FREE LUNCH IN DIFFUSION INVERSION: TENSORIAL CLASSIFIER-FREE GUIDANCE AT ITS ORIGIN

Anonymous authors
Paper under double-blind review

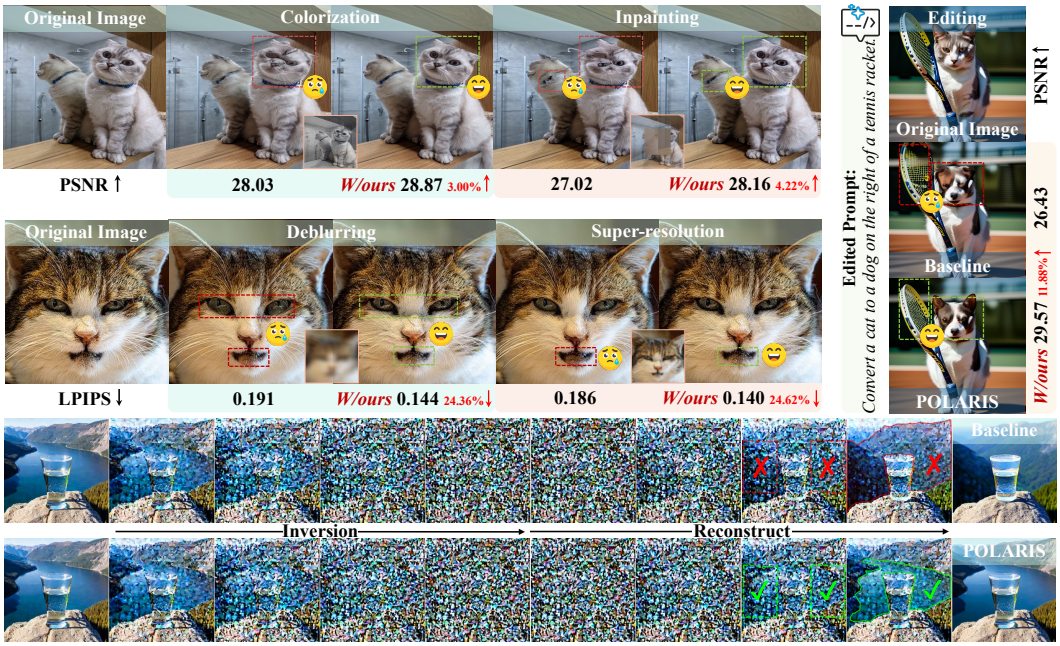

Figure 1: POLARIS serves as a superior inversion foundation, significantly enhancing the performance of various downstream diffusion tasks. (**Top two rows**) Image restoration: In colorization, inpainting, deblurring, and super-resolution tasks, all results show significant fidelity improvements when POLARIS is used as the inversion foundation. (**Right panel**) Image editing: In the "cat-to-dog" editing task, POLARIS enables the model to successfully execute this complex object replacement. (**Bottom row**) Reconstruction: POLARIS achieves near-perfect inversion and reconstruction, providing a robust and high-fidelity foundation for downstream tasks.

ABSTRACT

Inversion errors can compromise source preservation before semantic editing begins. We propose POLARIS, a lightweight, training-free guidance plugin for improving the source trajectory in diffusion-based editing. Building on reconstruction-residual control in diffusion inversion, POLARIS specializes the local inverse-consistency objective to the classifier-free guidance (CFG) coefficients. With adjacent network predictions held fixed, the deterministic DDIM cycle residual is a timestep-dependent scalar multiple of the adjacent guided-noise discrepancy. This yields a coordinate-wise least-squares update using existing conditional and unconditional predictions, without gradient-based optimization or an auxiliary network. Structured-guidance ablations and measured cycle residuals support the local analysis, while comparisons with both high-CFG and null-branch fixed-scale controls demonstrate improved reconstruction fidelity. Paired evaluations further show better background preservation when POLARIS is integrated with existing image, audio, and video editors. A separate transport-based extension improves background fidelity on Wan2.1 video editing without additional model evaluations. Across the measured image backbones, peak memory overhead is 2.0–3.2%. These results support adaptive guidance as a practical component for source preservation; improved inversion alone does not guarantee target-semantic editing.

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.

1 INTRODUCTION

Recent advances in diffusion and flow-based generative models Ho et al. (2020); Song et al. (2022); Rombach et al. (2022); Song et al. (2021); Lipman et al. (2023) have improved text-driven editing across images, videos, and audio Hertz et al. (2022); Cao et al. (2023); Parmar et al. (2023); Tumanyan et al. (2022); Meng et al. (2022); Couairon et al. (2022); Yang et al. (2022); Avrahami et al. (2023); Brooks et al. (2022). Many editing pipelines follow an *inversion–editing* paradigm: a source sample is inverted into a latent trajectory and then manipulated according to a target instruction. Inversion errors can therefore propagate into unintended changes in regions that should remain unchanged.

Prior work addresses inversion drift through conditioning adaptation Mokady et al. (2022); Dong et al. (2023); Miyake et al. (2024), trajectory correction Ju et al. (2023), and iterative re-noising Garibi et al. (2024). FreeInv Bao et al. (2025) reduces trajectory mismatch through matched random latent transformations during inversion and reconstruction, motivated by an ensemble interpretation. Building on this reconstruction-consistency perspective, we investigate analytic adaptation of classifier-free guidance Ho and Salimans (2022) (CFG) using existing conditional and unconditional predictions.

For deterministic DDIM, the one-step inversion–denoising cycle residual is proportional to the discrepancy between adjacent guided-noise predictions. A fixed CFG scale applies one coefficient across timesteps and latent coordinates, despite their heterogeneous discrepancies. Reducing this scalar alone cannot address the coordinate-dependent structure of inversion error.

We propose **POLARIS**, short for **P**rojection-**O**rthogonal **L**east Squ**a**res for **R**obust **I**nver**s**ion. POLARIS treats CFG coefficients as local inverse-consistency controllers. With adjacent network predictions held fixed, it solves a separable least-squares problem for a tensor-valued guidance update. The closed form reuses existing predictions without training, gradient-based optimization, or auxiliary networks. Its contribution is this analytic CFG controller within the established reconstruction-consistency framework.

Scalar, channel-wise, and spatial guidance constrain coordinates to share coefficients. Tensor guidance contains these restricted solutions and thus attains no larger minimum of the same frozen-prediction objective. Under a Lipschitz assumption on denoising transitions, accumulated local cycle defects also bound final latent reconstruction error. Measured residuals and guidance-granularity ablations support this local analysis. In DDIM-based editing, POLARIS improves the source-preservation trajectory while retaining the editor's target-branch controls; successful semantic changes still depend on the downstream editor.

We evaluate reconstruction on COCO2017 Lin et al. (2015) and Pick-a-Pic Kirstain et al. (2023), including a fixed-$\omega = 0$ null-branch control alongside high-CFG inversion. Paired comparisons on PIE-Bench Ju et al. (2023) cover DI+P2P, DCI+P2P, SAGE, RF-Edit, and DNAEdit, assessing masked-background preservation and target-prompt alignment separately. AudioEditor Jia et al. (2024) and OpenSora-based RF-Edit Wang et al. (2024b) extend the evaluation across modalities. For inversion-free FlowEdit Kulikov et al. (2025a), a distinct *Transport-POLARIS* extension improves background fidelity on Wan2.1 Team Wan et al. (2025) using a 100-edit FiVE subset Li et al. (2025a), without additional Wan model evaluations. Peak GPU-memory overhead on the tested image backbones is 2.0–3.2%.

Our **main contributions** are as follows:

i. We derive a coordinate-wise, closed-form CFG controller from local inversion consistency and analyze its relationship to structured guidance.

ii. We validate the controller through measured cycle defects, guidance-granularity ablations, null-branch controls, and paired editing comparisons.

iii. We demonstrate preservation gains across image, audio, and video pipelines, including a transport extension on Wan2.1, and characterize computational costs and applicability limits.

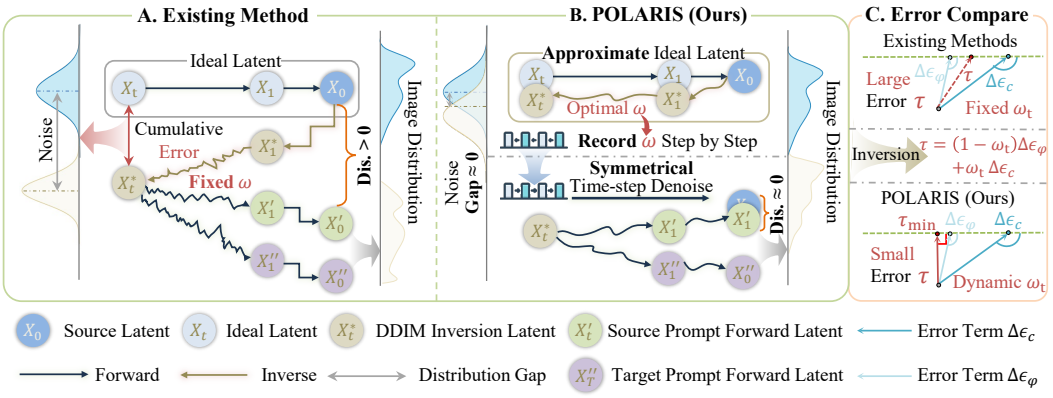

Figure 2: Overview of POLARIS. (A) Adjacent prediction mismatch causes local DDIM inversion errors. (B) Recorded tensor guidance is replayed at matching denoising timesteps for source reconstruction. (C) The controller minimizes a local least-squares residual with fixed endpoint predictions.

2 PRELIMINARY

Deterministic DDIM inversion. Let $\mathbf{x}_t$ denote the latent variable at timestep t, and let $\bar{\alpha}_t$ be the cumulative noise schedule. Define

$$\sigma_t = \sqrt{1-\bar{\alpha}_t}, \qquad \gamma_{t+1} = \frac{\sqrt{\bar{\alpha}_{t+1}}}{\sqrt{\bar{\alpha}_t}}, \qquad \eta_{t+1} = \sigma_{t+1} - \gamma_{t+1}\sigma_t. \tag{1}$$

For deterministic DDIM, the inversion step $\mathbf{x}_t \to \mathbf{x}_{t+1}$ and the corresponding denoising step $\mathbf{x}_{t+1} \to \mathbf{x}_t$ can be written as affine maps:

$$\mathfrak{I}_{t\to t+1}(\mathbf{x}, \varepsilon_0) = \gamma_{t+1}\mathbf{x} + \eta_{t+1}\varepsilon_0, \tag{2}$$

$$\mathfrak{D}_{t+1\to t}(\mathbf{x}, \varepsilon_1) = \gamma_{t+1}^{-1}\mathbf{x} + \bar{\eta}_{t+1}\varepsilon_1, \qquad \bar{\eta}_{t+1} = -\gamma_{t+1}^{-1}\eta_{t+1}. \tag{3}$$

Combining the two maps yields the one-step self-consistency residual:

$$\mathfrak{D}_{t+1\to t}\left(\mathfrak{I}_{t\to t+1}(\mathbf{x}, \varepsilon_0), \varepsilon_1\right) - \mathbf{x} = \bar{\eta}_{t+1}\left(\varepsilon_1 - \varepsilon_0\right). \tag{4}$$

Thus, local inversion drift is directly controlled by the adjacent noise-prediction mismatch $\varepsilon_1 - \varepsilon_0$, which motivates our error-origin objective in Sec. 3.

Classifier-free guidance. Let

$$\varepsilon_\varphi^t = \varepsilon_\theta(\tilde{\mathbf{x}}_t, \varphi, t), \qquad \varepsilon_c^t = \varepsilon_\theta(\tilde{\mathbf{x}}_t, c, t), \tag{5}$$

where φ and c denote the null and text conditions, respectively. Classifier-free guidance combines them as

$$\tilde{\varepsilon}_t(\omega) = (\mathbf{1} - \omega) \odot \varepsilon_\varphi^t + \omega \odot \varepsilon_c^t. \tag{6}$$

Standard CFG uses a scalar scale $\omega = s\mathbf{1}$, while POLARIS optimizes ω in the latent tensor space where the local inversion error is defined.

3 METHODOLOGY

Let $t = 0, \ldots, T$ index increasing noise levels and $\tilde{\mathbf{x}}_t$ denote the realized inversion latent. The unconditional and source-conditioned predictions are $\varepsilon_{\varphi,t} = \varepsilon_\theta(\tilde{\mathbf{x}}_t, t, \varphi)$ and $\varepsilon_{c,t} = \varepsilon_\theta(\tilde{\mathbf{x}}_t, t, c)$, respectively. The deterministic DDIM maps $\mathfrak{I}_{t\to t+1}$ and $\mathfrak{D}_{t+1\to t}$ take a latent and a noise estimate to the next noisier and less noisy latent, respectively, and are affine in these supplied inputs.

