.1109	0.1052	0.9228	0.9847
$r = 16$	5	0.1083	0.1031	0.9252	0.9865
$r = 16$	50	0.0982	0.0944	0.9442	1.0029
$r = 32$	1	0.1934	0.1664	0.8774	0.9576
$r = 32$	2	0.1222	0.1140	0.9122	0.9776
$r = 32$	3	0.1107	0.1052	0.9208	0.9832
$r = 32$	4	0.1058	0.1014	0.9251	0.9862
$r = 32$	5	0.1031	0.0992	0.9278	0.9882
$r = 32$	50	**0.0930**	**0.0906**	**0.9479**	1.0055

Table 10: CLIP metrics for the synthetic target-prompt probe.

Adapter	Steps	CLIP source ↓	CLIP target ↑	Directional CLIP ↑
None	0	35.08	67.90	29.85
$r = 8$	1	35.09	67.88	29.85
$r = 8$	2	33.84	70.10	31.22
$r = 8$	3	33.45	70.74	31.53
$r = 8$	4	33.12	71.07	31.71
$r = 8$	5	32.99	71.29	31.80
$r = 8$	50	32.07	72.05	32.05
$r = 16$	1	35.07	67.89	29.86
$r = 16$	2	33.75	70.24	31.29
$r = 16$	3	33.35	70.82	31.59
$r = 16$	4	33.07	71.21	31.77
$r = 16$	5	32.93	71.33	31.82
$r = 16$	50	32.04	72.13	**32.09**
$r = 32$	1	35.09	67.89	29.85
$r = 32$	2	33.75	70.24	31.29
$r = 32$	3	33.24	70.92	31.64
$r = 32$	4	33.02	71.29	31.80
$r = 32$	5	32.83	71.42	31.89
$r = 32$	50	**32.00**	**72.16**	32.07

E CYCLE-CONSISTENCY STRENGTH

Table 13 reports the complete rank-16 cycle-weight sweep. P2P benefits most from a small cycle weight, with $\lambda_{\text{cycle}} = 0.5$ giving the strongest preservation. PnP is nearly insensitive to the tested range. For MasaCtrl, changing λ_{cycle} does not close the preservation gap to DirectInv, although several settings improve CLIP alignment. DirectInv remains stronger than every rank-16 cycle setting for Pix2Pix-Zero, indicating that cycle strength alone does not resolve that editor's inversion–attention mismatch.

