RESTATE: NOISE-LEVEL-AWARE STATE CORRECTION FOR TRAINING-FREE OBJECT REMOVAL IN FEW-STEP DIFFUSION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Training-free object removal commonly redirects self-attention away from the target region, but on few-step distilled diffusion models this often leaves object-shaped residuals. We identify a temporal mismatch: a structured prediction forms inside the mask at the first forward, and the short trajectory offers few later chances to revise it, while attention rerouting acts mainly in later steps. Across four distilled backbones, magnitude-matched single-forward interventions show that background-directed state correction affects the final result more than attention rerouting. We therefore propose ReState, which corrects the masked clean prediction toward background evidence retrieved from the input using a noise-level-aware weight that vanishes at the final forward. ReState retains object-excluding attention as a complementary constraint and adds no denoiser evaluations. Across two paired benchmarks and two few-step backbones, ReState achieves the best masked PSNR, SSIM, and LPIPS among training-free methods, improving PSNR by 3.17 dB on average over the strongest baseline. On OBER-Test, it also outperforms the 50-step attention-forcing reference on all three metrics at $6.7\times$ lower latency. The same configuration transfers to two additional distilled backbones without retuning. Code will be released upon publication.

1 INTRODUCTION

Object removal aims to reconstruct the background occluded by a masked object. Training-free diffusion methods leverage pretrained generative priors without task-specific fine-tuning. Most existing training-free methods pursue removal by intervening in self-attention: they suppress attention to target-region keys so that the network draws information from the surrounding context (Jia et al., 2025; Sun et al., 2025; Liu, 2026; Vo et al., 2026). We refer to this family of interventions as *attention forcing* (Section 3.2). In their multi-step setting, these methods typically require tens of denoising forwards. Native few-step distilled models offer a substantially faster alternative, but attention forcing does not transfer cleanly to this regime: although it improves removal over no intervention, object contours and textures often remain inside the mask. The central question is therefore not only how to redirect attention, but why this intervention fails to change content that a short trajectory has already begun to commit.

To answer this question, we compare matched single-forward interventions. From the same latent at the same noise level, we either apply object-excluding attention for one forward or directly move the masked clean prediction toward retrieved background evidence by exactly the same ℓ_2 norm. Sampling then continues without further intervention. Across four distilled backbones, direct correction at the first forward produces a larger terminal removal effect than attention forcing. On the Turbo backbones, the one-step effect of attention forcing at the first forward is statistically not significantly different from zero, whereas the matched direct correction is significant. These results reveal a temporal mismatch: early in a few-step trajectory, changing the predicted state can affect the final image directly, while changing the attention route affects the trajectory only indirectly and may take effect too late.

So, we propose ReState, a training-free method that directly corrects the masked clean prediction toward background evidence retrieved from unmasked source regions. The correction is strongest

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.

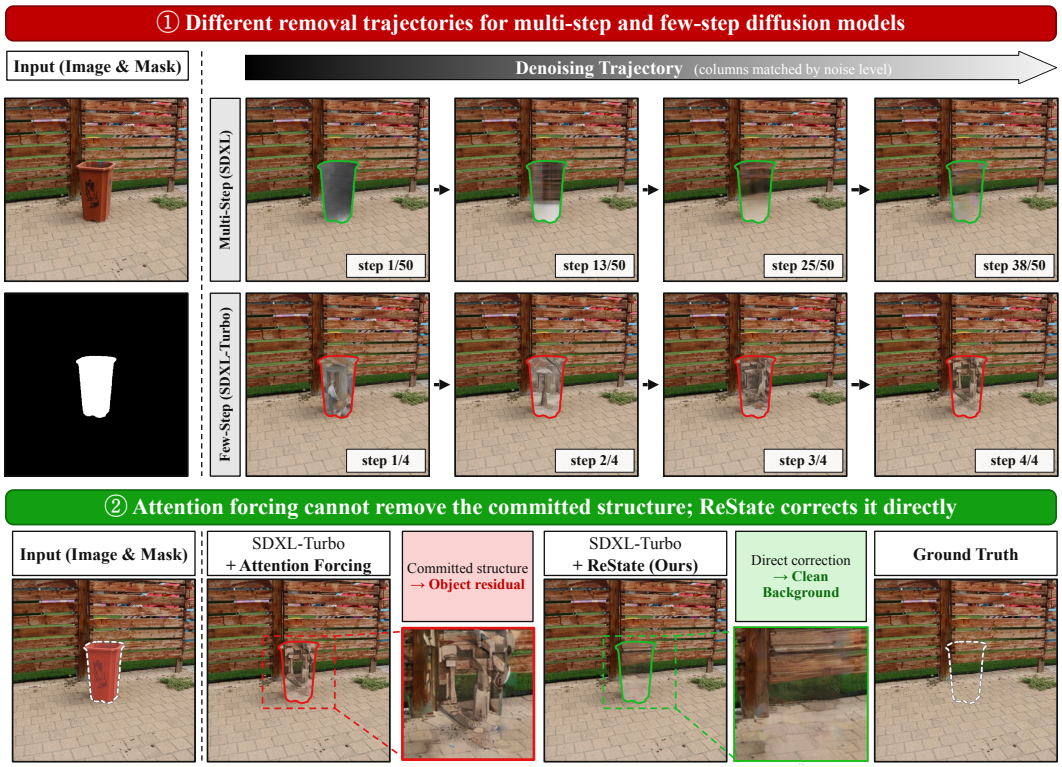

Figure 1: **Motivation.** With attention forcing applied, the multi-step model (SDXL) starts diffuse and forms the background gradually along the trajectory (top), while the few-step model (SDXL-Turbo) already fixes a structured image in the mask at the first forward, and the short trajectory does not absorb it. This committed structure is what attention forcing then fails to remove, leaving a residual (bottom); ReState instead corrects the prediction directly and removes it.

when noise remains and vanishes at the final forward, leaving the model free to synthesize final details. Object-excluding attention is applied throughout sampling as a complementary constraint; its additional benefit is concentrated near final synthesis. The retrieval features are shared with the first generation forward, so ReState preserves the native number of denoiser evaluations. Our contributions are:

- **Diagnosis.** We use matched single-forward interventions to show that background-directed direct correction has a larger first-forward terminal effect than attention forcing across four distilled backbones. On the Turbo family, the first-forward effect of attention forcing is not significantly different from zero.
- **Method.** We propose ReState, which corrects the masked clean prediction toward background evidence retrieved from the input image, with strength keyed to the remaining noise level, and keeps object-excluding attention as a route constraint. Among the compared training-free methods, ReState is the only one that writes an explicit background target into the predicted state, and it adds no denoiser evaluations beyond the native schedule.
- **Results.** Across two paired benchmarks and two main few-step backbones, ReState achieves the best masked PSNR, SSIM, and LPIPS among training-free methods, improving masked PSNR by 3.17 dB over the strongest baseline. On OBER-Test it surpasses every compared 50-step attention-forcing configuration in SSIM and LPIPS, runs 6.7× faster than the 50-step reference, and a single frozen configuration transfers to SDXL-Lightning and Hyper-SDXL without per-backbone tuning.

2 Related Work

Training-based and optimization-based removal. Training-based object removal methods learn removal-specific behavior through dedicated data or adaptation. PowerPaint learns task prompts for

Table 1: **ReState is the only training-free removal method that writes an explicit background target into the predicted state at the native few-step cost.** Cells classify each method as published (Appendix B.2); ; they state design properties, not measured performance, which Table 2 reports.

Property	DesignEdit AAAI'25	Attentive Eraser AAAI'25	AdaEraser ICML'26	PANDORA ICME'26	FuLLaMa WACV'26	**ReState** (Ours)
Writes an **explicit background target** into the predicted state	✗	✗	✗	✗	✗	✓
Adds **no forwards or external models** beyond the backbone	✓	✗	✗	✗	✗	✓
Operates at the **native few-step** schedule	✗	✗	✗	✗	✗	✓

removal (Zhuang et al., 2024), CLIPAway steers an inpainting latent away from the target identity (Ekin et al., 2024), and SmartEraser and EntityErasure use removal-specific supervision (Jiang et al., 2025; Zhu et al., 2025b). EraseLoRA adapts a low-rank adapter at test time (Jo et al., 2026), while EraDiff calibrates the diffusion trajectory for erasure (Liu et al., 2025). A separate line trains fast removal models that generate results in one to four forwards (Tang et al., 2026b; Zhou et al., 2026). These approaches can achieve high background fidelity, but they require removal supervision, test-time optimization, or dedicated training. ReState instead operates on a frozen pretrained model without removal-specific training.