3.1 TENSOR-LEVEL ERROR-ORIGIN OBJECTIVE

For $t \to t+1$, apply a trial guidance tensor ω to both endpoints while holding their network predictions fixed. Define the local cycle residual

$$\mathbf{e}_{t+1}(\omega) := \mathfrak{D}_{t+1\to t}\left(\mathfrak{I}_{t\to t+1}\left(\tilde{\mathbf{x}}_t, \tilde{\varepsilon}_t(\omega)\right), \tilde{\varepsilon}_{t+1}(\omega)\right) - \tilde{\mathbf{x}}_t. \tag{7}$$

Composing the affine DDIM maps gives

$$\mathbf{e}_{t+1}(\omega) = \bar{\eta}_{t+1}\tau_{t+1}(\omega), \tag{8}$$

where $\bar{\eta}_{t+1}$ is the denoising map's noise coefficient, independent of ω, and

$$\tau_{t+1}(\omega) = (\mathbf{1} - \omega) \odot \Delta\varepsilon_{\varphi,t+1} + \omega \odot \Delta\varepsilon_{c,t+1}. \tag{9}$$

Here,

$$\Delta\varepsilon_{\varphi,t+1} = \varepsilon_{\varphi,t+1} - \varepsilon_{\varphi,t}, \qquad \Delta\varepsilon_{c,t+1} = \varepsilon_{c,t+1} - \varepsilon_{c,t}. \tag{10}$$

For $\bar{\eta}_{t+1} \neq 0$,

$$\arg\min_{\omega} \|\mathbf{e}_{t+1}(\omega)\|_2^2 = \arg\min_{\omega} J_{t+1}(\omega), \qquad J_{t+1}(\omega) = \|\tau_{t+1}(\omega)\|_2^2. \tag{11}$$

Expanding the objective yields

$$J_{t+1}(\omega) = \left\|\Delta\varepsilon_{\varphi,t+1} - \omega \odot (\Delta\varepsilon_{\varphi,t+1} - \Delta\varepsilon_{c,t+1})\right\|_2^2. \tag{12}$$

Appendix A.1 derives this fixed-prediction identity and explains its distinction from the nonlinear rollout.

3.2 POLARIS: COORDINATE-WISE CLOSED-FORM GUIDANCE

For image latents, POLARIS uses

$$\omega_t \in \mathbb{R}^{C\times H\times W}, \tag{13}$$

with guided prediction

$$\tilde{\varepsilon}_t(\omega) = (\mathbf{1} - \omega) \odot \varepsilon_{\varphi,t} + \omega \odot \varepsilon_{c,t}. \tag{14}$$

The tensor shape follows the modality. Define

$$\delta_{t+1} = \Delta\varepsilon_{\varphi,t+1} - \Delta\varepsilon_{c,t+1}. \tag{15}$$

Then

$$J_{t+1}(\omega) = \left\|\Delta\varepsilon_{\varphi,t+1} - \omega \odot \delta_{t+1}\right\|_2^2. \tag{16}$$

Flattening into $d = CHW$ coordinates gives

$$J_{t+1}(\omega) = \sum_{i=1}^{d} \left(\Delta\varepsilon_{\varphi,t+1,i} - \omega_i\delta_{t+1,i}\right)^2, \tag{17}$$

so each coordinate solves

$$\min_{\omega_i} \left(\Delta\varepsilon_{\varphi,t+1,i} - \omega_i\delta_{t+1,i}\right)^2. \tag{18}$$

For $\delta_{t+1,i} \neq 0$, its minimizer is

$$\omega_{t+1,i}^{\mathrm{POLARIS}} = \frac{\Delta\varepsilon_{\varphi,t+1,i}}{\Delta\varepsilon_{\varphi,t+1,i} - \Delta\varepsilon_{c,t+1,i}}. \tag{19}$$

When $\delta_{t+1,i} = 0$, any coefficient minimizes the coordinate objective. For stability, we reuse the previous value when $|\delta_{t+1,i}| \leq \epsilon_{\mathrm{num}}$; this fallback need not be optimal for nonzero guarded denominators.

Relation to scalar CFG. Scalar CFG imposes $\omega = s\mathbf{1}$; channel-wise and spatial variants tie smaller groups of coordinates. Their feasible sets lie within tensor guidance, whose optimal residual is therefore no larger for fixed predictions. Channel-wise and spatial variants are not universally ordered (Appendix A.3). Only inference-time coefficients change, so no retraining is required.

Reconstruction error. Initialize reconstruction with $\widehat{\mathbf{x}}_T = \tilde{\mathbf{x}}_T$. If denoising transition $k \to k-1$ is L_k-Lipschitz and has actual cycle-defect norm r_k on the inversion trajectory, then

$$\|\widehat{\mathbf{x}}_0 - \tilde{\mathbf{x}}_0\|_2 \le \sum_{k=1}^{T} \left(\prod_{j=1}^{k-1} L_j \right) r_k. \tag{20}$$

Appendix A.2 proves this latent-space bound. It does not guarantee improvements in decoded-image metrics or target-semantic success.

Implementation. Initialize $\omega_0 = s_0\mathbf{1}$, update from cached adjacent predictions, and store $\{\omega_t^{\text{POLARIS}}\}_{t=1}^{T}$. Replay follows reverse timestep order on the source-preservation branch, retaining the editor's target guidance and semantic controls. Updates cost $O(d)$ per step, with $O(Td)$ guidance storage. The closed form requires distinct unconditional and conditional predictions.

3.3 Transport-POLARIS for Flow-Based Editing

FlowEdit Kulikov et al. (2025a) edits through a source–target velocity difference without constructing a DDIM inversion cycle. We extend the local least-squares principle to this difference with **Transport-POLARIS**, which jointly adjusts source and target guidance. Let k index successive sampling steps and let $\mathbf{v}^u_{b,k}$ and $\mathbf{v}^c_{b,k}$ be the unconditional and conditional flow predictions for branch $b \in \{s,t\}$. Define the guidance residuals and baseline transport direction as

$$\mathbf{f}_{b,k} = \mathbf{v}^c_{b,k} - \mathbf{v}^u_{b,k}, \quad \mathbf{d}^0_k = (\mathbf{v}^u_{t,k} + g^0_t \mathbf{f}_{t,k}) - (\mathbf{v}^u_{s,k} + g^0_s \mathbf{f}_{s,k}), \tag{21}$$

where $(g^0_s, g^0_t) = (5, 10)$ in the Wan2.1 experiments. For each latent channel, flatten the temporal and spatial coordinates and write the guidance adjustments as $g_s = g^0_s + a_s$ and $g_t = g^0_t + a_t$. With the current predictions and a cached transport predictor $\widehat{\mathbf{d}}_k$ held fixed,

$$\mathbf{r}_k = \mathbf{d}^0_k - \widehat{\mathbf{d}}_k, \quad G_k = [-\mathbf{f}_{s,k},\ \mathbf{f}_{t,k}], \quad \mathbf{d}_k(\mathbf{a}) = \mathbf{d}^0_k + G_k\mathbf{a}, \quad \mathbf{a} = (a_s, a_t)^\top. \tag{22}$$

Thus, matching the predictor reduces to a two-variable ridge problem:

$$\mathbf{a}^\star_k = \arg\min_{\mathbf{a}\in\mathbb{R}^2} \left\{ \|\mathbf{r}_k + G_k\mathbf{a}\|_2^2 + \lambda\|\mathbf{a}\|_2^2 \right\} = -(G_k^\top G_k + \lambda I_2)^{-1} G_k^\top \mathbf{r}_k, \qquad \lambda > 0. \tag{23}$$

The ridge term controls large coefficient changes when the two guidance residuals are nearly collinear. Let $\overline{\mathbf{a}}_k$ denote the coefficients after range constraints, inter-step change limits, and EMA smoothing. A bounded line search scales the resulting correction:

$$\mathbf{q}_k = G_k\overline{\mathbf{a}}_k, \quad \eta_k = \text{clip}_{[0,\eta_{\max}]}\left(-\frac{\langle \mathbf{r}_k, \mathbf{q}_k \rangle}{\|\mathbf{q}_k\|_2^2 + \epsilon_{\text{ls}}} \right), \quad \mathbf{d}^{\text{TP}}_k = \mathbf{d}^0_k + \eta_k\mathbf{q}_k. \tag{24}$$

Here $\epsilon_{\text{ls}} > 0$ stabilizes the denominator. The line search cannot increase this frozen-predictor residual (Appendix A.7). All predictions are reused from FlowEdit, so the correction adds no Wan DiT evaluations.

Standard POLARIS controls a DDIM cycle defect with coordinate-wise guidance; Transport-POLARIS controls a local transport defect with paired channel-wise coefficients. The latter is a flow-specific extension, with preservation gains demonstrated on Wan2.1 in Sec. 4.3. Its local descent property does not imply a global error guarantee for the UniPC rollout or successful completion of every semantic edit.

4 Experiments

Full protocols and results are in Appendix B. Reconstruction metrics use full images; on PIE-Bench, PSNR, MSE, SSIM, and LPIPS assess background preservation, and CLIP assesses target alignment.

4.1 Diffusion Inversion and Reconstruction

We compare POLARIS with DDIM at CFG 7.5 Mokady et al. (2022) and a fixed-$\omega = 0$ null-branch control using SD1.5 Rombach et al. (2022) on COCO2017 and Pick-a-Pic over 10–100

Table 1: Comprehensive comparison of runtime efficiency (s) and image quality metrics across various inference steps on COCO2017 and Pick-a-Pic datasets. The arrows (↓ and ↑) indicate whether lower or higher values are better, respectively.

Category / Dataset	Metric	Method	10	20	30	40	50	60	70	80	90	100
Efficiency	Runtime (s) (↓)	Fixed scale	1.6876	3.1628	4.6906	6.2423	7.7075	9.2745	10.5553	12.3849	13.7852	15.0787
		POLARIS	1.7113	3.2435	4.8286	6.3940	7.9618	9.5557	11.1370	12.6824	14.3093	15.9283
COCO2017	MSE (↓)	Fixed scale	1827	2285	2417	2614	2692	2429	2617	2492	2764	2866
		POLARIS	365	352	342	345	349	332	339	330	335	335
	PSNR (↑)	Fixed scale	15.940	14.940	14.680	14.320	14.190	14.670	14.330	14.550	14.080	13.920
		POLARIS	23.254	23.482	23.651	23.635	23.661	23.952	23.884	23.981	23.856	23.893
	SSIM (↑)	Fixed scale	0.545	0.508	0.496	0.485	0.481	0.487	0.479	0.482	0.472	0.469
		POLARIS	0.712	0.728	0.733	0.731	0.734	0.739	0.738	0.741	0.739	0.740
	LPIPS (↓)	Fixed scale	0.458	0.506	0.518	0.533	0.538	0.521	0.534	0.526	0.544	0.550
		POLARIS	0.165	0.150	0.143	0.144	0.142	0.135	0.137	0.133	0.136	0.135
Pick-a-Pic	MSE (↓)	Fixed scale	1875	2341	2482	2668	2755	2491	2673	2551	2810	2925
		POLARIS	391	378	368	371	375	358	365	356	361	362
	PSNR (↑)	Fixed scale	15.720	14.710	14.450	14.110	13.980	14.460	14.120	14.330	13.860	13.710
		POLARIS	22.742	22.965	23.138	23.121	23.145	23.435	23.371	23.468	23.342	23.381
	SSIM (↑)	Fixed scale	0.521	0.486	0.474	0.462	0.459	0.466	0.457	0.460	0.451	0.448
		POLARIS	0.691	0.707	0.712	0.710	0.713	0.718	0.717	0.720	0.718	0.719
	LPIPS (↓)	Fixed scale	0.475	0.521	0.534	0.549	0.554	0.537	0.551	0.542	0.561	0.568
		POLARIS	0.178	0.163	0.156	0.157	0.155	0.148	0.150	0.146	0.149	0.148



Figure 3: **Image Editing Visualizations.**

steps. POLARIS replays its controller (Sec. 3). We report 8-bit MSE, LPIPS Zhang et al. (2018), PSNR Fardo et al. (2016), and SSIM Wang et al. (2004).

POLARIS lowers MSE/LPIPS and raises PSNR/SSIM on both datasets (Table 1). Peak VRAM rises by 0.14 GB for SD1.5, 0.28 GB for SDXL, and 0.35 GB for SD3.5 (2.0–3.2%; Appendix B.5).

Null-branch control. At 50 steps on COCO2017, fixed $\omega = 0$ yields MSE 553.17, PSNR 20.69, SSIM 0.572, and LPIPS 0.156; POLARIS yields 349, 23.661, 0.734, and 0.142, respectively. On Pick-a-Pic, PSNR rises from 19.87 to 23.145 and LPIPS falls from 0.206 to 0.155. See Appendix B.3 for 10–100-step results; COCO2017 LPIPS ties at the reported precision at 100 steps. POLARIS requires distinct conditional and unconditional predictions.