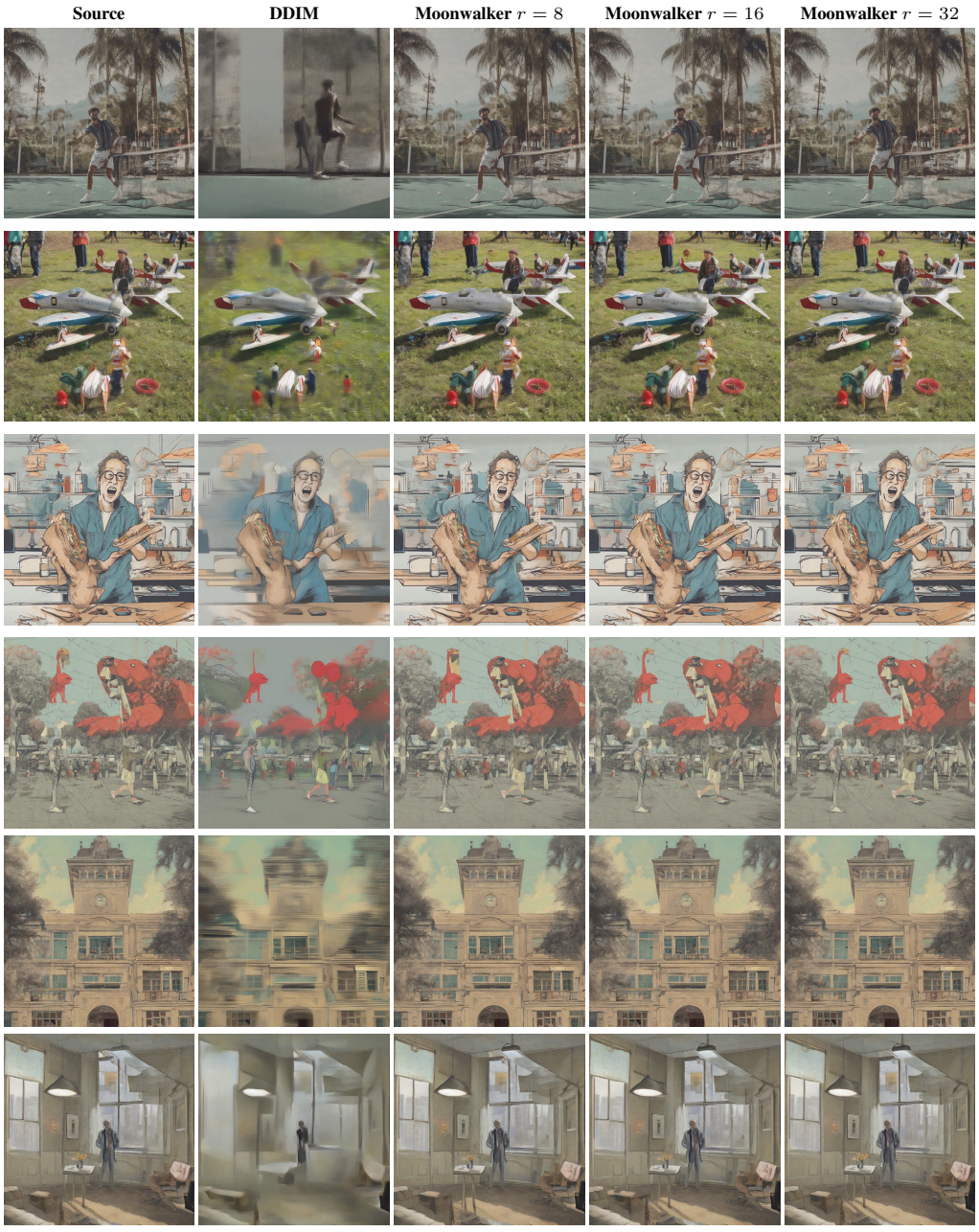


Figure 5: Qualitative SDXL reconstructions at CFG 1. Columns show the generated source image, DDIM reconstruction, and Moonwalker reconstructions for ranks 8, 16, and 32. All methods use 50 inversion and reconstruction steps; Moonwalker is active only during inversion.

Table 11: Applying Moonwalker ($r = 32$) at different inversion CFG values. All entries are means.

Method	CFG	Corr. ↓	KL $\times 10^2$ ↓	MAE ↓	LPIPS ↓	PSNR ↑	SSIM ↑
DDIM Inv.	1	0.1601	0.8566	0.02663	0.09617	28.32	0.8112
Moonwalker $r = 32$	1	0.0972	0.0504	**0.01970**	**0.05538**	**30.40**	**0.8743**
Moonwalker $r = 32$	2	**0.0955**	**0.0150**	0.07249	0.29056	21.16	0.6442
Moonwalker $r = 32$	3	0.0989	0.0607	0.26066	0.64225	11.36	0.3292
Moonwalker $r = 32$	4	0.1124	0.1500	0.32729	0.72783	9.10	0.2355
Moonwalker $r = 32$	5	0.1331	0.2499	0.34274	0.75096	8.52	0.2063

Table 12: CFG-4 inversion on CFG-4 generated data (100-sample diagnostic). Noise standard deviation is a recovered-latent metric, not an uncertainty term.