Training-free attention-based removal. Training-free methods keep the generator frozen and modify self-attention so that image-token attention suppresses keys in the target region and draws information from surrounding context. DesignEdit (Jia et al., 2025), Attentive Eraser (Sun et al., 2025), AdaEraser (Liu, 2026), and PANDORA (Vo et al., 2026) instantiate this idea through different suppression rules, schedules, or guidance mechanisms; FuLLaMa additionally searches sampling parameters with derivative-free optimization (Demir & Çiftçi, 2026); attention routing also extends to dense-scene removal (Tang et al., 2026a) and, combined with local in-mask perturbations, to training-free geometric editing (Zhu et al., 2025a); none of these supplies an explicit background target for the masked prediction. We use the hard object-excluding attention of this family as a complementary constraint and do not claim it as a contribution. The distinction of ReState is that it does not rely on attention rerouting alone: it directly corrects the masked clean prediction when the trajectory is most sensitive to that correction.

Direct state manipulation in diffusion sampling. Diffusion methods for inverse problems and inpainting often modify the sampling trajectory or clean prediction to enforce measurement consistency, for example through known-region replacement (Lugmayr et al., 2022; Avrahami et al., 2023), null-space projection (Wang et al., 2023), or diffusion posterior sampling (Chung et al., 2023). ReState likewise modifies a clean prediction, but direct state manipulation itself is not our novelty. Rather, matched intervention measurements identify direct correction toward retrieved background evidence as the effective early intervention for few-step removal and determine its noise-level-dependent schedule.

Few-step distilled diffusion. Adversarial and consistency distillation enable image generation in one to four forwards (Sauer et al., 2024; Lin et al., 2024; Ren et al., 2024; Luo et al., 2023). Prior work has primarily studied generation quality and speed in this regime. We instead study whether an external removal intervention applied at a single forward survives to the terminal image. This terminal-effect analysis reveals the distinct temporal roles of attention rerouting and direct prediction correction, and motivates their combination in ReState.

3 METHOD

Figure 2 overviews ReState. The method retrieves background evidence from the first generation forward, corrects the masked clean prediction during sampling, and retains object-excluding attention as a complementary constraint throughout the native schedule.

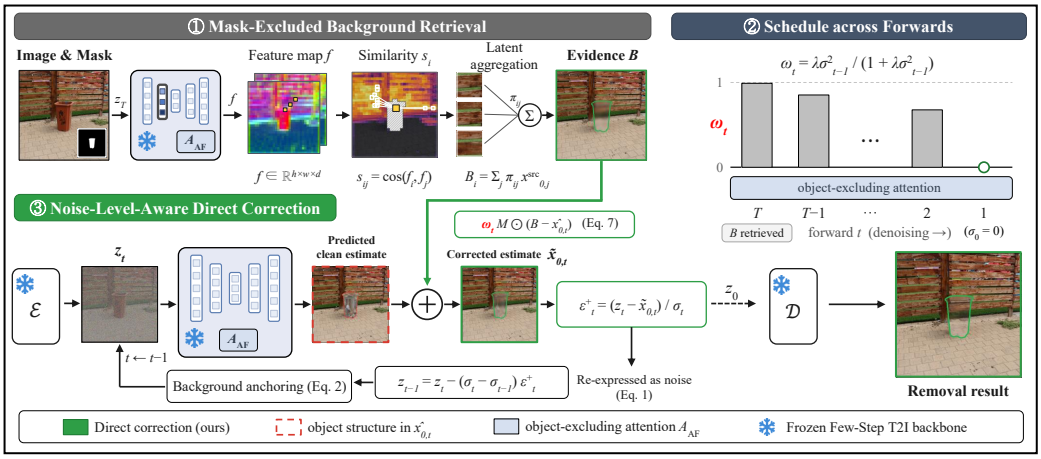

Figure 2: **Overview of ReState.** Features from the first generation forward retrieve background evidence B from unmasked source regions without an additional denoiser evaluation. At each step, ReState moves the masked clean prediction $\hat{x}_{0,t}$ toward B with a noise-level-aware weight ω_t. The weight vanishes at the final forward, while object-excluding attention remains active throughout sampling.

3.1 Problem Setup

Given a source image x^{src}, a binary removal mask M, a text condition c, and a frozen few-step model with native schedule $\sigma_T > \cdots > \sigma_0 = 0$, let Ω_m and Ω_b denote the masked and unmasked latent positions. We use M also for the corresponding latent-grid mask. Sampling starts from the noised source latent $z_T = \mathcal{E}(x^{\mathrm{src}}) + \sigma_T \epsilon$ (Rombach et al., 2022), with $\epsilon \sim \mathcal{N}(0, I)$. We use variance-exploding notation; the parameterization and scheduler conversions for all backbones are given in Appendix A.1.

For the ϵ-parameterized models used here, the denoiser induces the clean prediction

$$\hat{x}_{0,t} = z_t - \sigma_t \, \epsilon_\theta(z_t, \sigma_t, c; \mathcal{A}_t), \tag{1}$$

where $\mathcal{A}_t$ denotes the attention used at step t. At fixed z_t, correcting $\hat{x}_{0,t}$ is therefore equivalent to correcting the noise prediction consumed by the native sampler. After each transition, we anchor the unmasked region to the source at the matching noise level:

$$z_{t-1} \leftarrow M \odot z_{t-1} + (1 - M) \odot \left(\mathcal{E}(x^{\mathrm{src}}) + \sigma_{t-1}\epsilon\right). \tag{2}$$

Thus, all mechanisms below modify only the masked region.

3.2 Existing Object-Excluding Attention

Existing training-free removal methods modify self-attention with a mask-dependent bias on target-region keys. We refer to this family of interventions as *attention forcing*. Its hard and soft forms are

$$\Phi^{\mathrm{hard}}_{q,s} = \begin{cases} -\infty, & s \in \Omega_m, \\ 0, & \text{otherwise}, \end{cases} \qquad \Phi^{\mathrm{soft}}_{q,s} = \log \eta_t(s), \quad \eta_t(s) \in (0, 1], \tag{3}$$

where $\eta_t(s) = 1$ outside the mask and the hard form is the limit as $\eta_t(s) \to 0$. ReState uses the hard form at every forward.

Attention forcing prevents self-attention from reading target-region keys, but it does not identify which background content a masked prediction should approach. ReState therefore uses it as a route constraint and supplies the missing target through direct correction. Appendix A.2 specifies the attention layers, mask processing, and key-column implementation.

3.3 Mask-Excluded Background Retrieval

ReState retrieves an explicit background proposal from the source image. During the first generation forward, we capture a feature f_i at each latent position and reuse the same forward's denoiser

output for sampling. For each masked position i, we compare its feature with features at unmasked positions, retain the top fraction r, and aggregate their clean source latents:

$$\mathcal{N}_r(i) = \mathrm{TopK}_{j:\, M_j=0}(s_{ij}), \qquad s_{ij} = \frac{f_i^\top f_j}{\|f_i\|_2 \|f_j\|_2}, \qquad K = \lceil r N_{\mathrm{bg}} \rceil, \tag{4}$$

$$\pi_{ij} = \frac{\exp(s_{ij}/\tau)}{\sum_{k\in\mathcal{N}_r(i)} \exp(s_{ik}/\tau)}, \qquad B_i = \sum_{j\in\mathcal{N}_r(i)} \pi_{ij}\, x_{0,j}^{\mathrm{src}}, \qquad i \in \Omega_m. \tag{5}$$

We use $r = 0.01$. Because B is built from clean source latents outside the mask, it is a background proposal in the same space as $\hat{x}_{0,t}$ rather than a ground-truth reconstruction. The feature source, temperature, and implementation details are given in Appendix A.3.