Table 2: Quantitative comparison of image editing performance on PIE-Bench

Method	Model	DINO↓	PSNR↑	LPIPS↓	MSE↓	SSIM↑	CLIP↑
RF-Inv Rout et al. (2025)	FLUX	42.29	20.20	179.54	139.20	69.91	24.57
FireFlow Deng et al. (2024)	FLUX	29.03	23.33	133.40	70.83	81.22	**26.19**
FlowEdit Kulikov et al. (2025b)	FLUX	27.82	21.96	112.19	94.99	83.08	25.25
FlowEdit Kulikov et al. (2025b)	SD3	27.12	22.22	104.12	85.96	**93.22**	**26.53**
RF-Edit Wang et al. (2024b)	FLUX	21.79	24.83	113.15	52.46	83.38	25.57
RF-Edit Wang et al. (2024b)	SD3.5	19.46	25.32	101.04	46.86	84.91	25.62
+ POLARIS	SD3.5	17.92	25.88	92.35	42.10	85.68	25.78
DNAEdit Xie et al. (2025)	FLUX	18.87	24.99	95.06	50.45	85.71	25.79
DNAEdit Xie et al. (2025)	SD3.5	**14.19**	**26.66**	**74.57**	**32.76**	88.63	25.63
+ POLARIS	SD3.5	**13.76**	**26.94**	**69.48**	**31.15**	**89.02**	25.72
PnP Tumanyan et al. (2022)	SD1.5	27.35	22.31	112.76	82.95	79.25	25.41
MasaCtrl Cao et al. (2023)	SD1.4	27.12	22.19	105.44	86.37	79.91	24.03
DI Ju et al. (2023) +PnP	SD1.5	23.35	22.46	105.51	79.94	79.88	25.49
DI+MasaCtrl	SD1.4	23.58	22.68	87.41	80.63	81.51	24.39
InfEdit Xu et al. (2023)	LCM	19.31	27.31	56.32	47.80	85.30	24.90
InsP2P Brooks et al. (2022)	InsP2P	58.13	20.80	159.23	221.30	76.47	23.61
NTI Mokady et al. (2022) +P2P	SD1.4	13.44	27.03	60.67	35.86	84.11	24.75
NPI Miyake et al. (2024) +P2P Hertz et al. (2022)	SD1.4	16.17	26.21	69.01	39.73	83.40	24.61
SPDInv Li et al. (2024) +P2P	SD1.4	8.81	28.60	36.01	24.54	86.23	25.26
h-Edit Nguyen et al. (2025) +P2P	SD1.4	11.17	27.87	48.50	85.40	84.80	25.30
DI+P2P	SD1.4	11.65	27.22	54.55	32.86	84.76	25.02
+ POLARIS	SD1.4	11.32	27.52	49.80	29.95	85.12	25.20
SAGE Gomez-Trenado et al. (2025)	SD1.5	11.00	26.98	39.60	58.33	84.23	**25.76**
+ POLARIS	SD1.5	10.89	27.40	34.73	46.73	84.87	**25.76**
DCI+P2P Li et al. (2025b)	SD1.4	**6.07**	**29.38**	**33.01**	**21.28**	**87.14**	**25.52**
+ POLARIS	SD1.4	**5.25**	**29.89**	**28.58**	**18.42**	**87.53**	**25.52**

4.2 Image Editing

On PIE-Bench, we integrate POLARIS with DI+P2P, DCI+P2P, SAGE, RF-Edit, and DNAEdit. Following Direct Inversion Ju et al. (2023), POLARIS adapts the source-preservation trajectory while retaining the editor's target guidance.

POLARIS improves background preservation across the tested pipelines (Table 2). For DCI+P2P, PSNR rises from 29.38 to 29.89 and LPIPS falls from 33.01 to 28.58. PSNR, LPIPS, MSE, and SSIM also improve for DI+P2P and SD3.5 RF-Edit/DNAEdit. CLIP rises for DI+P2P, SAGE, RF-Edit, and DNAEdit, though edits can remain incomplete (Appendix B.11).

4.3 Wan2.1 Video Editing with Transport-POLARIS

Wan2.1-T2V-1.3B Team Wan et al. (2025) with FlowEdit Kulikov et al. (2025a) uses a coupled source–target transport field. We apply the paired channel-wise controller from Sec. 3.3, reusing the host editor's flow predictions to jointly adapt source and target guidance throughout video editing.

Matched evaluation. FiVE-100 is a deterministic round-robin subset of FiVE Li et al. (2025a) with 100 edits from 34 source videos across six categories. Paired runs share prompts, 41 uniformly sampled 832×480 frames, a 50-step UniPC Zhao et al. (2023) schedule with 34 active steps, and source/target CFG 5/10. Seeds, per-step probe noise, and source/mask/output frame indices are matched; no Wan DiT evaluations are added. Background metrics are macro-averaged over aligned frames and masks to assess source preservation and temporal stability in unedited regions.

Preservation and temporal consistency. Transport-POLARIS improves every reported background metric (Table 3), raising unedited-region PSNR by 4.540 dB and reducing MSE and temporal drift by 63.4% and 30.7%. Figure 4 shows preservation of the bear video's rock layout and the dog's

Table 3: Wan2.1/FlowEdit on FiVE-100 at matched model-evaluation budgets. Video pixel errors differ in scale from 8-bit image-reconstruction MSE. Values are rounded; relative changes may differ from displayed ratios.

Background metric	FlowEdit	Transport-POLARIS	Reported change
PSNR, unedited region ↑	24.037	**28.577**	+4.540 dB
LPIPS, unedited region ↓	0.124	**0.093**	−25.239%
MSE, unedited region ↓	0.005	**0.002**	−63.422%
SSIM, unedited region ↑	0.767	**0.833**	+0.066
Valid-background PSNR ↑	23.535	**28.075**	+4.540 dB
Valid-background MSE ↓	0.005	**0.002**	−63.420%
Valid-background MAE ↓	0.044	**0.028**	−36.217%
Background temporal drift (L1) ↓	0.048	**0.033**	−30.677%

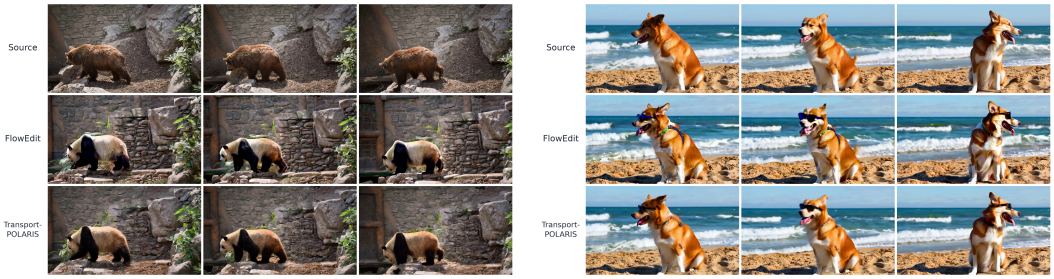

Figure 4: Wan video editing: bear (left) and dog (right). Rows: source, fixed CFG, and Transport-POLARIS. Transport-POLARIS better preserves the background across frames.

pose and beach composition. These metrics assess preservation, not semantic completion; target alignment and edit completeness require separate evaluation (Appendix B.4).

4.4 Additional Audio and Video Editing Results

We apply POLARIS to the preservation trajectory in AudioEditor and OpenSora/RF-Edit.

Audio editing setup. Following AudioEditor Jia et al. (2024), we use 300 manually selected AudioCaps Kim et al. (2019) test samples for addition, deletion, and replacement, comparing against SDEdit Meng et al. (2022) and TTA target-prompt regeneration. Frechet Distance (FD), Frechet Audio Distance (FAD), and Kullback–Leibler (KL) divergence assess preservation relative to the original wavs, with lower values indicating greater similarity to the source-audio distribution.

Video editing setup. We use OpenSora Zheng et al. (2024) with RF-Edit Wang et al. (2024b) and COVE-style data preparation. Baselines include FateZero Qi et al. (2023), Flatten Cong et al. (2023), COVE Wang et al. (2024a), RAVE Kara et al. (2024), and Token-Flow Geyer et al. (2023). We report four VBench Huang et al. (2024) metrics of video quality (Table 5).

Table 4: **AudioCaps source-preservation results.**

Editing Similarity with Original wavs			
Models	**FD↓**	**FAD↓**	**KL↓**
Regenerated wavs	55.48	6.45	2.61
SDEdit	44.13	5.68	**1.79**
AudioEditor	43.48	4.95	1.93
+ POLARIS	**42.87**	**4.65**	1.81

Audio editing results. POLARIS improves all three metrics over AudioEditor (Table 4), achieving the lowest FD and FAD among the compared methods. SDEdit remains slightly better on KL (1.79 versus 1.81).

Video editing results. POLARIS modestly improves all four VBench metrics over RF-Edit, without uncertainty estimates (Table 5). Figure 5 shows better background preservation across frames.

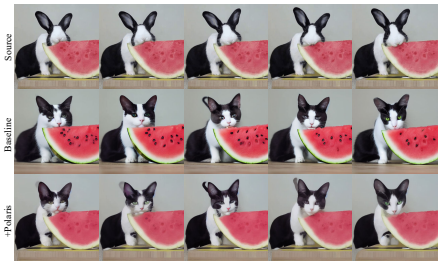

Figure 5: **Qualitative comparison of video editing.**

	FateZero	Flatten	COVE	RAVE	TokenFlow	RF-Edit	**+POLARIS**
SC (↑)	0.9382	0.9420	0.9433	0.9292	0.9439	0.9501	**0.9542**
MS (↑)	0.9611	0.9528	0.9697	0.9519	0.9632	0.9712	**0.9755**
AQ (↑)	0.6092	0.6329	0.6717	0.6586	0.6742	0.6796	**0.6843**
IQ (↑)	0.6898	0.7024	0.7163	0.6917	0.7128	0.7207	**0.7243**

Subject Consistency (SC), Motion Smoothness (MS), Aesthetic Quality (AQ), and Imaging Quality (IQ); higher is better.

Table 5: **Quantitative results of video editing.** VBench metrics under the RF-Edit protocol.

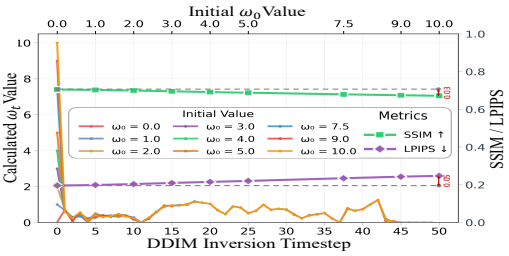

Figure 6: Initialization sensitivity for $\omega_0 = s_0\mathbf{1}$, $s_0 \in [0, 10]$. Curves show average guidance $\bar{\omega}_t$.

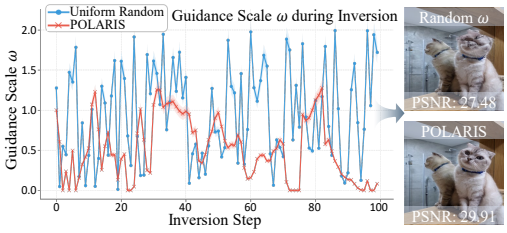

Figure 7: POLARIS versus random scalar guidance $\omega_t^{\text{rand}} = r_t\mathbf{1}$, $r_t \sim U(0, 2)$. Curves show coordinate averages.

These metrics assess preservation and perceptual quality, but do not establish whether the semantic edits are realized. FLUX.1-dev requires separate treatment, as discussed in Appendix A.6.

4.5 Ablation Study

Scale initialization. We run 50-step DDIM reconstruction on 100 random COCO2017 validation images, initializing $\omega_t \in \mathbb{R}^{C\times H\times W}$ with $\omega_0 = s_0\mathbf{1}$, $s_0 \in [0, 10]$, and updating via Eq. equation 19. We plot the coordinate average $\bar{\omega}_t = \frac{1}{d}\sum_{i=1}^{d}\omega_{t,i}, \quad d = CHW$. SSIM and LPIPS show limited sensitivity to the tested initializations (Figure 6). The coordinate average summarizes the tensor-valued controller; coordinates can receive different guidance.

Random-guidance control. Over 50 steps, random guidance $\omega_t^{\text{rand}} = r_t\mathbf{1}$, $r_t \sim U(0, 2)$, mimics fluctuations around 1.0 without solving Eq. equation 18 and yields lower reconstruction quality (Figure 7). Thus, arbitrary scalar fluctuations do not reproduce the gains of the residual-driven controller. Guidance granularity is tested separately below.