Method	Noise std. $\rightarrow 1$	Corr. ↓	KL ↓	MAE ↓	LPIPS ↓	PSNR ↑	SSIM ↑
DDIM Inv.	0.8812	0.3669	2.2262	0.1725	0.5900	13.76	0.3566
Moonwalker (shared)	**0.9828**	**0.3599**	**0.0838**	**0.0557**	**0.1939**	**22.09**	**0.7522**
Moonwalker (paired)	0.9562	0.3611	0.3306	0.0737	0.2489	20.40	0.7026

Table 13: Complete cycle-loss sweep for rank-16 Moonwalker+DirectInv on PIE-Bench. Bold denotes the best result within each editor, including the reproduced SD1.5 DirectInv control. Italics denote the best cycle-loss setting for each metric.

Editor	λ_{cycle}	Structure ($\times 10^3$) ↓	PSNR ↑	LPIPS ($\times 10^3$) ↓	MSE ($\times 10^4$) ↓	SSIM ($\times 10^2$) ↑	CLIP target ↑	CLIP edit ↑
P2P	DirectInv	31.35	22.57	125.51	91.27	78.54	25.07	22.22
P2P	0.5	***14.00***	***26.05***	***71.76***	***43.57***	***83.60***	25.28	22.31
P2P	1	14.45	25.92	74.51	45.37	83.37	25.27	22.31
P2P	2	14.41	25.92	74.10	44.87	83.39	***25.32***	***22.36***
P2P	4	14.23	26.04	73.38	44.43	83.44	***25.32***	22.30
P2P	8	14.36	25.95	73.39	44.87	83.45	***25.32***	22.32
P2P	16	14.24	26.01	72.52	43.61	83.56	25.31	22.31
PnP	DirectInv	**23.33**	**22.46**	**105.51**	**79.97**	**80.22**	25.63	22.67
PnP	0.5	*24.38*	*22.36*	110.27	81.87	*79.88*	25.81	***22.83***
PnP	1	*24.38*	*22.36*	110.28	81.87	*79.88*	25.81	***22.83***
PnP	2	*24.38*	*22.36*	110.27	81.87	*79.88*	25.81	***22.83***
PnP	4	*24.38*	*22.36*	110.27	81.87	*79.88*	25.81	22.82
PnP	8	24.39	*22.36*	110.26	81.86	*79.88*	***25.82***	***22.83***
PnP	16	24.39	*22.36*	*110.25*	*81.85*	*79.88*	25.81	***22.83***
MasaCtrl	DirectInv	**23.57**	**22.82**	**86.83**	**79.19**	**82.02**	24.51	21.47
MasaCtrl	0.5	28.52	21.89	98.93	99.02	80.50	24.61	21.47
MasaCtrl	1	28.49	21.89	98.91	98.95	80.50	***24.63***	21.48
MasaCtrl	2	28.53	*21.90*	98.83	99.05	80.50	24.62	***21.50***
MasaCtrl	4	28.51	*21.90*	98.89	*98.91*	80.50	24.62	***21.50***
MasaCtrl	8	*28.47*	*21.90*	*98.77*	*98.91*	*80.52*	24.59	21.47
MasaCtrl	16	28.49	21.89	98.88	*98.91*	80.50	***24.63***	***21.50***
Pix2Pix-Zero	DirectInv	**49.99**	**21.47**	**142.46**	**133.49**	**77.49**	**23.53**	**21.09**
Pix2Pix-Zero	0.5	52.67	21.26	145.88	138.21	77.07	*23.43*	21.04
Pix2Pix-Zero	1	52.16	21.22	147.00	140.33	76.95	*23.43*	20.98
Pix2Pix-Zero	2	53.06	21.25	146.22	139.69	76.97	23.33	20.87
Pix2Pix-Zero	4	52.06	21.27	146.70	139.76	77.02	23.41	*21.08*
Pix2Pix-Zero	8	52.38	*21.29*	146.18	139.26	77.03	*23.43*	20.93
Pix2Pix-Zero	16	*51.41*	21.26	*145.77*	*137.38*	*77.12*	23.33	20.89

Table 14: Reconstruction quality and edit alignment of DDIM and mixed-CFG-trained Moonwalker variants with equal inversion and generation CFG. MSE, MAE and CLIP-S are minimized; PSNR, SSIM, CLIP-T and Dir. CLIP are maximized. All results are means. Bold denotes the best reported result within each CFG setting.