3.4 Noise-Level-Aware Direct Correction

The correction is defined as the solution of a proximal objective. At step t, we seek a clean prediction that remains close to the model prediction while moving the mask interior toward B, with the fidelity term weighted by the noise remaining after the upcoming transition:

$$\tilde{x}_{0,t} = \arg\min_x \ \frac{1}{2\sigma_{t-1}^2}\|x - \hat{x}_{0,t}\|_2^2 + \frac{\lambda}{2}\|M \odot (x - B)\|_2^2. \tag{6}$$

The objective separates over latent positions. Outside the mask, its minimizer is the original prediction; inside the mask, setting the coordinate-wise gradient to zero gives

$$\tilde{x}_{0,t} = \hat{x}_{0,t} + \omega_t M \odot (B - \hat{x}_{0,t}), \qquad \omega_t = \frac{\lambda\sigma_{t-1}^2}{1 + \lambda\sigma_{t-1}^2}. \tag{7}$$

The update has three immediate operational properties. First, it leaves the mask exterior unchanged. Second, it contracts the masked distance to the evidence by the factor $1-\omega_t$. Third, $0 \le \omega_t < 1$, and the correction strength increases monotonically with the noise remaining after the transition. These properties characterize the update itself. Using σ_{t-1} rather than σ_t is essential. Early in the schedule, when the prediction can still be revised, the correction is strong. At the final transition, $\sigma_{t-1} = \sigma_0 = 0$ and hence $\omega_t = 0$: the final forward is not directly corrected toward the averaged evidence B, but instead synthesizes final details under object-excluding attention alone. ReState therefore uses a continuous noise-level-dependent weight rather than a hand-designed temporal gate. Appendix A.4 gives the complete derivation, and Appendix D.1 and Appendix D.2 verify the strength profile and final-forward boundary. Algorithm 1 summarizes the complete procedure. ReState retrieves B once from the first generation forward, applies the correction of Eq. 7 at every step, maps the corrected clean prediction back to a noise prediction, and advances the native sampler. Sharing the first-forward features adds no denoiser evaluation.

4 Experiments

Experimental setup. We evaluate on OBER-Test (Zhao et al., 2026) (163 GT-paired images) and RORD (Sagong et al., 2022) (343 GT-paired images), both of which provide a source image, removal mask, and object-free background. Main removal results use SDXL-Turbo and SD-Turbo (Sauer et al., 2024); SDXL at 50 steps (Podell et al., 2024) is the multi-step reference. We additionally evaluate the intervention analysis and full removal transfer on SDXL-Lightning (Lin et al., 2024) and Hyper-SDXL (Ren et al., 2024). All training-free methods use the same forward-only removal pipeline: an empty prompt, guidance scale 0, seed 0, no inversion, and source-latent anchoring outside the mask at every step. We compare DesignEdit (Jia et al., 2025), Attentive Eraser (Sun et al., 2025), and AdaEraser (Liu, 2026) under the fully shared four-step pipeline; their official multi-step procedures are evaluated separately at their native step count in Appendix E.3, and training- and optimization-based comparisons are reported in Appendix **??**. We report masked PSNR, SSIM, and LPIPS (Zhang et al., 2018) against the paired background, together with median end-to-end latency and effective denoiser evaluations. Metric definitions and preprocessing are in Appendix B.1. All metric differences are paired at the image level and tested with a two-sided sign-flip permutation

Algorithm 1 ReState sampling

Require: source x^{src}, mask M, condition c; frozen denoiser ϵ_θ with object-excluding attention $\mathcal{A}^{\mathrm{AF}}$; native schedule $\sigma_T > \cdots > \sigma_0 = 0$; autoencoder $(\mathcal{E}, \mathcal{D})$; parameters λ, r

Ensure: removal result $\hat{I}$

$z_T \leftarrow \mathcal{E}(x^{\mathrm{src}}) + \sigma_T \epsilon, \quad \epsilon \sim \mathcal{N}(0, I)$
$\hat{\epsilon}_T, \{f_i\} \leftarrow \epsilon_\theta(z_T, \sigma_T, c; \mathcal{A}^{\mathrm{AF}})$ ▷ shared first forward
$B \leftarrow \textsc{Retrieve}(\{f_i\}, M)$ ▷ Eq. 5; no additional denoiser evaluation
for $t = T, \ldots, 1$ **do**
  if $t < T$ **then** $\hat{\epsilon}_t \leftarrow \epsilon_\theta(z_t, \sigma_t, c; \mathcal{A}^{\mathrm{AF}})$
  $\hat{x}_{0,t} \leftarrow z_t - \sigma_t \hat{\epsilon}_t$ ▷ Eq. 1
  $\omega_t \leftarrow \lambda \sigma_{t-1}^2 / (1 + \lambda \sigma_{t-1}^2)$
  $\tilde{x}_{0,t} \leftarrow \hat{x}_{0,t} + \omega_t \, M \odot (B - \hat{x}_{0,t})$ ▷ Eq. 7
  $z_{t-1} \leftarrow z_t - (\sigma_t - \sigma_{t-1}) \, (z_t - \tilde{x}_{0,t}) / \sigma_t$ ▷ native Euler step on the corrected prediction
  $z_{t-1} \leftarrow M \odot z_{t-1} + (1 - M) \odot (\mathcal{E}(x^{\mathrm{src}}) + \sigma_{t-1} \epsilon)$ ▷ Eq. 2
end for
return $\hat{I} \leftarrow \mathcal{D}(z_0)$

Table 2: **Object removal on OBER-Test and RORD: on the same frozen backbone and the same four denoiser evaluations, ReState opens a large gap on every masked metric.** Rows with + apply the method on the backbone of their block; bold is best per column. Latency is the median per image and VRAM the peak allocation in GB, both on OBER-Test (Appendix B). Training-based comparisons are in Appendix E.1.

	OBER-Test			RORD				
Method	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	Lat. (s)↓	VRAM↓
SD-Turbo (4 steps)	11.27	0.203	0.380	9.86	0.162	0.438	0.174	4.58
+ Attentive Eraser AAAI'25	7.28	0.063	0.516	7.22	0.060	0.555	0.200	4.58
+ DesignEdit AAAI'25	14.74	0.316	0.283	13.48	0.263	0.355	0.184	4.58
+ AdaEraser ICML'26	12.27	0.237	0.347	10.48	0.177	0.420	0.843	5.08
+ **ReState (Ours)**	**18.19**	**0.485**	**0.186**	**16.59**	**0.440**	**0.280**	0.305	4.58
SDXL-Turbo (4 steps)	9.86	0.133	0.435	7.55	0.074	0.516	0.247	7.1
+ Attentive Eraser AAAI'25	7.09	0.046	0.514	7.35	0.065	0.538	0.338	7.1
+ DesignEdit AAAI'25	15.50	0.348	0.269	14.34	0.326	0.334	0.281	7.1
+ AdaEraser ICML'26	11.71	0.205	0.372	9.47	0.122	0.462	1.145	7.1
+ **ReState (Ours)**	**18.75**	**0.498**	**0.194**	**17.21**	**0.451**	**0.273**	0.373	7.1

test using 10,000 permutations; the four first-forward backbone comparisons additionally use Holm correction.

Main removal results. Table 2 reports the main comparison. ReState achieves the highest masked PSNR and SSIM and the lowest LPIPS in all four backbone–benchmark settings. Relative to DesignEdit, the strongest training-free baseline, it improves masked PSNR by 3.17 dB on average. Attentive Eraser falls below the no-intervention baseline at four steps because its SARG term amplifies the noise prediction at every forward. This mechanism is calibrated for 50-step sampling: under the shared 50-step reference configuration, Attentive Eraser reaches 18.79 dB on OBER-Test, versus 7.09 dB for its four-step port (Appendix E.3). With only four forwards, the same amplification is too large for a trajectory in which each step commits a substantial fraction of the final image. The identical ReState configuration ($\lambda = 8$, $r = 0.01$, probe layer, and schedule keying) also transfers to SDXL-Lightning and Hyper-SDXL without per-backbone tuning. On both benchmarks, it improves over DesignEdit by 1.46 to 2.39 dB in masked PSNR, with gains on every masked metric; all stated gaps are significant under the protocol of Appendix B.1. Full results appear in Appendix E.2.