5 Conclusion

We introduced POLARIS, a training-free guidance plugin that specializes local inversion-consistency control to the CFG coefficients. Its coordinate-wise least-squares update reuses existing network predictions without gradient-based optimization or an auxiliary model. The analysis establishes a local advantage over restricted guidance parameterizations under frozen predictions, while measured cycle defects, structured-guidance ablations, and null-branch controls support its practical reconstruction benefits. Paired image, audio, and video evaluations show improved source preservation, and a separate transport-based extension demonstrates gains on Wan2.1/FlowEdit without extra model evaluations. These findings support adaptive guidance as a complementary component of existing editing systems. Its limitations include nonzero guidance-storage cost, dependence on accessible prediction branches, and the distinction between source fidelity and successful target editing. Large semantic changes, incorrect edit localization, and under-editing remain challenges for the underlying editor. Extending the controller to other guidance interfaces and evaluating target-semantic success alongside preservation are directions for future work.

REFERENCES

Omri Avrahami, Ohad Fried, and Dani Lischinski. Blended latent diffusion. *ACM Transactions on Graphics*, 42(4):1–11, July 2023. ISSN 1557-7368. doi: 10.1145/3592450. URL http://dx.doi.org/10.1145/3592450.

Yuxiang Bao, Huijie Liu, Xun Gao, Huan Fu, and Guoliang Kang. FreeInv: Free lunch for improving DDIM inversion. In *Advances in Neural Information Processing Systems*, volume 38, pages 32525–32550, 2025. URL https://papers.nips.cc/paper_files/paper/2025/hash/2eb74636e69ca26ab6cba8b724b0776e-Abstract-Conference.html.

Black Forest Labs. FLUX.1 [dev] model card. https://huggingface.co/black-forest-labs/FLUX.1-dev, 2024.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. *arXiv preprint arXiv:2211.09800*, 2022.

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)*, pages 22560–22570, October 2023.

Hyungjin Chung, Jeongsol Kim, Michael T. Mccann, Marc L. Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems, 2024. URL https://arxiv.org/abs/2209.14687.

Yuren Cong, Mengmeng Xu, Christian Simon, Shoufa Chen, Jiawei Ren, Yanping Xie, Juan-Manuel Perez-Rua, Bodo Rosenhahn, Tao Xiang, and Sen He. Flatten: optical flow-guided attention for consistent text-to-video editing. *arXiv preprint arXiv:2310.05922*, 2023.

Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance, 2022. URL https://arxiv.org/abs/2210.11427.

Yingying Deng, Xiangyu He, Changwang Mei, Peisong Wang, and Fan Tang. Fireflow: Fast inversion of rectified flow for image semantic editing, 2024. URL https://arxiv.org/abs/2412.07517.

Wenkai Dong, Song Xue, Xiaoyue Duan, and Shumin Han. Prompt tuning inversion for text-driven image editing using diffusion models, 2023. URL https://arxiv.org/abs/2305.04441.

Fernando A. Fardo, Victor H. Conforto, Francisco C. de Oliveira, and Paulo S. Rodrigues. A formal evaluation of psnr as quality measurement parameter for image segmentation algorithms, 2016. URL https://arxiv.org/abs/1605.07116.

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*, 2024. URL https://arxiv.org/abs/2403.14602.

Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. *arXiv preprint arXiv:2307.10373*, 2023.

Guillermo Gomez-Trenado, Pablo Mesejo, Oscar Cordón, and Stéphane Lathuilière. Don't forget your inverse ddim for image editing. *IEEE Computational Intelligence Magazine*, 20(3):10–18, August 2025. ISSN 1556-6048. doi: 10.1109/mci.2025.3563859. URL http://dx.doi.org/10.1109/MCI.2025.3563859.

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

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022. URL https://arxiv.org/abs/2207.12598.

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

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818, 2024.

Yuhang Jia, Yang Chen, Jinghua Zhao, Shiwan Zhao, Wenjia Zeng, Yong Chen, and Yong Qin. Audioeditor: A training-free diffusion-based audio editing framework, 2024. URL https://arxiv.org/abs/2409.12466.

Xuan Ju, Ailing Zeng, Yuxuan Bian, Shaoteng Liu, and Qiang Xu. Direct inversion: Boosting diffusion-based editing with 3 lines of code, 2023. URL https://arxiv.org/abs/2310.01506.

Ozgur Kara, Bariscan Kurtkaya, Hidir Yesiltepe, James M Rehg, and Pinar Yanardag. Rave: Randomized noise shuffling for fast and consistent video editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6507–6516, 2024.

Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. In *Advances in Neural Information Processing Systems*, 2022.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. Audiocaps: Generating captions for audios in the wild. In *NAACL-HLT*, 2019.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation, 2023. URL https://arxiv.org/abs/2305.01569.

Vladimir Kulikov, Matan Kleiner, Inbar Huberman-Spiegelglas, and Tomer Michaeli. FlowEdit: Inversion-free text-based editing using pre-trained flow models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19721–19730, 2025a. URL https://arxiv.org/abs/2412.08629.

Vladimir Kulikov, Matan Kleiner, Inbar Huberman-Spiegelglas, and Tomer Michaeli. Flowedit: Inversion-free text-based editing using pre-trained flow models, 2025b. URL https://arxiv.org/abs/2412.08629.

Minghan Li, Chenxi Xie, Yichen Wu, Lei Zhang, and Mengyu Wang. FiVE: A fine-grained video editing benchmark for evaluating emerging diffusion and rectified flow models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025a. URL https://arxiv.org/abs/2503.13684.

Ruibin Li, Ruihuang Li, Song Guo, and Lei Zhang. Source prompt disentangled inversion for boosting image editability with diffusion models, 2024. URL https://arxiv.org/abs/2403.11105.

Zixiang Li, Haoyu Wang, Wei Wang, Chuangchuang Tan, Yunchao Wei, and Yao Zhao. Dci: Dual-conditional inversion for boosting diffusion-based image editing, 2025b. URL https://arxiv.org/abs/2506.02560.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. URL https://arxiv.org/abs/1405.0312.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling, 2023. URL https://arxiv.org/abs/2210.02747.

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 *International Conference on Learning Representations*, 2022.

Daiki Miyake, Akihiro Iohara, Yu Saito, and Toshiyuki Tanaka. Negative-prompt inversion: Fast image inversion for editing with text-guided diffusion models, 2024. URL https://arxiv.org/abs/2305.16807.

Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. *arXiv preprint arXiv:2211.09794*, 2022.

Toan Nguyen, Kien Do, Duc Kieu, and Thin Nguyen. h-edit: Effective and flexible diffusion-based editing via doob's h-transform, 2025. URL https://arxiv.org/abs/2503.02187.

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

Chenyang Qi, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang, Ying Shan, and Qifeng Chen. Fatezero: Fusing attentions for zero-shot text-based video editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15932–15942, 2023.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. URL https://arxiv.org/abs/2112.10752.

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.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models, 2022. URL https://arxiv.org/abs/2010.02502.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations, 2021. URL https://arxiv.org/abs/2011.13456.

Team Wan et al. Wan: Open and advanced large-scale video generative models, 2025. URL https://arxiv.org/abs/2503.20314.

Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation, 2022. URL https://arxiv.org/abs/2211.12572.

Jiangshan Wang, Yue Ma, Jiayi Guo, Yicheng Xiao, Gao Huang, and Xiu Li. Cove: Unleashing the diffusion feature correspondence for consistent video editing. *Advances in Neural Information Processing Systems*, 37:96541–96565, 2024a.

Jiangshan Wang, Junfu Pu, Zhongang Qi, Jiayi Guo, Yue Ma, Nisha Huang, Yuxin Chen, Xiu Li, and Ying Shan. Taming rectified flow for inversion and editing. *arXiv preprint arXiv:2411.04746*, 2024b.

Yinhuai Wang, Jiwen Yu, and Jian Zhang. Zero-shot image restoration using denoising diffusion null-space model. *The Eleventh International Conference on Learning Representations*, 2023.

Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. doi: 10.1109/TIP.2003.819861.

Chenxi Xie, Minghan Li, Shuai Li, Yuhui Wu, Qiaosi Yi, and Lei Zhang. Dnaedit: Direct noise alignment for text-guided rectified flow editing, 2025. URL https://arxiv.org/abs/2506.01430.

Sihan Xu, Yidong Huang, Jiayi Pan, Ziqiao Ma, and Joyce Chai. Inversion-free image editing with natural language, 2023. URL https://arxiv.org/abs/2312.04965.

Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. Paint by example: Exemplar-based image editing with diffusion models. *arXiv preprint arXiv:2211.13227*, 2022.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric, 2018. URL https://arxiv.org/abs/1801.03924.

Wenliang Zhao, Lujia Bai, Yongming Rao, Jie Zhou, and Jiwen Lu. UniPC: A unified predictor-corrector framework for fast sampling of diffusion models. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://arxiv.org/abs/2302.04867.

Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all, 2024. URL https://arxiv.org/abs/2412.20404.

A Additional Analysis and Scope

A.1 Exact Local DDIM Cycle Identity

Let $\bar{\alpha}_t$ denote the cumulative signal coefficient at the selected noise level t, and assume $0 < \bar{\alpha}_{t+1} < \bar{\alpha}_t \leq 1$. Define the deterministic DDIM coefficients

$$\begin{aligned} a_{t+1} &= \sqrt{\frac{\bar{\alpha}_{t+1}}{\bar{\alpha}_t}}, \\ b_{t+1} &= \sqrt{1-\bar{\alpha}_{t+1}} - a_{t+1}\sqrt{1-\bar{\alpha}_t}, \\ \bar{\eta}_{t+1} &= -b_{t+1}/a_{t+1}. \end{aligned} \tag{25}$$

The affine maps, when supplied with a noise estimate $\mathbf{v}$, are

$$\begin{aligned} \mathfrak{I}_{t\to t+1}(\mathbf{x}, \mathbf{v}) &= a_{t+1}\mathbf{x} + b_{t+1}\mathbf{v}, \\ \mathfrak{D}_{t+1\to t}(\mathbf{y}, \mathbf{v}) &= a_{t+1}^{-1}\mathbf{y} + \bar{\eta}_{t+1}\mathbf{v}. \end{aligned} \tag{26}$$

Their composition at two supplied predictions $\mathbf{v}_t$ and $\mathbf{v}_{t+1}$ satisfies

$$\begin{aligned} &\mathfrak{D}_{t+1\to t}\big(\mathfrak{I}_{t\to t+1}(\mathbf{x}, \mathbf{v}_t), \mathbf{v}_{t+1}\big) - \mathbf{x} \\ &\quad = \frac{b_{t+1}}{a_{t+1}}\mathbf{v}_t + \bar{\eta}_{t+1}\mathbf{v}_{t+1} = \bar{\eta}_{t+1}(\mathbf{v}_{t+1} - \mathbf{v}_t). \end{aligned} \tag{27}$$

Using the same trial tensor ω to combine the fixed conditional and unconditional predictions at the two endpoints gives

$$\mathbf{v}_{t+1}(\omega) - \mathbf{v}_t(\omega) = (\mathbf{1} - \omega) \odot \Delta\varepsilon_{\varphi,t+1} + \omega \odot \Delta\varepsilon_{c,t+1} = \tau_{t+1}(\omega). \tag{28}$$

Consequently,

$$\|\mathbf{e}_{t+1}(\omega)\|_2^2 = \bar{\eta}_{t+1}^2 J_{t+1}(\omega). \tag{29}$$

For $\bar{\eta}_{t+1} \neq 0$, the two objectives have identical minimizers. For a zero-length transition, the cycle residual vanishes regardless of the coefficient and imposes no guidance preference.

What is held fixed. The algebraic cycle identity is exact for the supplied endpoint predictions. The least-squares solution treats those predictions as constants when varying ω. The full neural-network composition is different: a new guidance coefficient can change the latent, and therefore a freshly evaluated endpoint prediction. The closed form does not differentiate through or solve that nonlinear dependence. In addition, controllers can differ across neighboring implemented steps and near-zero coordinates use a numerical fallback. Thus, a vanishing unconstrained frozen-prediction residual does not imply an exact inverse for the complete rollout. The measured residuals in Appendix B.1 assess the implemented pipeline directly.