$w_{\text{inv}} = w_{\text{gen}}$	Inversion	MSE ↓	MAE ↓	PSNR ↑	SSIM ↑	CLIP-S ↓	CLIP-T ↑	Dir. CLIP ↑
2.0	DDIM	0.00827	0.0561	23.16	0.6563	33.49	74.08	34.08
	Moonwalker shared, $\mathcal{L}_{\text{cfg}}$	**0.00347**	**0.0341**	**26.09**	**0.8149**	30.14	80.20	**36.94**
	Moonwalker shared, $\mathcal{L}_{\text{branch}}$	0.00355	0.0349	25.61	0.8075	**29.67**	**80.47**	36.93
	Moonwalker paired, $\mathcal{L}_{\text{cfg}}$	0.02803	0.1352	15.70	0.5356	39.88	40.02	18.29
	Moonwalker paired, $\mathcal{L}_{\text{branch}}$	0.00531	0.0460	23.71	0.7347	29.85	79.80	36.79
3.0	DDIM	0.03172	0.1248	16.76	0.4171	35.83	72.24	32.61
	Moonwalker shared, $\mathcal{L}_{\text{cfg}}$	**0.00613**	**0.0427**	**24.49**	**0.7952**	29.60	81.94	37.81
	Moonwalker shared, $\mathcal{L}_{\text{branch}}$	0.00758	0.0503	22.76	0.7687	**28.80**	**82.36**	**37.97**
	Moonwalker paired, $\mathcal{L}_{\text{cfg}}$	0.02984	0.1374	15.44	0.5448	39.23	40.15	19.23
	Moonwalker paired, $\mathcal{L}_{\text{branch}}$	0.01657	0.0792	19.99	0.6474	30.06	80.47	36.98
4.0	DDIM	0.05467	0.1736	13.68	0.3464	36.72	67.66	30.82
	Moonwalker shared, $\mathcal{L}_{\text{cfg}}$	**0.01351**	**0.0658**	**21.40**	**0.7392**	**28.76**	82.58	38.40
	Moonwalker shared, $\mathcal{L}_{\text{branch}}$	0.01616	0.0764	19.77	0.6985	28.78	**83.37**	**38.60**
	Moonwalker paired, $\mathcal{L}_{\text{cfg}}$	0.03115	0.1373	15.26	0.5440	40.15	43.05	20.54
	Moonwalker paired, $\mathcal{L}_{\text{branch}}$	0.04813	0.1477	15.48	0.5134	30.88	77.53	36.03
5.0	DDIM	0.07277	0.2047	12.07	0.3355	36.02	64.94	30.04
	Moonwalker shared, $\mathcal{L}_{\text{cfg}}$	**0.02578**	**0.0977**	**18.49**	**0.6567**	29.48	81.98	38.13
	Moonwalker shared, $\mathcal{L}_{\text{branch}}$	0.03562	0.1221	16.33	0.5979	**29.24**	**82.74**	**38.28**
	Moonwalker paired, $\mathcal{L}_{\text{cfg}}$	0.03417	0.1413	14.89	0.5414	39.74	48.14	22.66
	Moonwalker paired, $\mathcal{L}_{\text{branch}}$	0.09919	0.2389	11.30	0.3538	33.69	69.37	32.19
7.5	DDIM	0.10139	0.2468	10.30	0.3163	37.10	60.04	28.53
	Moonwalker shared, $\mathcal{L}_{\text{cfg}}$	0.06526	0.1767	13.39	0.5101	**30.94**	**77.90**	**36.70**
	Moonwalker shared, $\mathcal{L}_{\text{branch}}$	0.10671	0.2462	10.62	0.3962	32.53	71.98	33.87
	Moonwalker paired, $\mathcal{L}_{\text{cfg}}$	**0.04216**	**0.1552**	**14.01**	**0.5401**	38.92	69.17	30.96
	Moonwalker paired, $\mathcal{L}_{\text{branch}}$	0.18254	0.3574	7.52	0.1869	37.73	51.68	24.42
10.0	DDIM	0.12182	0.2711	9.37	0.2946	37.99	55.59	26.80
	Moonwalker shared, $\mathcal{L}_{\text{cfg}}$	0.11570	0.2603	9.78	**0.3787**	**34.58**	68.99	**32.48**
	Moonwalker shared, $\mathcal{L}_{\text{branch}}$	0.15604	0.3131	8.37	0.2888	35.92	62.21	29.44
	Moonwalker paired, $\mathcal{L}_{\text{cfg}}$	**0.09063**	**0.2320**	**10.72**	0.3506	40.19	**70.83**	30.83
	Moonwalker paired, $\mathcal{L}_{\text{branch}}$	0.19553	0.3669	7.15	0.1688	37.75	45.89	22.54