Intervention asymmetry and timing. At a chosen forward t, we start the base and intervened runs from the same state z_t. The attention-forcing arm applies object-excluding attention at that forward only and induces $\Delta_t^{\mathrm{AF}} = M \odot (\hat{x}_{0,t}^{\mathrm{AF}} - \hat{x}_{0,t}^{\mathrm{base}})$. The competing arm moves the same clean prediction toward retrieved background evidence by the matched norm. With $u_t = M \odot (B - \hat{x}_{0,t}^{\mathrm{base}})$ and a

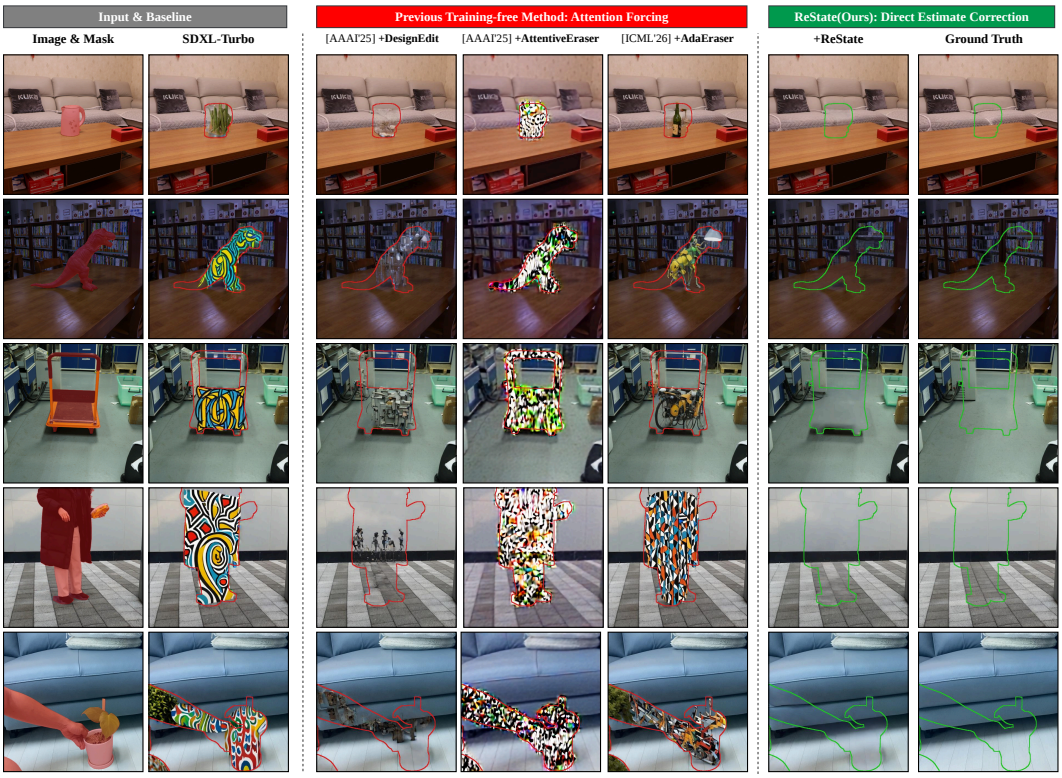

Figure 3: **Qualitative results on GT-paired benchmarks.** Previous training-free methods often leave object-shaped residuals in the mask, whereas ReState more closely reconstructs the paired background.

small $\varepsilon > 0$,

$$\Delta_t^{\text{direct}} = \frac{\|\Delta_t^{\text{AF}}\|_2}{\|u_t\|_2 + \varepsilon} u_t. \tag{8}$$

Exactly one perturbation is applied, after which sampling continues without further intervention. Letting $p(I, M)$ measure the source-to-GT directional residual of the output, the terminal removal effect is

$$E_k(t) = p\big(I_0^{\text{base}}, M\big) - p\big(I_0^{k@t}, M\big), \qquad k \in \{\text{AF}, \text{direct}\}, \tag{9}$$

where positive E means that the one-time intervention reduces this residual. Appendix C.1 defines p and gives the complete protocol. At the first SDXL-Turbo forward, correction toward retrieved background yields $E = +0.0251$ (permutation $p = 1.0 \times 10^{-4}$), whereas random and background-orthogonal directions yield -0.0004 (permutation $p = 0.53$) and -0.0006 (permutation $p = 0.36$), respectively. Thus, the comparison isolates the direction of state correction rather than perturbation magnitude. Table 3 shows that matched direct correction has a larger first-forward terminal effect than attention forcing on all four distilled backbones. The first-forward gaps remain significant after Holm correction (all corrected $p \leq 4 \times 10^{-4}$). On the Turbo family, attention forcing is not significantly from zero at the first forward ($p = 0.27$ for SDXL-Turbo and $p = 0.52$ for SD-Turbo), whereas direct correction is significant. On Lightning and Hyper-SDXL, attention forcing has a positive first-forward effect, but direct correction remains larger. For SDXL-Turbo, the terminal effect of a one-step attention intervention increases from $+0.002$ to $+0.033$ as noise decreases, whereas direct correction is effective from the first forward. A fixed-count control on the 50-step reference likewise favors late over high-noise attention forcing (Appendix C.3). These results motivate a correction that is strong while the state remains revisable and decays before final synthesis.

Table 3 reports E across noise levels and backbones; effect magnitudes are interpreted within each backbone. At the first forward, the magnitude-matched direct correction has a larger terminal removal effect than attention forcing on every tested backbone, and the permutation tests of the appendix place every gap outside sample variation. The contrast is strongest for the Turbo family,

Table 3: **Direct correction acts from the first forward.** Terminal effect E of a single matched intervention across noise levels and distilled backbones. Direct correction (E_{direct}, bold) exceeds attention forcing (E_{AF}) at every noise level and backbone.

	SDXL-Turbo		SD-Turbo		SDXL-Lightning		Hyper-SDXL	
σ	E_{AF}	E_{direct}	E_{AF}	E_{direct}	E_{AF}	E_{direct}	E_{AF}	E_{direct}
14.6	$+.002$	$\mathbf{+.025}$	$-.001$	$\mathbf{+.029}$	$+.034$	$\mathbf{+.072}$	$+.030$	$\mathbf{+.052}$
4.1	$+.016$	$\mathbf{+.039}$	$+.018$	$\mathbf{+.045}$	$+.052$	$\mathbf{+.059}$	$+.049$	$\mathbf{+.062}$
1.6	$+.028$	$\mathbf{+.040}$	$+.026$	$\mathbf{+.036}$	$+.029$	$\mathbf{+.041}$	$+.027$	$\mathbf{+.042}$
0.7	$+.033$	$\mathbf{+.049}$	$+.021$	$\mathbf{+.040}$	$+.023$	$\mathbf{+.046}$	$+.020$	$\mathbf{+.046}$

Table 4: **Component and timing ablation.** Direct correction supplies most of the gain; attention forcing contributes mainly at the last forward.

Direct correction	AF (all forwards)	AF (last only)	PSNR↑	SSIM↑	LPIPS↓
✗	✗	✗	9.86	0.133	0.435
✗	✗	✓	12.15	0.182	0.370
✗	✓	✗	15.50	0.348	0.269
✓	✗	✗	17.62	0.477	0.237
✓	✗	✓	18.65	0.495	0.196
✓	✓	✗	**18.75**	**0.498**	**0.194**

where the first-forward effect of attention forcing is not significantly different from zero while the same-norm direct correction is clearly significant; on Lightning and Hyper both effects are positive and the gap remains. Across the schedule, the terminal effect of attention forcing rises monotonically as noise decreases ($+0.002$ to $+0.033$ on SDXL-Turbo), while the direct correction is effective from the very first forward. A fixed-count control on the 50-step reference agrees: at the same count, late placements outperform high-noise placements (Appendix C.3). The effect of attention forcing thus arrives late, which motivates decaying the direct correction as the trajectory approaches final synthesis, exactly what the weight of Eq. 7 does.

Component and timing ablation. Table 4 separates the two components on SDXL-Turbo. Direct correction alone reaches 17.62 dB, compared with 15.50 dB for full-schedule attention forcing alone. Adding attention forcing only at the final forward raises the result to 18.65 dB, and applying it throughout reaches 18.75 dB. Conversely, last-forward attention forcing without direct correction reaches only 12.15 dB, below full-schedule attention forcing. Attention forcing therefore contributes on top of a state that direct correction has already prepared; its additional benefit is concentrated late, although the full schedule retains a small significant gain (Appendix D.2). The gain is also not explained by the retrieved proposal alone: at the same B and the same four forwards, the per-forward correction outperforms writing B into the latent once (Appendix D.4).