A.2 Accumulated Cycle Defects and Final Reconstruction

Let $\tilde{\mathbf{x}}_0, \ldots, \tilde{\mathbf{x}}_T$ be a fixed inversion trajectory. Let D_k denote the actual denoising transition from noise level k to $k-1$, including network evaluation with the recorded controller for that transition. Reconstruction follows $\widehat{\mathbf{x}}_{k-1} = D_k(\widehat{\mathbf{x}}_k)$ and starts at $\widehat{\mathbf{x}}_T = \tilde{\mathbf{x}}_T$. Define

$$\mathbf{r}_k = D_k(\tilde{\mathbf{x}}_k) - \tilde{\mathbf{x}}_{k-1}, \qquad r_k = \|\mathbf{r}_k\|_2, \qquad E_k = \|\widehat{\mathbf{x}}_k - \tilde{\mathbf{x}}_k\|_2. \tag{30}$$

Assume D_k is L_k-Lipschitz on a region containing the two relevant latent states. The triangle inequality gives

$$\begin{aligned} E_{k-1} &\leq \|D_k(\widehat{\mathbf{x}}_k) - D_k(\tilde{\mathbf{x}}_k)\|_2 + \|D_k(\tilde{\mathbf{x}}_k) - \tilde{\mathbf{x}}_{k-1}\|_2 \\ &\leq L_k E_k + r_k. \end{aligned} \tag{31}$$

Since $E_T = 0$, unrolling this recursion yields

$$E_0 \leq r_1 + L_1 r_2 + \cdots + (L_1 \cdots L_{T-1}) r_T = \sum_{k=1}^{T} \left(\prod_{j=1}^{k-1} L_j \right) r_k, \tag{32}$$

where an empty product equals one. This establishes Eq. equation 20.

For fixed transition constants, smaller actual local defects tighten this upper bound. However, the closed-form local objective uses frozen predictions, and changing a controller can also change the trajectory and its Lipschitz constants. The result therefore does not prove a universal ordering of final errors between two nonlinear rollouts. It concerns latent-space reconstruction; decoded-image metrics additionally depend on the decoder, and editing metrics depend on the target branch. These distinctions motivate reporting both local defects and downstream measurements.

A.3 Scalar, Channel-wise, Spatial, and Tensor Guidance

Suppress the timestep index and write

$$\mathbf{a} = \Delta\varepsilon_{\varphi,t+1}, \qquad \mathbf{d} = \Delta\varepsilon_{c,t+1} - \Delta\varepsilon_{\varphi,t+1} = -\delta_{t+1}. \tag{33}$$

Then the common objective is

$$J(\omega) = \|\mathbf{a} + \omega \odot \mathbf{d}\|_2^2. \tag{34}$$

For a group of coordinates G that shares one coefficient ω_G, differentiating $\|\mathbf{a}_G + \omega_G \mathbf{d}_G\|_2^2$ gives

$$\omega_G^\star = -\frac{\langle \mathbf{a}_G, \mathbf{d}_G \rangle}{\|\mathbf{d}_G\|_2^2}, \qquad \|\mathbf{d}_G\|_2 > 0. \tag{35}$$

The four variants use different groups:

$$\text{Scalar:} \quad s^\star = -\frac{\langle \mathbf{a}, \mathbf{d} \rangle}{\|\mathbf{d}\|_2^2}, \qquad \omega_{j,h,w} = s^\star, \tag{36}$$

$$\text{Channel:} \quad \omega_j^\star = -\frac{\langle \mathbf{a}_{j,:,:}, \mathbf{d}_{j,:,:} \rangle}{\|\mathbf{d}_{j,:,:}\|_2^2}, \qquad \omega_{j,h,w} = \omega_j^\star, \tag{37}$$

$$\text{Spatial:} \quad \omega_{h,w}^\star = -\frac{\langle \mathbf{a}_{:,h,w}, \mathbf{d}_{:,h,w} \rangle}{\|\mathbf{d}_{:,h,w}\|_2^2}, \qquad \omega_{j,h,w} = \omega_{h,w}^\star, \tag{38}$$

$$\text{Tensor:} \quad \omega_i^\star = -\frac{a_i}{d_i}, \qquad d_i \neq 0. \tag{39}$$

Channel-wise inner products run over spatial positions, whereas spatial inner products run over channels. The tensor solution is identical to Eq. equation 19; its minus sign comes from $\mathbf{d} = -\delta_{t+1}$. A zero group denominator makes that group objective independent of its coefficient. The previous value is reused in this case, and a numerical ϵ_{num} guard is used for near-degenerate divisions.

Let $\mathcal{W}_{\text{scalar}}$, $\mathcal{W}_{\text{channel}}$, $\mathcal{W}_{\text{spatial}}$, and $\mathcal{W}_{\text{tensor}}$ denote the corresponding unconstrained feasible sets. They obey

$$\begin{aligned} \mathcal{W}_{\text{scalar}} &\subseteq \mathcal{W}_{\text{channel}} \subseteq \mathcal{W}_{\text{tensor}}, \\ \mathcal{W}_{\text{scalar}} &\subseteq \mathcal{W}_{\text{spatial}} \subseteq \mathcal{W}_{\text{tensor}}. \end{aligned} \tag{40}$$

Minimizing the same fixed objective over these nested sets yields

$$\begin{aligned} J^\star_{\text{tensor}} &\leq J^\star_{\text{channel}} \leq J^\star_{\text{scalar}}, \\ J^\star_{\text{tensor}} &\leq J^\star_{\text{spatial}} \leq J^\star_{\text{scalar}}. \end{aligned} \tag{41}$$

Channel-wise and spatial sets are generally not nested, so their relative performance is empirical. These inequalities describe the exact optima for common fixed predictions. Numerical guards and changes to later network predictions can prevent the same ordering from holding for final image metrics.

A.4 Source Preservation and the Editing Displacement

The connection to editing can be made explicit using the reinterpretation of inversion-based editing in FlowEdit Kulikov et al. (2025a). Let Z_t^{src} and Z_t^{tar} be the source inversion and target reverse trajectories sharing their terminal noise. The associated direct path is

$$Z_t^{\text{inv}} = Z_0^{\text{src}} + Z_t^{\text{tar}} - Z_t^{\text{src}}, \qquad dZ_t^{\text{inv}} = \left[V^{\text{tar}}(Z_t^{\text{tar}}, t) - V^{\text{src}}(Z_t^{\text{src}}, t)\right] dt. \tag{42}$$

This identity shows that the source trajectory participates in the editing displacement. Perturbing it can therefore alter preservation behavior. It does not prove that more accurate inversion guarantees the target edit: the target field and the source–target coupling also matter. FlowEdit itself uses an inversion-free transport construction; the identity above describes its analysis of conventional inversion-based editing. Our Wan experiment uses the distinct transport extension in Appendix A.7.

A.5 Positioning Relative to Inversion Methods

Reconstruction consistency and trajectory mismatch are established concerns in diffusion inversion. Null-text Inversion Mokady et al. (2022) optimizes unconditional embeddings, Direct Inversion Ju et al. (2023) stores trajectory corrections, and ReNoise Garibi et al. (2024) uses iterative re-noising to refine inversion. FreeInv Bao et al. (2025) is particularly close in its goal of inexpensive trajectory-error reduction. It matches random latent transformations between corresponding inversion and reconstruction steps and interprets the resulting process as efficient trajectory ensembling. POLARIS instead optimizes the coefficients mixing existing conditional and unconditional predictions in a local least-squares problem. Its novelty is the resulting analytic CFG controller, not the general idea of controlling reconstruction residuals. These are different mechanisms for a shared problem. The paired editor experiments establish the incremental effect of POLARIS on the tested pipelines; they do not establish numerical superiority over FreeInv or ReNoise, for which no matched comparison is reported here.

A.6 Branch Availability and Guidance Distillation

The current DDIM derivation needs separately available unconditional and conditional predictions. Standard FLUX.1-dev uses guidance distillation Black Forest Labs (2024): its ordinary distilled-guidance path conditions one prediction on a guidance input instead of computing the two-branch CFG combination used in our derivation. That input is not directly interchangeable with ω in Eq. equation 14. A configuration that explicitly evaluates additional prompt branches would require its own derivation, computational accounting, and evaluation. We therefore do not claim direct validation on standard FLUX.1-dev or LTX. SD3.5-based editing and the Wan transport experiment provide evidence for the specific tested flow-model integrations, rather than unrestricted compatibility with all generative architectures.

A.7 Transport-POLARIS for FlowEdit

Frozen transport and coefficient parameterization. The FlowEdit direction is the target guided velocity minus the source guided velocity Kulikov et al. (2025a). All four conditional/unconditional predictions are held fixed during the local correction. For one latent channel, write each flow tensor as a vector in $\mathbb{R}^n$, where n counts its temporal and spatial coordinates. With the definitions in Eqs. equation 21–equation 22, changing the source and target scales by (a_s, a_t) gives

$$\begin{aligned}\mathbf{d}_k(\mathbf{a}) &= \mathbf{v}^u_{t,k} + (g^0_t + a_t)\mathbf{f}_{t,k} - \mathbf{v}^u_{s,k} - (g^0_s + a_s)\mathbf{f}_{s,k} \\ &= \mathbf{d}^0_k - a_s\mathbf{f}_{s,k} + a_t\mathbf{f}_{t,k} = \mathbf{d}^0_k + G_k\mathbf{a}.\end{aligned} \tag{43}$$

A one-step predictor $\widehat{\mathbf{d}}_k$ is constructed from the cached preceding transport direction and held fixed in the solve. The defect, expressed in velocity units, is therefore

$$\mathbf{e}^{\mathrm{tr}}_k(\mathbf{a}) = \mathbf{d}_k(\mathbf{a}) - \widehat{\mathbf{d}}_k = \mathbf{r}_k + G_k\mathbf{a}. \tag{44}$$

This is a predictor-consistency surrogate. Its interpretation differs from the exact frozen-prediction DDIM cycle identity: it measures mismatch to a cached transport estimate, whose own accuracy is not guaranteed.

Ridge solution. For $J^{\mathrm{tr}}_k(\mathbf{a}) = \|\mathbf{r}_k + G_k\mathbf{a}\|_2^2 + \lambda\|\mathbf{a}\|_2^2$, $\lambda > 0$, stationarity gives

$$\nabla J^{\mathrm{tr}}_k(\mathbf{a}) = 2G_k^\top(\mathbf{r}_k + G_k\mathbf{a}) + 2\lambda\mathbf{a} = 0. \tag{45}$$

The Hessian $2(G_k^\top G_k + \lambda I_2)$ is positive definite, yielding the unique solution in Eq. equation 23. In channel-wise inner products, the system is

$$\begin{bmatrix} \|\mathbf{f}_{s,k}\|_2^2 + \lambda & -\langle \mathbf{f}_{s,k}, \mathbf{f}_{t,k}\rangle \\ -\langle \mathbf{f}_{s,k}, \mathbf{f}_{t,k}\rangle & \|\mathbf{f}_{t,k}\|_2^2 + \lambda \end{bmatrix} \mathbf{a}^\star_k = \begin{bmatrix} \langle \mathbf{f}_{s,k}, \mathbf{r}_k\rangle \\ -\langle \mathbf{f}_{t,k}, \mathbf{r}_k\rangle \end{bmatrix}. \tag{46}$$

Only vector inner products and a 2×2 solve are required per channel. Since the zero adjustment is feasible for this unconstrained ridge problem,

$$\|\mathbf{r}_k + G_k\mathbf{a}^\star_k\|_2^2 + \lambda\|\mathbf{a}^\star_k\|_2^2 \le \|\mathbf{r}_k\|_2^2. \tag{47}$$

Thus the ridge solution cannot increase the local unregularized defect relative to the baseline direction, for fixed predictions and predictor.

Table 6: Measured local cycle-residual norms. Each row averages 100 COCO2017 samples at one transition. Norms are displayed to three decimal places; the reported reductions need not equal ratios recomputed from rounded entries.

Transition	Fixed	POLARIS	Reported reduction
0	3.490	**3.423**	1.9%
1	0.795	**0.773**	2.7%
10	0.146	**0.136**	6.7%
20	0.099	**0.072**	27.3%
30	0.077	**0.041**	46.1%
40	0.040	**0.024**	40.5%
48	0.021	**0.011**	47.5%
Mean over transitions	0.188	**0.169**	$\approx 10\%$
Median over transitions	0.086	**0.065**	24.4%

Stabilization and bounded line search. Coefficient-range constraints, a maximum change between sampling steps, and EMA smoothing produce $\overline{\mathbf{a}}_k$. These operations can alter the exact ridge minimizer. Set $\mathbf{q}_k = G_k\overline{\mathbf{a}}_k$ and consider the one-dimensional objective

$$\min_{0\le\eta\le\eta_{\max}} Q_k(\eta), \qquad Q_k(\eta) = \|\mathbf{r}_k + \eta\mathbf{q}_k\|_2^2 + \epsilon_{\mathrm{ls}}\eta^2, \quad \eta_{\max} \ge 0. \tag{48}$$

Differentiating gives $Q_k'(\eta) = 2\langle \mathbf{r}_k, \mathbf{q}_k\rangle + 2\eta(\|\mathbf{q}_k\|_2^2 + \epsilon_{\mathrm{ls}})$. Clipping its unconstrained minimizer to the interval gives Eq. equation 24. Because $\eta = 0$ is available,

$$\|\mathbf{d}_k^{\mathrm{TP}} - \widehat{\mathbf{d}}_k\|_2^2 = \|\mathbf{r}_k + \eta_k\mathbf{q}_k\|_2^2 \le Q_k(\eta_k) \le Q_k(0) = \|\mathbf{r}_k\|_2^2. \tag{49}$$

This descent property holds for any stabilized direction $\mathbf{q}_k$, including one that no longer equals the unconstrained ridge correction. If $\mathbf{q}_k = 0$, the stabilized formula returns $\eta_k = 0$. The line search controls the local defect; it does not by itself enforce all constraints on the final inter-step guidance sequence.