Table 15: Inversion quality and edit alignment of DDIM and guided Moonwalker variants. Correlation, KL divergence and LPIPS are minimized; PSNR, SSIM and CLIP scores are maximized. All results are means. Bold denotes the best reported mean within each generation-guidance setting.

CFG gen.	Inversion	Corr. ↓	KL ↓	LPIPS ↓	PSNR ↑	SSIM ↑	CLIP-T ↑	Dir. CLIP ↑
5	DDIM	0.18093	2.42355	0.18965	19.89	0.7444	80.50	37.17
	Moonwalker single-pass	0.12247	0.21056	0.18793	19.77	0.7482	84.28	39.03
	Moonwalker shared, $\mathcal{L}_{\text{cfg}}$	**0.11746**	**0.03973**	0.18818	19.91	0.7468	**84.60**	**39.08**
	Moonwalker shared, $\mathcal{L}_{\text{branch}}$	0.12135	0.23944	**0.17634**	**20.31**	**0.7612**	84.04	39.00
	Moonwalker paired, $\mathcal{L}_{\text{cfg}}$	0.11895	0.13086	0.21203	19.31	0.7261	84.23	38.93
	Moonwalker paired, $\mathcal{L}_{\text{branch}}$	0.12108	0.19438	0.17797	20.27	0.7595	84.20	38.99
7.5	DDIM	0.18670	2.85674	0.22280	17.73	0.6931	81.76	37.97
	Moonwalker single-pass	0.12402	0.26300	0.21872	17.79	0.6975	85.04	39.62
	Moonwalker shared, $\mathcal{L}_{\text{cfg}}$	**0.11798**	**0.04718**	0.22357	17.81	0.6918	**85.36**	39.59
	Moonwalker shared, $\mathcal{L}_{\text{branch}}$	0.12282	0.30069	**0.20854**	**18.17**	**0.7095**	85.27	**39.69**
	Moonwalker paired, $\mathcal{L}_{\text{cfg}}$	0.11960	0.13733	0.24335	17.35	0.6748	85.23	39.61
	Moonwalker paired, $\mathcal{L}_{\text{branch}}$	0.12257	0.25404	0.21113	18.13	0.7078	85.17	39.60
10	DDIM	0.19018	3.12596	0.25287	16.30	0.6473	82.12	38.16
	Moonwalker single-pass	0.12523	0.30958	0.24565	16.44	0.6507	85.40	**39.84**
	Moonwalker shared, $\mathcal{L}_{\text{cfg}}$	**0.11833**	**0.05570**	0.25186	16.41	0.6448	**85.75**	**39.84**
	Moonwalker shared, $\mathcal{L}_{\text{branch}}$	0.12400	0.35233	**0.23606**	**16.74**	**0.6620**	85.21	39.75
	Moonwalker paired, $\mathcal{L}_{\text{cfg}}$	0.12022	0.14363	0.27316	15.94	0.6267	85.57	39.79
	Moonwalker paired, $\mathcal{L}_{\text{branch}}$	0.12368	0.30700	0.23881	16.71	0.6606	85.43	39.79

Table 16: Effect of separating inversion and generation guidance for DDIM. LPIPS is minimized; PSNR, SSIM and CLIP scores are maximized. All results are means. Bold denotes the best result within each generation-guidance setting.