Efficiency and multi-step comparison. Table 5 reports hook-measured effective denoiser evaluations and end-to-end latency. ReState remains at the native four evaluations because the retrieval probe shares the first generation forward; feature matching and the masked update are its only additional operations. Against the 50-step attention-forcing reference (SDXL with DesignEdit), ReState improves all three masked metrics at $6.7\times$ lower latency. Appendix E.3 extends this comparison to every compared attention-forcing baseline at its native 50 steps: four-step ReState exceeds all of them in SSIM and LPIPS, and its PSNR is within 0.04 dB of the best.

5 Discussion

Why ReState combines a route constraint and a state correction. The combination follows from a measured temporal mismatch: a few-step trajectory commits to its masked prediction early, while the effect of attention forcing arrives only late. At equal perturbation norm, a background-directed correction of the masked clean prediction is effective from the first forward, whereas the first-forward effect of attention forcing is not significantly different from zero on the Turbo family and rises from $+0.002$ to $+0.033$ only as noise decreases; random and background-orthogonal

Table 5: **Cost and comparison against the multi-step reference.** Latency is median per image; forwards counts the effective model passes.

Method	Forwards	Latency (s)↓	PSNR↑	SSIM↑	LPIPS↓
SDXL-Turbo	4	0.247	9.86	0.133	0.435
+ Attentive Eraser AAAI'25	8	0.338	7.09	0.046	0.514
+ DesignEdit AAAI'25	4	0.281	15.50	0.348	0.269
+ AdaEraser ICML'26	8	1.145	11.71	0.205	0.372
+ ReState (Ours)	4	0.373	**18.75**	**0.498**	**0.194**
SDXL + DesignEdit AAAI'25	50	2.50	17.86	0.441	0.209
SDXL + ReState (Ours; App. E.3)	50	2.84	18.83	0.485	0.181

perturbations of the same norm have no significant terminal effect, attributing the difference to the direction and timing of the intervention rather than to its magnitude (Table 3). Commitment is equally measured: a first-forward perturbation survives to the output with slope 1.38 on the native four-step trajectory but only 0.30 on the 50-step reference (Appendix C.4, Appendix C.2). The direct correction therefore supplies the background direction, which attention forcing cannot specify, while generative freedom remains; attention forcing constrains late synthesis, and the timing ablation confirms this division: 1.03 of its 1.13 dB added benefit arrives at the final forward (Table 4, Appendix D.2). The removal state itself is therefore established by the direct correction before final synthesis; attention forcing completes rather than creates it.

The native few-step schedule is an operating point, not a degraded setting. Attention-forcing baselines that degrade under the four-step port operate correctly at their native step count (Attentive Eraser: $7.09 \rightarrow 18.79$ dB at 50 steps), so the few-step failures of Section 4 are consistent with a regime effect rather than a port artifact (Appendix E.3). Conversely, ReState transferred unchanged to 50 steps yields 17.57 dB, since the σ_{t-1}-keyed correction stays strong across the many low-noise forwards where fine texture is synthesized; a diagnostic window ($\sigma_{t-1} \geq 1.61$) recovers 18.83 dB, whereas on the native schedule the closed-form boundary $\omega \rightarrow 0$ already leaves the final low-noise forward uncorrected (Appendix E.3). The native four-step schedule is therefore the operating point of the method.

Scope and limitations. The asymmetry tracks the distilled denoiser and its short trajectory rather than the step count alone: original SDXL weights on the same four-step grid do not reproduce it, and densifying the Turbo schedule only shrinks the direct effect; we read these as diagnostic controls, not a causal factorization (Appendix C.3). Each claim is scoped to where it was measured: the zero first-forward effect of attention forcing to the Turbo family, the direct-correction advantage to the four distilled backbones tested, and the removal claims to two paired benchmarks and two main backbones, with full removal further verified on SDXL-Lightning and Hyper-SDXL (Appendix E.2). The clearest method limitation is the quality of the background evidence: an oracle background raises masked PSNR from 18.75 to 22.32 dB, a headroom of $+3.57$ dB (Appendix D.3), and part of this headroom is already recoverable by updating the correspondence at every forward (Appendix D.6), so improving background correspondence and adding uncertainty estimation appear more promising than further attention machinery. A second, smaller residual is textural: the fill remains under-textured at $0.51\times$ the high-band energy of the true background (Appendix C.5), and Appendix E.3 indicates that part of this reflects the limited free-synthesis budget of the coarse schedule rather than the correction itself.

6 Conclusion

Attention-based object removal on few-step distilled diffusion often leaves object-shaped residuals. Matched single-forward interventions show that background-directed clean-prediction correction has a larger terminal effect from the first forward, while attention forcing grows more effective at later noise levels. ReState therefore corrects the masked prediction toward retrieved background evidence with a noise-level-aware weight and retains object-excluding attention for late synthesis. Across two paired benchmarks and two few-step backbones, it achieves the best masked metrics among training-free methods without additional denoiser evaluations. In the settings we evaluate, few-step removal requires direct state correction while generative freedom remains, rather than attention rerouting alone.

AI USE STATEMENT

Large language models (LLMs) were used to edit author-written text for grammar, concision, readability, and academic tone, to assist with literature retrieval by suggesting search terms and candidate related work, and to assist with the implementation of evaluation and experiment scripts under author-specified protocols. The authors independently located and read the original sources, verified all citations and literature-based statements, and reviewed all AI-assisted code and text. Research ideation, method development, experimental design, and the interpretation of results were carried out by the authors, who take full responsibility for the final content of the paper.

REPRODUCIBILITY STATEMENT

ReState is training-free and requires no fine-tuning. The main text specifies the complete sampling procedure, the closed-form correction, and its operational properties; Appendix A documents per-backbone parameterizations, the attention convention, retrieval details, and the full derivation. Appendix B documents dataset construction, baseline reproduction settings, metric definitions, statistical tests, the hook-measured effective-forward accounting, and the runtime protocol and environment. Code will be released upon publication.

REFERENCES

Omri Avrahami, Ohad Fried, and Dani Lischinski. Blended latent diffusion. *ACM Transactions on Graphics*, 42(4), 2023.

Black Forest Labs. Flux.1 fill [dev]. `https://huggingface.co/black-forest-labs/FLUX.1-Fill-dev`, 2024.

Hyungjin Chung, Jeongsol Kim, Michael T Mccann, Marc L Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *International Conference on Learning Representations*, 2023.

İlke Demir and Umur Aybars Çiftçi. FuLLaMa: Training-free diffusion-based object removal with context preservation. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 8563–8573, 2026. doi: 10.1109/WACV61042.2026.00826.

Yiğit Ekin, Ahmet Burak Yildirim, Erdem Eren Caglar, Aykut Erdem, Erkut Erdem, and Aysegul Dundar. Clipaway: Harmonizing focused embeddings for removing objects via diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 17572–17601, 2024.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Yueru Jia, Aosong Cheng, Yuhui Yuan, Chuke Wang, Ji Li, Huizhu Jia, and Shanghang Zhang. Designedit: Unify spatial-aware image editing via training-free inpainting with a multi-layered latent diffusion framework. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 3958–3966, 2025.

Longtao Jiang, Zhendong Wang, Jianmin Bao, Wengang Zhou, Dongdong Chen, Lei Shi, Dong Chen, and Houqiang Li. Smarteraser: Remove anything from images using masked-region guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24452–24462, 2025.

Sanghyun Jo, Donghwan Lee, Eunji Jung, Seong Je Oh, and Kyungsu Kim. Eraselora: Mllm-driven foreground exclusion and background subtype aggregation for dataset-free object removal. In *European Conference on Computer Vision (ECCV)*, volume 17069 of *Lecture Notes in Computer Science*, pp. 376–393. Springer, 2026. doi: 10.1007/978-3-032-37023-5_21.

Shanchuan Lin, Anran Wang, and Xiao Yang. Sdxl-lightning: Progressive adversarial diffusion distillation. *arXiv preprint arXiv:2402.13929*, 2024.

Dingming Liu. Adaeraser: Training-free object removal via adaptive attention suppression. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 306 of *Proceedings of Machine Learning Research*, 2026.

Yi Liu, Hao Zhou, Benlei Cui, Wenxiang Shang, and Ran Lin. Erase diffusion: Empowering object removal through calibrating diffusion pathways. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11461–11471, 2022.

Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*, 2023.

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.

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 *International Conference on Learning Representations (ICLR)*, 2024.

Yuxi Ren, Xin Xia, Yanzuo Lu, Jiacheng Zhang, Jie Wu, Pan Xie, Xing Wang, and Xuefeng Xiao. Hyper-sd: Trajectory segmented consistency model for efficient image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 117340–117362, 2024.