Connection to standard POLARIS and scope. Both variants fit guidance coefficients using existing model predictions. Standard POLARIS uses one coefficient per latent coordinate for an adjacent DDIM noise discrepancy. Transport-POLARIS uses two coefficients per channel for the coupled source–target direction, with ridge regularization and a bounded correction. Its final direction is supplied to the host sampling update. The analysis conditions on the current predictions and predictor; it does not establish a UniPC truncation-error bound or a semantic-editing guarantee. The Wan2.1 results in Sec. 4.3 provide empirical support for this extension at unchanged model-evaluation cost.

B More Experiment Details

B.1 Measured Local Cycle Residuals

We measure the absolute ℓ_2 norm of the implemented one-step cycle residual on 100 COCO2017 samples using SD1.5. For each of the 49 measured adjacent transitions, we average over the evaluated samples. Table 6 reports representative transitions. The mean across all transition averages decreases from 0.188 to 0.169 (approximately 10%), and the median decreases from 0.086 to 0.065 (24.4%). The reported reduction reaches 47.5% at the last transition. These measurements show that the benefit remains observable beyond the frozen-prediction calculation. They do not establish that every coordinate or every individual sample improves.

B.2 Guidance-Granularity Ablation

We compare fixed guidance with Scalar-, Channel-, Spatial-, and Tensor-POLARIS under the same 50-step DDIM inversion setting. Table 7 shows a substantial improvement from scalar adaptation, with further gains from finer parameterizations. Tensor-POLARIS achieves the best reconstruction metrics

Table 7: Guidance-granularity comparison under the same 50-step DDIM setting.

Method	MSE ↓	PSNR ↑	SSIM ↑	LPIPS ↓
Fixed	2692	14.190	0.481	0.538
Scalar-POLARIS	980	17.800	0.610	0.280
Channel-POLARIS	482	21.740	0.715	0.182
Spatial-POLARIS	560	20.650	0.685	0.190
Tensor-POLARIS	**349**	**23.661**	**0.734**	**0.142**

Table 8: Null-branch fixed-$\omega = 0$ control versus POLARIS with SD1.5. Values are reproduced at their reported precision; COCO2017 LPIPS ties at 100 steps. This control is distinct from the fixed-$\omega = 1$ experiment.

Dataset	Metric	Method	10	20	30	40	50	60	70	80	90	100
COCO2017	MSE	Fixed $\omega = 0$	596.82	586.91	576.48	564.32	553.17	541.84	532.67	524.83	517.96	511.42
COCO2017	MSE	POLARIS	365	352	342	345	349	332	339	330	335	335
COCO2017	PSNR	Fixed $\omega = 0$	20.18	20.31	20.42	20.56	20.69	20.82	20.94	21.06	21.15	21.23
COCO2017	PSNR	POLARIS	23.254	23.482	23.651	23.635	23.661	23.952	23.884	23.981	23.856	23.893
COCO2017	SSIM	Fixed $\omega = 0$	0.521	0.539	0.551	0.563	0.572	0.580	0.587	0.592	0.596	0.600
COCO2017	SSIM	POLARIS	0.712	0.728	0.733	0.731	0.734	0.739	0.738	0.741	0.739	0.740
COCO2017	LPIPS	Fixed $\omega = 0$	0.198	0.181	0.171	0.162	0.156	0.150	0.145	0.141	0.138	0.135
COCO2017	LPIPS	POLARIS	0.165	0.150	0.143	0.144	0.142	0.135	0.137	0.133	0.136	0.135
Pick-a-Pic	MSE	Fixed $\omega = 0$	648.21	634.72	620.86	608.17	596.42	583.31	575.62	567.34	559.91	552.88
Pick-a-Pic	MSE	POLARIS	391	378	368	371	375	358	365	356	361	362
Pick-a-Pic	PSNR	Fixed $\omega = 0$	19.28	19.43	19.58	19.72	19.87	20.01	20.13	20.24	20.34	20.45
Pick-a-Pic	PSNR	POLARIS	22.742	22.965	23.138	23.121	23.145	23.435	23.371	23.468	23.342	23.381
Pick-a-Pic	SSIM	Fixed $\omega = 0$	0.433	0.452	0.465	0.477	0.487	0.496	0.503	0.509	0.513	0.518
Pick-a-Pic	SSIM	POLARIS	0.691	0.707	0.712	0.710	0.713	0.718	0.717	0.720	0.718	0.719
Pick-a-Pic	LPIPS	Fixed $\omega = 0$	0.252	0.237	0.225	0.215	0.206	0.199	0.193	0.188	0.184	0.179
Pick-a-Pic	LPIPS	POLARIS	0.178	0.163	0.156	0.157	0.155	0.148	0.150	0.146	0.149	0.148

among the tested variants. Channel-POLARIS outperforms Spatial-POLARIS in this experiment, but their feasible sets are not nested and no universal ordering is implied. The result is consistent with the importance of heterogeneous residuals across both channels and spatial positions. The local feasible-set argument in Appendix A.3 and this final-reconstruction experiment provide different forms of evidence.

B.3 Null-Branch Reconstruction Control

To test whether the benefit depends solely on a high positive baseline CFG scale, we compare with fixed $\omega = 0$, which selects the unconditional (null-prompt) prediction. Table 8 reports both this control and POLARIS over 10–100 steps on COCO2017 and Pick-a-Pic. At 50 steps, COCO2017 MSE decreases from 553.17 to 349 and PSNR increases from 20.69 to 23.661 dB; Pick-a-Pic MSE decreases from 596.42 to 375 and PSNR increases from 19.87 to 23.145 dB. The fixed null-branch baseline is much stronger than the high-CFG stress test, but POLARIS retains an advantage in pixel and structural fidelity. At 100 steps on COCO2017, LPIPS is tied at the reported precision (0.135), while MSE, PSNR, and SSIM favor POLARIS.

This is a null-branch *baseline* control. It does not establish that the adaptive method can operate with no distinct conditioning information: if the two predictions coincide, the denominator in Eq. equation 19 vanishes and guidance cannot change their combination.

B.4 Wan2.1 Video Editing on FiVE-100

Matched evaluation protocol. We evaluate Wan2.1-T2V-1.3B Team Wan et al. (2025) with FlowEdit Kulikov et al. (2025a), using the Transport-POLARIS extension in Appendix A.7. The evaluation uses a deterministic round-robin subset of FiVE Li et al. (2025a), denoted FiVE-100: 100 editing cases from 34 source videos covering all six FiVE editing categories. FiVE-100 is the subset used here, not a claim of evaluating the full benchmark. Both methods use matched source and

Table 9: Measured peak GPU memory for paired runs within each backbone.

Model	Baseline	+ POLARIS	Increase
SD1.5	5.82 GB	5.96 GB	0.14 GB (2.4%)
SDXL	8.74 GB	9.02 GB	0.28 GB (3.2%)
SD3.5	17.86 GB	18.21 GB	0.35 GB (2.0%)

target prompts, 41 uniformly sampled frames at 832×480 resolution, a 50-step UniPC Zhao et al. (2023) schedule with 34 active steps, source CFG 5, and target CFG 10. The random seed, per-step probe-noise sequence, and source, mask, and output frame indices are matched. The number of Wan DiT evaluations is unchanged. Background metrics are macro-averaged over the aligned frames and masks.

Results and interpretation. The main-text results in Table 3 (Sec. 4.3) show improved unedited-region fidelity and lower background temporal drift. The PSNR increase is 4.540 dB, while reported LPIPS and temporal-drift reductions are approximately 25.2% and 30.7%. These results extend the preservation evidence to the tested Wan backbone. They do not by themselves establish better target-semantic success: background similarity and temporal preservation measure different properties from completion of the requested edit. No target-semantic accuracy, uncertainty intervals, or multi-seed variability are reported for this subset.

B.5 Computational and Memory Overhead

The DDIM guidance update uses element-wise operations on predictions already available to the CFG pipeline. For d latent coordinates, this costs $O(d)$ arithmetic per update without additional network evaluations. Replay stores T guidance tensors, requiring $O(Td)$ memory; the scalar, channel-wise, and spatial variants use fewer guidance coefficients. Transport-POLARIS instead solves a small channel-wise system using flow predictions already computed by FlowEdit. No-extra-evaluation accounting assumes that the host already evaluates the required branches; a fixed null-branch implementation may omit a branch and has a different cost baseline. These statements concern incremental controller cost; they do not remove any optimization or model evaluations already used by a host editor.

Table 9 reports measured peak GPU memory with and without POLARIS under otherwise identical settings within each backbone. The observed increase is 0.14–0.35 GB, or 2.0–3.2%. Retaining the guidance trajectory and temporary controller tensors accounts for additional storage. Thus, negligible added network computation does not mean zero memory cost.

B.6 Generalization to Other Architectures

Table 10: Quantitative results of image reconstruction quality on the COCO 2017 validation set across different models.

Model	**Method**	**MSE** $\downarrow$	**PSNR** $\uparrow$	**SSIM** $\uparrow$	**LPIPS** $\downarrow$
SD 3.5 Medium	Euler	7195.15	10.37	0.41	0.64
	POLARIS	**1012.90**	**18.94**	**0.64**	**0.33**
SD1.5 + LCM	DDIM	1405.97	16.65	0.52	0.62
	POLARIS	**686.30**	**19.77**	**0.63**	**0.50**

As shown in Table 10, POLARIS significantly outperforms standard baselines across all metrics for image reconstruction on the COCO 2017 validation set. This consistent improvement across both Stable Diffusion 3.5 Medium and SD1.5 with LCM LoRA demonstrates the strong generalizability and effectiveness of our method.

B.7 CONDITIONAL RECONSTRUCTION

As a control experiment, we conduct reconstruction tests under standard DDIM, where inversion and sampling guidance scales are fixed to $\omega \equiv 1$. Using the Stable Diffusion v1.5 backbone, we evaluate reconstruction performance from 10 to 100 inference steps at intervals of 10. The results show that this baseline consistently underperforms POLARIS across all metrics and step counts.

Table 11: Reconstruction metrics comparison on COCO2017 and Pick-a-Pic with reconstruction scale fixed to $\omega \equiv 1$.

Metric	Dataset	10	20	30	40	50	60	70	80	90	100
MSE (↓)	COCO2017	621.45	612.81	605.93	594.17	585.22	573.99	568.14	560.29	555.71	552.01
	Pick-a-Pic	672.81	661.90	650.15	638.49	627.31	615.82	610.93	604.18	598.22	594.19
PSNR (↑)	COCO2017	19.85	19.98	20.07	20.21	20.35	20.48	20.59	20.68	20.75	20.81
	Pick-a-Pic	18.99	19.11	19.23	19.35	19.47	19.59	19.66	19.74	19.82	19.88
SSIM (↑)	COCO2017	0.5011	0.5188	0.5290	0.5381	0.5453	0.5512	0.5549	0.5580	0.5601	0.5615
	Pick-a-Pic	0.4123	0.4297	0.4411	0.4503	0.4588	0.4654	0.4688	0.4719	0.4740	0.4752
LPIPS (↓)	COCO2017	0.2144	0.1981	0.1905	0.1842	0.1798	0.1743	0.1712	0.1689	0.1670	0.1659
	Pick-a-Pic	0.2691	0.2553	0.2478	0.2411	0.2350	0.2298	0.2265	0.2231	0.2209	0.2190

B.8 ABLATIONS ON SAMPLING SCHEDULERS

Table 12 compares four sampling-time guidance strategies. The recorded inversion-order sequence is $\omega_1, \ldots, \omega_T$; if the reconstruction loop index increases from 1 to T, matching the original noise levels requires ω_{T-t+1}. This reverse-order replay has the best reported metrics among the four strategies. The alternatives replay the array in forward order (ω_t), replace it with cosine decay from 1 to 0, or use $\omega \equiv 1$ during sampling. The experiment tests timestep alignment of the stored sequence in addition to adaptivity. The sampling-time fixed-$\omega = 1$ ablation should be distinguished from the separate experiment that fixes guidance to one during both inversion and reconstruction.