CFG inv.	CFG gen.	Inversion	LPIPS ↓	PSNR ↑	SSIM ↑	CLIP-T ↑	Dir. CLIP ↑
5	5	DDIM	0.6285	12.07	0.3355	64.94	30.04
2	5	DDIM	**0.1871**	**20.07**	**0.7450**	**80.10**	**37.08**
7.5	7.5	DDIM	0.6484	10.30	0.3163	60.04	28.53
2	7.5	DDIM	**0.2237**	**17.79**	**0.6899**	**82.10**	**38.21**
10	10	DDIM	0.6670	9.37	0.2946	55.59	26.80
2	10	DDIM	**0.2539**	**16.38**	**0.6429**	**81.92**	**38.62**

Table 17: Noise diagnostics, reconstruction and editing quality for FLUX inversion at different guidance settings. Correlation, KL divergence, MAE, LPIPS and source CLIP are minimized; the remaining metrics are maximized. All results are means. Bold denotes the best reported mean within each block.

Inversion	Corr. ↓	KL ↓	MAE ↓	LPIPS ↓	PSNR ↑	SSIM ↑	CLIP-S ↓	CLIP-T ↑	Dir. CLIP ↑
$g_{\text{src}} = g_{\text{inv}} = g_{\text{rec}} = g_{\text{edit}} = 1$									
Euler	0.13795	0.3796	0.08970	0.50005	18.97	0.63897	33.949	78.456	33.486
Nudging $\times 1.5$	0.28697	6.1737	0.03841	0.17837	26.17	0.84902	63.451	47.924	15.849
Moonwalker	**0.13650**	**0.0397**	**0.01733**	**0.07333**	**32.43**	**0.95713**	**32.033**	**80.378**	**34.845**
$g_{\text{src}} = g_{\text{inv}} = g_{\text{rec}} = g_{\text{edit}} = 3.5$									
Euler	0.18743	3.9522	0.21394	0.72322	11.20	0.43764	**33.071**	**73.917**	**32.813**
Nudging $\times 1.5$	0.37150	3.2313	**0.06833**	**0.15837**	**22.93**	**0.80523**	72.051	40.391	9.840
Moonwalker	**0.14628**	**0.8322**	0.26107	0.75262	9.47	0.39155	33.718	67.161	30.308
$g_{\text{src}} = g_{\text{rec}} = g_{\text{edit}} = 3.5,\quad g_{\text{inv}} = 1$									
Euler	0.17916	4.2695	0.06236	0.17645	20.85	**0.81535**	31.463	78.916	35.252
Nudging $\times 1.5$	0.36462	3.2371	0.09074	**0.15229**	19.56	0.78944	78.477	36.375	6.636
Moonwalker	**0.14830**	**1.4746**	**0.06115**	0.17746	**20.97**	0.80636	**29.028**	**80.415**	**36.655**

Table 18: Full-image reconstruction, background preservation after editing, and edit alignment for FLUX on PIE-Bench. LPIPS, MSE and MAE are minimized; PSNR, SSIM and CLIP scores are maximized. All results are means. Bold denotes the best result in each column.

	Full-image reconstruction					Background after editing				CLIP-T	
Inversion	PSNR ↑	LPIPS ↓	MSE ↓	SSIM ↑	MAE ↓	PSNR ↑	LPIPS ↓	MSE ↓	SSIM ↑	Whole ↑	Mask ↑
Euler	16.88	329.96	284.88	60.14	0.11741	17.28	245.79	281.38	69.99	23.68	21.44
Moonwalker	20.95	194.15	180.07	75.42	0.07561	18.59	197.14	257.52	75.98	**24.33**	**21.73**
Nudging $\times 1.5$	**22.10**	**82.27**	**115.19**	**79.11**	**0.06248**	**23.55**	**67.29**	**101.47**	**84.94**	23.20	20.38