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, 2022.

Min-Cheol Sagong, Yoon-Jae Yeo, Seung-Won Jung, and Sung-Jea Ko. RORD: A real-world object removal dataset. In *British Machine Vision Conference (BMVC)*, 2022.

Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *European Conference on Computer Vision (ECCV)*, pp. 87–103. Springer, 2024.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

Wenhao Sun, Xue-Mei Dong, Benlei Cui, and Jingqun Tang. Attentive eraser: Unleashing diffusion model's object removal potential via self-attention redirection guidance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 20734–20742, 2025.

Haitong Tang, Haipeng Liu, and Yang Wang. DORS: Dynamic attention routing for diffusion-based object removal in dense scenes. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, 2026a.

Yixin Tang, Jiawei Guo, Junxian Li, Zhiteng Li, Jixin Zhao, Bingya Zhang, Chenbo Wang, Yulun Zhang, and Shangchen Zhou. FlashClear: Ultra-fast image content removal via efficient step distillation and feature caching, 2026b.

Dinh-Khoi Vo, Van-Loc Nguyen, Tam V. Nguyen, Minh-Triet Tran, and Trung-Nghia Le. PANDORA: Pixel-wise attention dissolution and latent guidance for zero-shot object removal. In *IEEE International Conference on Multimedia and Expo (ICME)*, 2026.

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

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

Jixin Zhao, Zhouxia Wang, Peiqing Yang, and Shangchen Zhou. Precise object and effect removal with adaptive target-aware attention. In *CVPR*, 2026.

Qinming Zhou, Chenxi Sun, Deyang Kong, Junhao He, Xiangheng Tang, Peike Yu, Haotian Wu, Leilei Cao, and Linfeng Zhang. OSOR: One-step diffusion inpainting for effect-aware object removal. In *European Conference on Computer Vision (ECCV)*, 2026.

Hanshen Zhu, Zhen Zhu, Kaile Zhang, Yiming Gong, Yuliang Liu, and Xiang Bai. Training-free geometric image editing on diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19130–19140, 2025a.

Yixing Zhu, Qing Zhang, Yitong Wang, Yongwei Nie, and Wei-Shi Zheng. Entityerasure: Erasing entity cleanly via amodal entity segmentation and completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28274–28283, 2025b.

Junhao Zhuang, Yanhong Zeng, Wenran Liu, Chun Yuan, and Kai Chen. A task is worth one word: Learning with task prompts for high-quality versatile image inpainting. In *European Conference on Computer Vision (ECCV)*, pp. 195–211. Springer, 2024.

Appendix

Contents

A Method Details 14

A.1 Exact Diffusion Parameterization 14

A.2 Object-Excluding Attention Details 14

A.3 Background Retrieval Details 15

A.4 Derivation of the Closed-Form Correction 15

B Experimental Details 16

B.1 Metrics and Statistics 17

B.2 Auxiliary-Pass Accounting of Baselines 18

C Intervention Analysis 18

C.1 Complete First-Forward Analysis 18

C.2 Multi-Step Reference Intervention 19

C.3 Diagnostic Schedule Controls 19

C.4 Retention Analysis 21

C.5 Residual-Frequency Analysis 21

D Method Ablations 22

D.1 Correction Strength 22

D.2 Correction Timing and Schedule Structure 22

D.3 Background Evidence Source 23

D.4 Per-Forward Correction versus One-Shot Latent Initialization 23

D.5 Retrieval Ratio 24

D.6 Probe Feature Source 24

E Additional Results 26

E.1 Full Quantitative Tables 26

E.2 Full Removal on Additional Distilled Backbones 26

E.3 Multi-Step Baseline Comparison 27

E.4 Qualitative Results on OBER 29

E.5 Qualitative Results on RORD 30

A METHOD DETAILS

A.1 EXACT DIFFUSION PARAMETERIZATION

We state the clean-prediction and sampler relations of Section 3.1 in a backbone-independent form. A clean latent is corrupted as

$$z_t = x_0 + \sigma_t\,\epsilon \ \text{(VE)}, \qquad z_t = \sqrt{\bar\alpha_t}\,x_0 + \sqrt{1-\bar\alpha_t}\,\epsilon \ \text{(VP)}, \tag{A1}$$

and the VP form matches the VE form under $\sigma_t = \sqrt{(1-\bar\alpha_t)/\bar\alpha_t}$ after rescaling. The network output enters only through its parameterization, and the clean prediction is

$$\hat x_{0,t} = \begin{cases} z_t - \sigma_t\,\epsilon_\theta(z_t) & (\epsilon\text{-prediction}),\\ \sqrt{\bar\alpha_t}\,z_t - \sqrt{1-\bar\alpha_t}\,v_\theta(z_t) & (v\text{-prediction, VP coordinates}),\\ x_\theta(z_t) & (\text{sample-prediction}). \end{cases} \tag{A2}$$

In every case the correction of Eq. 7 is applied to $\hat x_{0,t}$ in this clean space, and the corrected prediction $\tilde x_{0,t}$ is mapped back to the native output type by inverting the same relation, e.g.

$$\epsilon_t^+ = \frac{z_t - \tilde x_{0,t}}{\sigma_t} \tag{A3}$$

for an ϵ-parameterization, so the scheduler transition to z_{t-1} runs unmodified. The method is therefore defined at the level of the clean prediction and is agnostic to the output parameterization and to the deterministic or stochastic form of the sampler.

Per-backbone instantiation. All four distilled backbones, SDXL-Turbo, SD-Turbo, SDXL-Lightning, and Hyper-SDXL,[1] are ϵ-parameterized and run the identical native four-step schedule of the Euler discrete scheduler with trailing timesteps $(999, 749, 499, 249)$, giving $\sigma = 14.617,\ 4.082,\ 1.613,\ 0.693$ with $\sigma_0 = 0$; these are the four noise levels of the intervention grid in Table 3. SDXL-Turbo and SD-Turbo operate at 512^2 and SDXL-Lightning and Hyper-SDXL at 1024^2. The scheduler exposes VE-form noise levels directly, so no explicit VP-to-VE conversion is performed at sampling time. At the final transition $\sigma_0 = 0$ no special-casing is required: the correction weight vanishes by the structure of the objective (Appendix A.4), and the sampler returns the clean prediction itself.

A.2 OBJECT-EXCLUDING ATTENTION DETAILS

With projected queries, keys, and values Q, K, V and key dimension d, an attention-forcing layer adds the bias Φ of Eq. 3 to the logits and reads

$$\tilde A = \mathrm{softmax}\Big(QK^\top/\sqrt{d} + \Phi\Big), \qquad h_q = \sum_s \tilde a_{q,s}\,v_s, \tag{A4}$$

where $\tilde a_{q,s}$ are the entries of $\tilde A$; the members of Section 3.2 differ only in Φ.

Attention convention. As implemented, the bias Φ acts on the key axis: the columns of object image keys, $s \in \Omega_m$ at the layer's token grid, receive $\phi_{\cdot,s} = \log \eta_t(s)$, which is $-\infty$ in the hard form, and background key columns receive zero. The same key-column bias is shared by all query rows, so background queries are likewise prevented from referencing object keys, and it is broadcast across all heads of a layer. Cross-attention layers are never modified.

The intervention covers every self-attention layer operating on the image-token grid; at 512^2 this is 70 layers of the SDXL UNet, 10 at the 32^2 grid and 60 at the 16^2 grid, and the 1024^2 backbones are covered at their correspondingly scaled grids. The latent-resolution mask is downsampled to each token grid by max-pooling, so a cell counts as object if any of its pixels is masked. ReState applies the hard form at all four forwards; the soft form appears only in the baselines of Section 3.2. Algorithm 2 summarizes the per-layer computation.

[1] Checkpoints: `stabilityai/sdxl-turbo`, `stabilityai/sd-turbo`, `ByteDance/SDXL-Lightning` (4-step UNet), and `ByteDance/Hyper-SD` (4-step SDXL LoRA).