Table 12: Performance comparison of POLARIS and alternative sampling strategies across various inference steps on COCO2017.

Dataset	Metric	Schedulers	10	20	30	40	50	60	70	80	90	100
COCO2017	MSE (↓)	ω_{T-t+1}	506	498	490	493	498	467	477	464	478	476
		ω_t	688	651	635	639	642	610	615	608	619	617
		$\omega \equiv 1$	1254	1198	1153	1160	1171	1120	1135	1115	1142	1138
		COSINE↓	891	845	822	829	833	798	804	790	810	805
	PSNR (↑)	ω_{T-t+1}	22.04	22.24	22.36	22.36	22.34	22.62	22.54	22.66	22.53	22.55
		ω_t	20.75	20.99	21.11	21.08	21.05	21.28	21.23	21.30	21.21	21.22
		$\omega \equiv 1$	18.15	18.34	18.51	18.48	18.43	18.64	18.58	18.66	18.55	18.57
		COSINE↓	19.63	19.86	19.98	19.95	19.92	20.11	20.07	20.16	20.04	20.06
	SSIM (↑)	ω_{T-t+1}	0.6922	0.7043	0.7085	0.7087	0.7088	0.7135	0.7122	0.7143	0.7126	0.7133
		ω_t	0.6615	0.6730	0.6781	0.6779	0.6780	0.6831	0.6820	0.6845	0.6822	0.6830
		$\omega \equiv 1$	0.5890	0.5995	0.6044	0.6039	0.6041	0.6088	0.6075	0.6101	0.6077	0.6085
		COSINE↓	0.6251	0.6368	0.6415	0.6410	0.6412	0.6466	0.6452	0.6480	0.6455	0.6463
	LPIPS (↓)	ω_{T-t+1}	0.2231	0.2039	0.1957	0.1950	0.1955	0.1870	0.1897	0.1862	0.1892	0.1886
		ω_t	0.2685	0.2451	0.2358	0.2349	0.2352	0.2255	0.2281	0.2240	0.2275	0.2268
		$\omega \equiv 1$	0.4102	0.3855	0.3710	0.3701	0.3715	0.3601	0.3640	0.3588	0.3633	0.3621
		COSINE↓	0.3350	0.3106	0.2988	0.2975	0.2981	0.2870	0.2905	0.2855	0.2899	0.2888

B.9 SCALING POLARIS TO LARGER DIFFUSION MODELS

We further extend POLARIS to Stable Diffusion XL, demonstrating that our method is not restricted to SD v1.5 but also scales effectively to larger, higher-capacity diffusion models. As shown in Table 13, POLARIS achieves strong results on this larger-parameter backbone, further confirming its scalability.

B.10 DIAGNOSTIC RESTORATION EVALUATION

We further conduct a diagnostic restoration evaluation to examine whether the improved inversion trajectory of POLARIS benefits downstream inverse-problem solvers. This source-informed diag-

Table 13: Performance comparison of POLARIS and fixed-scale inversion across various inference steps on Pick-a-Pic and COCO2017 using SDXL.

Dataset	Metric	Method	10	20	30	40	50	60	70	80	90	100
COCO2017	MSE (↓)	Fixed scale	1558	1932	2092	2348	2436	2013	2316	2071	2516	2668
		POLARIS	479	435	401	385	366	292	310	258	274	283
	PSNR (↑)	Fixed scale	16.52	15.58	15.27	14.76	14.60	15.43	14.82	15.32	14.45	14.19
		POLARIS	21.97	22.54	23.03	23.34	23.61	24.64	24.49	25.18	25.03	24.85
	SSIM (↑)	Fixed scale	0.5950	0.5856	0.5744	0.5575	0.5500	0.5711	0.5527	0.5661	0.5405	0.5322
		POLARIS	0.7113	0.7478	0.7574	0.7651	0.7698	0.7854	0.7828	0.7967	0.7950	0.7933
	LPIPS (↓)	Fixed scale	0.4636	0.4914	0.5050	0.5205	0.5279	0.5004	0.5207	0.5053	0.5354	0.5467
		POLARIS	0.2832	0.2568	0.2489	0.2366	0.2293	0.2021	0.2054	0.1815	0.1840	0.1872
Pick-a-Pic	MSE (↓)	Fixed scale	1149	1499	1604	1842	1928	1643	1868	1742	2068	2214
		POLARIS	309	342	326	337	326	269	268	255	252	240
	PSNR (↑)	Fixed scale	18.07	16.84	16.60	16.01	15.81	16.53	15.96	16.26	15.49	15.21
		POLARIS	24.01	24.22	24.59	24.60	24.83	25.88	25.88	26.38	25.98	26.30
	SSIM (↑)	Fixed scale	0.6884	0.6735	0.6695	0.6530	0.6455	0.6628	0.6462	0.6538	0.6305	0.6222
		POLARIS	0.8014	0.8181	0.8297	0.8323	0.8359	0.8478	0.8523	0.8552	0.8579	0.8625
	LPIPS (↓)	Fixed scale	0.3306	0.3745	0.3778	0.4002	0.4080	0.3846	0.4032	0.3928	0.4247	0.4360
		POLARIS	0.1673	0.1745	0.1641	0.1636	0.1557	0.1353	0.1298	0.1241	0.1226	0.1150

Table 14: Quantitative comparison. (MSE↓, PSNR↑, SSIM↑). Our method, POLARIS, consistently improves performance across various tasks and baseline models. The "—" indicates the Colorization task is not supported by the original DDNM architecture.

Baseline	Method	Deblur.			SR			Color.			Inpaint.		
		MSE	PSNR	SSIM	MSE	PSNR	SSIM	MSE	PSNR	SSIM	MSE	PSNR	SSIM
DPS	Fixed	327.32	23.98	0.80	365.76	23.51	0.80	761.29	20.09	**0.72**	690.69	20.45	0.72
	POLARIS	**300.25**	**24.38**	**0.81**	**286.64**	**24.58**	**0.82**	**755.96**	**20.14**	0.71	**632.57**	**20.85**	**0.72**
DDRM	Fixed	284.43	24.57	0.82	329.97	24.00	0.81	703.60	20.61	0.73	566.30	21.39	0.74
	POLARIS	**247.31**	**25.18**	**0.83**	**246.24**	**25.33**	**0.83**	**639.17**	**21.00**	**0.74**	**489.51**	**22.06**	**0.75**
DDNM	Fixed	330.44	23.95	0.78	344.62	23.64	0.80	—	—	—	735.80	20.21	0.70
	POLARIS	**320.26**	**24.12**	**0.79**	**290.28**	**24.41**	**0.82**	—	—	—	**700.72**	**20.49**	**0.71**

nostic differs from the standard restoration setting used by DPS Chung et al. (2024), DDRM Kawar et al. (2022), or DDNM Wang et al. (2023). In their original formulation, these solvers typically start from randomly sampled Gaussian noise and generate the restored image under the constraint of a degraded observation y. Such stochastic initialization may introduce generative hallucinations and make pixel-level metrics less directly attributable to inversion quality.

To isolate the role of inversion fidelity, we instead initialize the restoration process from the inverted noise latent z_T obtained from the source image, rather than from random Gaussian noise. The degraded observation y is used only to provide the task-specific data-consistency constraint. Under this setting, the restoration solver acts as a diagnostic decoder of the information stored in z_T: if the inverted latent preserves more source information, the final restored image should better match the original image.

The original source is available to the inversion stage in this diagnostic, so the inverted latent carries information that would not be available from the degraded observation alone. POLARIS and the baseline use the same restoration solver and data-consistency operation, while their inversion and replay guidance differ. The comparison therefore measures the combined inversion–reconstruction configuration, rather than isolating the terminal latent as the only changed variable. It does not establish superiority in a standard restoration task with access only to the degraded observation.

All experiments are conducted with Stable Diffusion v1.5 at 512×512 resolution using 50 DDIM steps. For a fair comparison, the task-specific restoration solver and data-consistency operations are kept unchanged. POLARIS uses the dynamic tensor guidance sequence $\{\omega_t\}_{t=1}^{T}$ during inversion and reconstruction, whereas the baseline uses a fixed guidance scale $\omega_t \equiv 1$. As shown in Table 14, POLARIS consistently reduces MSE and improves PSNR and SSIM across different restoration

tasks and solvers. These results further validate that POLARIS produces a more informative and error-resilient inverted noise latent z_T, which can serve as a stronger initialization for downstream reconstruction-oriented solvers.

B.11 Failure Cases and Preservation–Editing Trade-offs

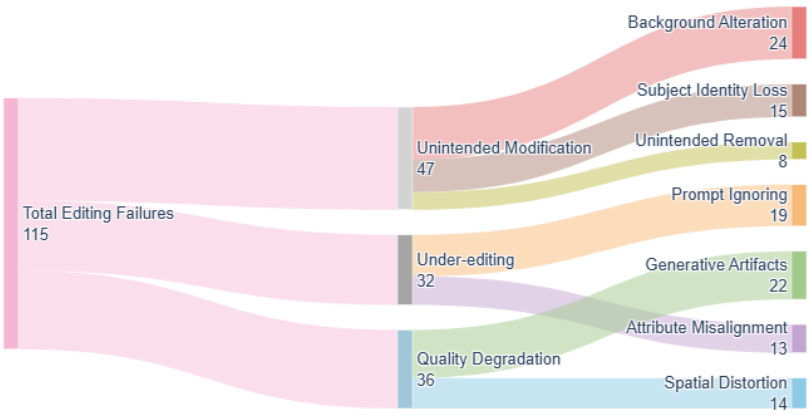

Figure 8: Detailed breakdown of failure cases in image editing. We categorize the 115 observed failures into three primary types: Under-editing, Unintended Modification, and Quality Degradation. The Sankey diagram illustrates the distribution of specific error sub-types, highlighting the most common failure modes addressed by POLARIS.

The 115 observed failures in Figure 8 are grouped into under-editing, unintended modification, and quality degradation. POLARIS primarily addresses source drift and unintended background changes. It cannot by itself repair incorrect edit localization, a mistaken semantic direction, or artifacts introduced by the target branch. If the editor leaks source-preservation constraints into the requested edit region, stronger preservation can also expose or worsen under-editing.

Simultaneous person and object replacement. In a case changing a woman examining a designer handbag into a man examining a sports bag, the store background and overall composition can remain faithful while feminine facial attributes or handbag-like structures persist. Correct background reconstruction does not ensure replacement of both semantic entities.

Large appearance and scene transformation. For an edit from a white horse crossing a lush green meadow to a black horse crossing a barren desert, preserving the horse's identity and motion can coexist with remnants of the original color or meadow. The requested transformation changes both the subject appearance and the scene, making the preservation constraint potentially conflict with the desired edit.

These cases explain why source fidelity and target fulfillment should be evaluated separately. They also limit the interpretation of background-only audio and video measurements: better preservation is useful, but is not a complete measure of editing success.

C Experimental Details

C.1 Baselines

Fixed-scale DDIM inversion. We use standard deterministic DDIM inversion with a fixed guidance scale as the main reconstruction baseline. This baseline uses the same pretrained diffusion backbone and scheduler as POLARIS, but keeps the guidance controller fixed across all timesteps and latent coordinates.

Prompt-to-Prompt (P2P). Hertz et al. (2022) Prompt-to-Prompt is a training-free text-guided image editing method based on cross-attention control. It injects attention maps from the source generation process into the target generation process, helping preserve the spatial layout of the source image while applying prompt-level semantic changes.

Direct Inversion (DI). Ju et al. (2023) Direct Inversion improves real-image editing by explicitly recording the discrepancy between inversion and reconstruction trajectories. During editing, these recorded error maps are injected back into the denoising process to improve source preservation.

DCI. Li et al. (2025b) DCI introduces a dual-conditional inversion strategy for diffusion-based image editing. It separates source-preserving and target-editing conditions, making it a strong baseline for evaluating whether POLARIS can further improve the preservation branch.

SAGE. Gomez-Trenado et al. (2025) SAGE uses self-attention guidance for text-guided image editing. It records self-attention maps during inversion and enforces consistency during denoising, improving structural preservation without requiring additional model training.

RF-Edit. Wang et al. (2024b) RF-Edit is a rectified-flow-based editing framework built upon RF-Solver. It uses feature sharing during inversion and editing to preserve the structural information of the source image or video. We combine POLARIS with RF-Edit to evaluate whether our inversion correction remains effective for rectified-flow-based editing.

DNAEdit. Xie et al. (2025) DNAEdit performs text-guided rectified-flow editing through direct noise alignment. We use it as another strong rectified-flow-based editing baseline to test whether POLARIS can provide orthogonal improvements beyond diffusion-based pipelines.

AudioEditor.Jia et al. (2024) AudioEditor is a training-free audio editing framework based on a pretrained text-to-audio diffusion model. It combines DDIM inversion, Null-text Optimization, and EOT-suppression to perform audio editing while preserving non-edited audio components.