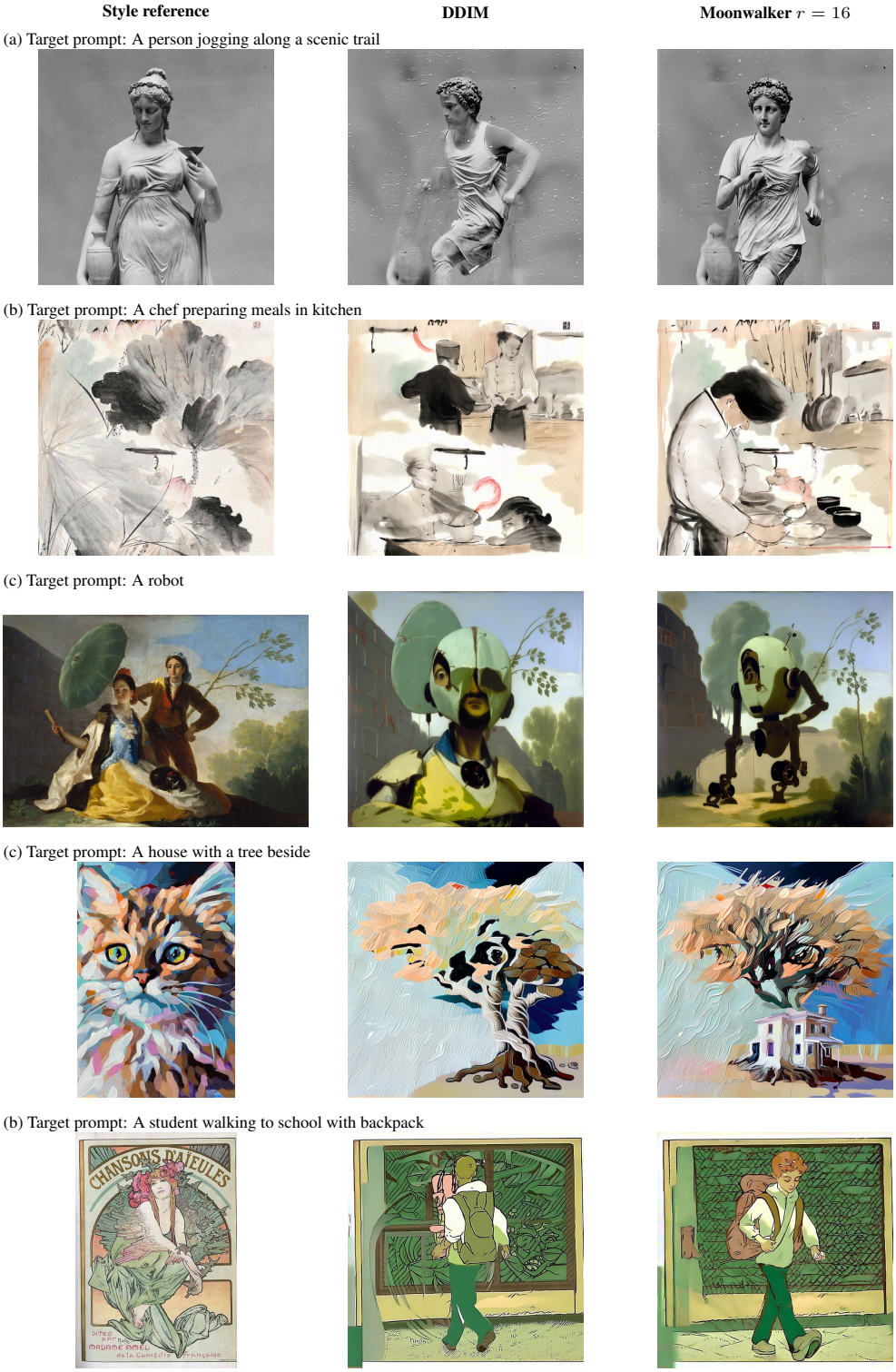


Figure 6: Inversion-InstantStyle at $s = 0$.