Algorithm 2 Object-excluding self-attention at one layer

Require: projected Q, K, V on the image-token grid; token-grid mask m; suppression η_t
1: $b_s \leftarrow \log \eta_t(s)$ for keys s with $m_s = 1$, else $b_s \leftarrow 0$ ▷ $-\infty$ in the hard form
2: **return** $\mathrm{softmax}\Big(QK^\top/\sqrt{d} + \mathbf{1}\, b^\top\Big)\, V$ ▷ key-column bias, shared by all query rows and heads

A.3 Background Retrieval Details

Feature source. Features f_i are taken from the output of the first up-block of the UNet during the probe forward. The probe is the first generation forward itself: its features are captured by a hook and its noise prediction is reused at the first sampling step, so retrieval adds no auxiliary pass and the pipeline stays at four denoiser evaluations. The native resolution of this map at 512^2 is 32×32; it is bilinearly resampled to the 64×64 grid of the latent, so each latent position receives one feature vector for matching.

Matching. For each masked position $i \in \Omega_m$, cosine similarities between the ℓ_2-normalized feature f_i and all background features $\{f_j : j \in \Omega_b\}$ are computed. The top $r = 1\%$ of background positions are retained, $k = \lceil 0.01\,|\Omega_b| \rceil$, which is between 26 and 41 positions with a median of 38 on OBER-Test, and their weights are the softmax of the retained similarities at temperature $\tau = 0.1$.

Evidence space. The evidence is $B_i = \sum_k w_k\, \mathcal{E}(x^{\mathrm{src}})_{s_k}$, the weighted average of the clean source latents at the retrieved positions. B therefore lives in the same clean latent space as $\hat{x}_{0,t}$, and Eq. 7 mixes commensurable quantities.

A.4 Derivation of the Closed-Form Correction

Setting. Fix a step t with $\sigma_{t-1} > 0$ and drop the step index. The objective of Eq. 6 is

$$J(x) \;=\; \frac{1}{2\sigma_{t-1}^2}\,\|x - \hat{x}_0\|_2^2 \;+\; \frac{\lambda}{2}\,\|M \odot (x - B)\|_2^2. \tag{A5}$$

Proposition 1 (Closed-form correction)**.** *For $\sigma_{t-1} > 0$, J is strictly convex and has a unique minimizer x^*, given coordinate-wise by $x_i^* = \hat{x}_{0,i}$ for $i \notin \Omega_m$ and*

$$x_i^* \;=\; \hat{x}_{0,i} + \omega_t\,(B_i - \hat{x}_{0,i})\,, \qquad \omega_t \;=\; \frac{\lambda\sigma_{t-1}^2}{1 + \lambda\sigma_{t-1}^2}, \tag{A6}$$

for $i \in \Omega_m$, which is Eq. 7. Moreover:

- *(i)* Mask-exterior invariance: $(1 - M) \odot x^* = (1 - M) \odot \hat{x}_0$.
- *(ii)* Contraction: $x_i^* - B_i = (1 - \omega_t)\,(\hat{x}_{0,i} - B_i)$ *for $i \in \Omega_m$, with the fixed factor* $1 - \omega_t = 1/(1 + \lambda\sigma_{t-1}^2)$.
- *(iii)* Bounds and monotonicity: $0 \le \omega_t < 1$*; ω_t is increasing in σ_{t-1}, equals* 0 *at $\sigma_{t-1} = 0$, and approaches* 1 *only as $\sigma_{t-1} \to \infty$.*

Proof. J is a sum of two convex quadratics whose Hessian is diagonal with entries $1/\sigma_{t-1}^2 + \lambda M_i > 0$, so J is strictly convex and the unique minimizer is characterized by $\nabla J(x^*) = 0$. The objective separates over coordinates. For $i \notin \Omega_m$ the second term vanishes and $x_i^* = \hat{x}_{0,i}$, which gives (i). For $i \in \Omega_m$, stationarity reads

$$\frac{x_i^* - \hat{x}_{0,i}}{\sigma_{t-1}^2} + \lambda\,(x_i^* - B_i) = 0 \quad\Longrightarrow\quad x_i^* = \frac{\hat{x}_{0,i} + \lambda\sigma_{t-1}^2 B_i}{1 + \lambda\sigma_{t-1}^2}, \tag{A7}$$

which rearranges to Eq. A6. Subtracting B_i from both sides gives (ii), and (iii) follows from $\omega_t = 1 - 1/(1 + \lambda\sigma_{t-1}^2)$. □

Remark (Final transition). At the last step $\sigma_{t-1} = \sigma_0 = 0$ the fidelity weight $1/\sigma_{t-1}^2$ diverges and the minimizer degenerates to $x^* = \hat{x}_0$; equivalently $\omega_t \to 0$ in the closed form. The final forward

therefore receives no correction by the structure of the objective rather than by an added rule, and synthesis at that step is constrained only by the object-excluding attention of Eq. A4.

With the schedule of SDXL-Turbo and $\lambda = 8$ the realized weights are $\omega = 0.993,\ 0.954,\ 0.794,\ 0$ over the four forwards. Sensitivity of removal quality to λ, together with the realized ω profile at each value, is reported with the method ablations in Table D1. The necessity of the boundary itself is verified in Table D2: keying ω to σ_t leaves the final forward a correction and collapses perceptual quality.

B Experimental Details

Datasets. OBER-Test is the GT-paired removal benchmark released with ObjectClear (Zhao et al., 2026): each sample pairs a source image containing the target object with a ground-truth background image in which the object is absent, together with the removal mask. We use the released set as is, without filtering or re-annotation; the 163 images constitute the full benchmark.

RORD (Sagong et al., 2022) is released as 1920×1080 video sequences: each sequence contains frames with the removable objects present, a single object-free ground-truth frame of the same scene, and per-frame polygon annotations of those objects. From each of the 343 sequences we select one frame containing a single removal target, filtering candidates with the released annotations and confirming each choice by manual review, and pair it with the sequence's ground-truth frame; the removal mask is the binary union of the selected frame's object polygons.

Each triplet is then converted to 512^2 by a mask-centered square crop rather than a full-frame downsample: the crop side is the smallest square that contains the mask with a 10% margin per side, floored at 512 pixels, the identical window is applied to the image, the mask, and the ground truth, and the crop is area-resized to 512^2. Small objects are therefore evaluated at native resolution, and large objects incur only the minimal downsampling.

Baselines. All three training-free baselines run on the identical forward-only removal pipeline as ReState, so the comparison isolates the intervention: the prompt is the empty string, few-step sampling is guidance-free (scale 0.0), the seed is fixed to 0, and no inversion is used. The official DesignEdit aligns the source through DDIM inversion (Song et al., 2021), while AdaEraser constructs its reference from the noised source latent at the matching noise level; our few-step port re-injects the clean source latent outside the mask at every step, which anchors the background in latent space while only the masked region is synthesized.

DesignEdit is instantiated as the hard key-masking of Section 3.2 over all self-attention layers and forwards (Appendix A.2). Attentive Eraser retains its official hyperparameters on the few-step backbone; SARG replaces classifier-free guidance (Ho & Salimans, 2022), so its guidance is disabled by construction. AdaEraser runs its adaptive suppression with a reference forward alongside each target forward; the official implementation concatenates the two into one batched call, whereas our port runs them sequentially, which is mathematically identical since batch elements do not interact in the denoiser, and is counted as two forwards in Appendix B.2. Training-based comparisons use released checkpoints with their official default inference settings; the exact checkpoints, steps, and guidance scales are listed with the full tables in Appendix E.1.

Training data of the comparisons. Table B1 lists the training or adaptation data of every training-based removal method compared in Appendix E.1; the two base inpainting models (SDXL-Inpainting and FLUX.1-Fill) are omitted (Black Forest Labs, 2024)), as they carry web-scale pretraining rather than removal supervision. Overlap with a benchmark is judged at the dataset-family level regardless of split, from the training data stated in each paper or from a release that precedes the benchmark: OBER (2025) postdates every method except OSOR, and RORD (2022) precedes every method, so the RORD column rests on the stated training data alone.

Table B1: **Training data of the training-based removal methods and its overlap with the evaluation benchmarks.** ✓: stated to include the benchmark family; ✗: no stated overlap.