SDEdit. Meng et al. (2022) For audio editing, following AudioEditor, we use SDEdit as a baseline. SDEdit performs editing by perturbing the source input with noise and denoising it under the target condition.

Video editing baselines. For video editing, we follow the RF-Edit evaluation protocol and compare with representative video editing methods, including FateZero Qi et al. (2023), Flatten Cong et al. (2023), COVE Wang et al. (2024a), RAVE Kara et al. (2024), and TokenFlow Geyer et al. (2023). These baselines cover attention fusion, optical-flow-guided attention, feature correspondence, randomized noise shuffling, and diffusion feature propagation.

FreeInv and ReNoise. FreeInv Bao et al. (2025) and ReNoise Garibi et al. (2024) are closely related inversion methods discussed in Appendix A.5. They are included for methodological positioning; no new quantitative comparison with them is claimed.

Wan/FlowEdit. FlowEdit Kulikov et al. (2025a) is an inversion-free flow editing method. We evaluate its Wan2.1 implementation with a transport-defect controller, rather than treating it as a DDIM inversion baseline.

Restoration solvers. For the diagnostic restoration evaluation, we use DPS Chung et al. (2024), DDRM Kawar et al. (2022), and DDNM Wang et al. (2023) as reconstruction-oriented inverse-problem solvers. In our setting, they are not initialized from random Gaussian noise; instead, they start from the inverted noise latent z_T to diagnose how much source information is retained by the inversion trajectory.

C.2 DATASETS

COCO 2017. Lin et al. (2015) For reconstruction evaluation, we use the COCO 2017 validation set. It contains diverse natural images with human-written captions, making it suitable for testing inversion fidelity under complex visual scenes.

Pick-a-Pic. Kirstain et al. (2023) We also evaluate reconstruction on Pick-a-Pic, a large-scale text-to-image preference dataset. Compared with COCO, Pick-a-Pic contains more diverse prompt-image pairs from generative models, allowing us to evaluate reconstruction robustness under a broader distribution.

PIE-Bench. Ju et al. (2023) For image editing, we use PIE-Bench, a comprehensive benchmark for evaluating text-guided image editing. It provides source images, source prompts, target prompts, and editing masks, enabling quantitative evaluation of both semantic editability and background preservation.

AudioCaps. Kim et al. (2019) For audio editing, following AudioEditor, we evaluate on 300 manually selected samples from the AudioCaps test set. These samples are grouped into addition, deletion, and replacement editing cases, with target captions manually crafted to guide the edits.

Video editing setup. For video editing, we follow the RF-Edit protocol, where the video data preparation mainly follows the COVE-style video editing setup. Since RF-Edit does not define a new named video dataset, we report results under the same protocol and evaluate with VBench metrics.

FiVE-100. For the Wan experiment, we use a deterministic 100-edit subset of FiVE Li et al. (2025a), containing 34 source videos and all six editing categories. Sampling, masking alignment, and inference settings are described in Appendix B.4.

C.3 METRICS

We evaluate reconstruction fidelity, source preservation, and semantic alignment separately. Reconstruction metrics compare the full reconstructed image with its source. On PIE-Bench, MSE, PSNR, SSIM, and LPIPS are evaluated on the masked non-edited/background region, while CLIP measures alignment with the target prompt. Masking conventions matter: computing an error after zeroing the edited region and normalizing over the whole frame is not equivalent to averaging only over valid background pixels. The Wan results therefore retain the separate unedited-region and valid-background labels used in the evaluation. Image-reconstruction MSE below is expressed in the 8-bit pixel scale. The Wan table retains the separate video evaluation scale and distinguishes its unedited-region and valid-background metrics; these values should not be compared numerically with 8-bit image MSE.

MSE. Mean Squared Error measures pixel-level reconstruction error. For the image-reconstruction tables, intensities are represented on the 8-bit scale $[0, 255]$ and errors are accumulated in floating point. Lower MSE indicates more accurate reconstruction.

$$\text{MSE} = \frac{1}{mn} \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} [I(i,j) - K(i,j)]^2. \tag{50}$$

PSNR. Fardo et al. (2016) Peak Signal-to-Noise Ratio measures reconstruction quality based on MSE. Higher PSNR indicates better reconstruction fidelity.

$$\text{PSNR} = 10 \cdot \log_{10} \left(\frac{\text{MAX}_I^2}{\text{MSE}} \right), \tag{51}$$

where $\text{MAX}_I = 255$ for 8-bit images.

SSIM. Wang et al. (2004) Structural Similarity Index Measure evaluates structural similarity between two images. It compares luminance, contrast, and structure, and a higher SSIM indicates better structural preservation.

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}. \tag{52}$$

LPIPS. Zhang et al. (2018) Learned Perceptual Image Patch Similarity measures perceptual distance between two images using deep features. Lower LPIPS indicates stronger perceptual similarity.

$$\mathrm{LPIPS}(x, x_0) = \sum_l \frac{w_l}{H_l W_l} \sum_{h,w} \left\| \hat{y}^l_{hw} - \hat{y}^l_{0,hw} \right\|_2^2 . \tag{53}$$

DINO. For image editing, DINO measures feature-level distance between the edited image and the source image, especially reflecting preservation of source structure and background regions. Lower DINO indicates better preservation.

CLIP. CLIP score measures semantic alignment between the edited image and the target prompt. Higher CLIP score indicates stronger consistency with the desired edit.

FD, FAD, and KL. For audio editing, we report Frechet Distance (FD), Frechet Audio Distance (FAD), and Kullback–Leibler divergence (KL) with respect to the original wavs. These metrics measure distributional similarity between edited audio and source audio. Lower values indicate greater distributional similarity to the source audio; they do not localize the preserved component or establish completion of the requested edit.

VBench metrics. Huang et al. (2024) For video editing, we follow VBench and report Subject Consistency (SC), Motion Smoothness (MS), Aesthetic Quality (AQ), and Imaging Quality (IQ). SC evaluates whether the subject remains consistent across frames, MS evaluates temporal motion smoothness, AQ evaluates visual appeal, and IQ evaluates frame-level imaging quality. Higher values indicate better video editing quality.

D Additional Visualizations

Fig. 9 presents additional qualitative results on text-guided image editing. These examples show that POLARIS better preserves background details and non-edited regions while allowing the original editing backbone to perform the desired semantic modifications. Figs. 10–13 provide additional qualitative results for diagnostic restoration, including Gaussian deblurring, $8\times$ super-resolution, central-mask inpainting, and image colorization. These examples illustrate source-informed diagnostic reconstruction and should not be interpreted as standard restoration results from degraded observations alone.

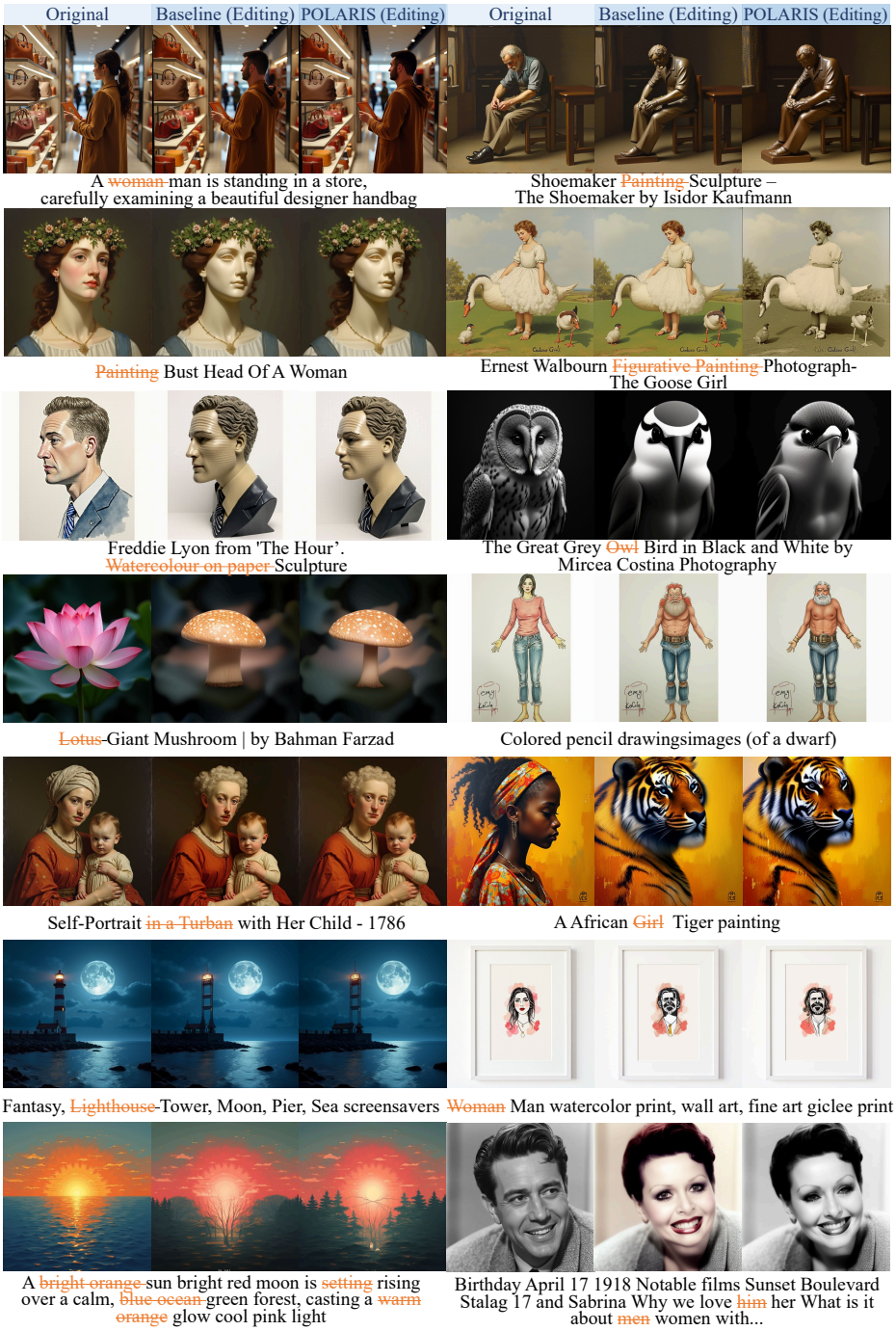

Figure 9: **Additional image editing visualizations.** POLARIS improves source preservation and reduces unintended background changes while maintaining the intended semantic edits.



Figure 10: **Gaussian deblurring visualizations.** Compared with the baseline, POLARIS better preserves source structures and reduces reconstruction artifacts under the diagnostic restoration setting.



Figure 11: **8× super-resolution visualizations.** POLARIS provides a more informative inverted noise latent, leading to sharper and more faithful reconstruction results.



Figure 12: **Central-mask inpainting visualizations.** POLARIS improves the consistency between restored regions and the original image, while better preserving non-masked content.

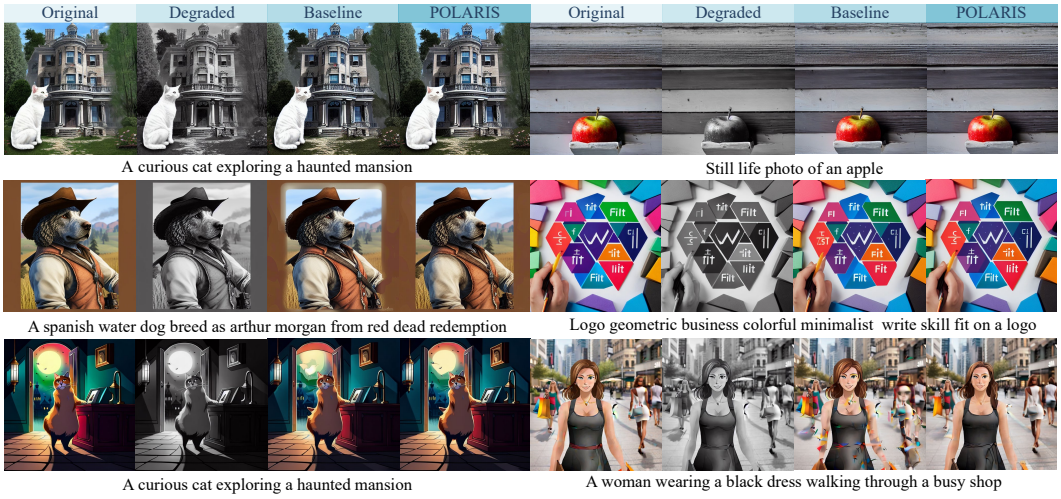

Figure 13: **Image colorization visualizations.** POLARIS better retains source structure and produces more faithful colorized reconstructions under the diagnostic restoration protocol.