Method	Training or adaptation data	OBER	RORD
PowerPaint ECCV'24	OpenImages v6 (seg. subset) + LAION-Aesthetics v2 5+	✗	✗
CLIPAway NeurIPS'24	frozen base; embedding projection trained on COCO	✗	✗
EntityErasure CVPR'25	Open Images V7, about 1.75M (pseudo-labeled)	✗	✗
SmartEraser CVPR'25	Syn4Removal, 1M (Open Images v7, SAM, COCONut)	✗	✗
EraseLoRA ECCV'26	none (dataset-free test-time adaptation)	✗	✗
OSOR (SDXL-Inp.) ECCV'26	CORNE, 280K pairs (curated from NHR-Edit)	✗	✗
OSOR (FLUX.1-Fill) ECCV'26	CORNE + refinement on the captured subset of OBER	✓	✗

Runtime protocol. Latencies in Table 5 are measured on a single NVIDIA A100 80GB in fp16 at batch size 1 with seed 0. Each entry is the median per-image end-to-end wall-clock over OBER-Test with the first three images excluded as warmup: the timed call spans the full generation, including the VAE encode, the text-encoder pass, all denoising forwards, and the VAE decode, and excludes only one-time model loading. Forward counts are measured with a pre-forward hook on the denoiser backbone: the nominal count is the number of backbone calls, and the effective count sums the batch sizes of those calls, so batch duplication inside one call is counted (Appendix B.2). All experiments run in fp16 with PyTorch 2.7.1, diffusers 0.37.0.dev0, and CUDA 12.8; LPIPS uses the AlexNet backbone of Zhang et al. (2018) at its released weights, and every run fixes seed 0.

B.1 Metrics and Statistics

Metrics. All metrics are computed at each backbone's operating resolution (Appendix A.1): the result R, the ground truth G, and the mask M are resized to that resolution, with area interpolation for images and nearest-neighbor for the mask, which is then binarized. A 4-pixel border is excluded from all PSNR and SSIM computations. Masked PSNR averages the squared error over masked pixels and the three RGB channels on the 8-bit scale,

$$\mathrm{MSE}_M = \frac{1}{3\,|M|} \sum_{i:\, M_i=1} \sum_{c=1}^{3} (R_{i,c} - G_{i,c})^2\,, \qquad \mathrm{PSNR}_M = 10 \log_{10} \frac{255^2}{\mathrm{MSE}_M}, \tag{B1}$$

where $|M|$ is the number of masked pixels. Masked SSIM (Wang et al., 2004) computes the standard SSIM map S_c of each channel on the full image, using an 11×11 Gaussian window with $\sigma = 1.5$ and stabilizers $C_1 = (0.01 \cdot 255)^2$, $C_2 = (0.03 \cdot 255)^2$, and averages the map over the masked positions,

$$\mathrm{SSIM}_M = \frac{1}{3} \sum_{c=1}^{3} \frac{1}{|M'|} \sum_{i:\, M'_i=1} S_c(i), \tag{B2}$$

where M' restricts the mask to the valid region of the windowed map. LPIPS admits no pixel-wise restriction, so masked LPIPS is computed on a composite that differs from the ground truth only inside the mask,

$$X = M \odot R + (1 - M) \odot G, \qquad \mathrm{LPIPS}_M = \mathrm{LPIPS}\big(X|_\beta,\, G|_\beta\big), \tag{B3}$$

where β is the bounding box of M padded by 32 pixels with a minimum side of 64, and LPIPS uses the AlexNet backbone (Zhang et al., 2018). All metrics are computed per image and averaged over the benchmark. Since masked PSNR and SSIM read only masked pixels and the LPIPS composite equals the ground truth outside the mask, discrepancies outside the mask, such as lighting drift or passers-by in RORD's separately captured ground-truth frames, do not enter any reported metric.

Statistics. All effects and metric differences are paired at the image level. Significance statements in the main text and this appendix use a paired sign-flip permutation test over images with 10,000 permutations; where a family of related tests is reported (the four backbone gaps of Table C1), Holm correction is applied and all conclusions survive it.

B.2 Auxiliary-Pass Accounting of Baselines

Effective forwards. We count the number of network forward passes each method performs per removal, measured with a forward hook on the denoiser rather than taken from the nominal step count. Table B2 reports the hook-measured counts for the methods run in our pipeline on the native four-step schedule. Attentive Eraser duplicates the batch to evaluate the suppressed and unsuppressed branches of its redirection guidance inside one call, doubling the effective computation at every step; AdaEraser runs a separate reference forward alongside each target forward; DesignEdit modifies attention inside the existing forwards and adds none.

ReState reuses the first-forward features for retrieval, so it likewise stays at the native count. For the two methods of Table 1 that are not run in our pipeline, the auxiliary-pass classification follows the published procedures: PANDORA combines a stored-latent pass with its latent guidance, and FuLLaMa refines an initial removal through repeated diffusion passes, scoring each iterate with LVLM and CLIP evaluations during its runtime parameter search.

Table B2: **ReState adds no auxiliary pass and stays at the native forward count.** Hook-measured on the denoiser under the native four-step schedule: nominal counts backbone calls and effective sums their batch sizes (Appendix B), so the two doubling mechanisms differ; Attentive Eraser duplicates the batch, AdaEraser calls twice. ReState's retrieval probe is shared with the first generation forward and adds none.

Method	Nominal	Effective	Auxiliary pass
SDXL-Turbo	4	4	none
+ Attentive Eraser AAAI'25	4	8	SARG branch pair
+ DesignEdit AAAI'25	4	4	none
+ AdaEraser ICML'26	8	8	reference forward
+ **ReState (Ours)**	4	4	none

C Intervention Analysis

C.1 Complete First-Forward Analysis

Protocol. For a chosen forward t, the base and intervened runs start from the identical state z_t under the identical seed. The attention-forcing arm replaces $\mathcal{A}_t$ by $\mathcal{A}^{\mathrm{AF}}$ at that single forward; the direct arm perturbs the clean prediction by $\Delta_t^{\mathrm{direct}}$, rescaled to $\|\Delta_t^{\mathrm{AF}}\|_2$ as in the main text, and the perturbed prediction is re-expressed as a noise prediction through Eq. 1 before the sampler update. All subsequent forwards run without any intervention. The source-to-GT directional residual $p(I, M)$ involves no learned scorer; the only model is the VAE encoder. With $z = \mathcal{E}(I)$, $z_{\mathrm{src}} = \mathcal{E}(x^{\mathrm{src}})$, and z_{gt} the encoding of the object-free ground truth, it is the scalar projection of the output onto the source-to-ground-truth direction over the latent positions of the mask,

$$p(I, M) \;=\; \frac{\big\langle z - z_{\mathrm{gt}},\; z_{\mathrm{src}} - z_{\mathrm{gt}} \big\rangle_M}{\|z_{\mathrm{src}} - z_{\mathrm{gt}}\|_M^2}, \tag{C1}$$

so $p = 1$ when the output matches the source and $p = 0$ when the output matches the object-free ground truth. Effects are averaged over $n = 163$ images. All p-values in this appendix are from a paired sign-flip permutation test over images with 10,000 permutations; $p = 1.0 \times 10^{-4}$ is the smallest value expressible at this count.

Direction controls. The matched norm alone does not produce removal. Keeping the perturbation norm fixed and changing only its direction gives, at the first forward on SDXL-Turbo: toward retrieved background $+0.0251$ ($p = 1.0 \times 10^{-4}$); toward the ground-truth background (oracle direction) $+0.0487$ ($p = 1.0 \times 10^{-4}$); random direction -0.0004 ($p = 0.53$); direction orthogonal to the background direction -0.0006 ($p = 0.36$). Equal-norm perturbations therefore reduce the source-to-GT residual only when they point toward background content, which validates the norm-matched comparison of the main text. The terminal effect also tracks realized quality at the image level: across the first-forward runs, the per-image gap ΔE between the direct and attention arms

correlates with the corresponding masked PSNR and LPIPS changes of the same outputs (Spearman $\rho = 0.46$ and 0.38 over n=163 images, both $p < 10^{-6}$), and among the 133 of 163 images with $\Delta E > 0$, 94.7% also improve in masked PSNR and 93.2% in LPIPS.

Full first-forward results. Table C1 reports the means of Table 3 with their tests: the permutation p-value that the effect of attention forcing is zero, and that the gap d is zero. The gaps of all four distilled backbones remain significant after Holm correction across the four tests (all corrected p-values $\leq 4 \times 10^{-4}$), while the first-forward effect of attention forcing on the Turbo family is not significantly different from zero. With the original SDXL weights on the same grid both interventions act symmetrically and the gap vanishes. Figure C1 shows the same comparison at the image level: on the four distilled backbones 71 to 82 percent of images lie above the diagonal, while the weight control sits at 49 percent, not significantly different from chance.

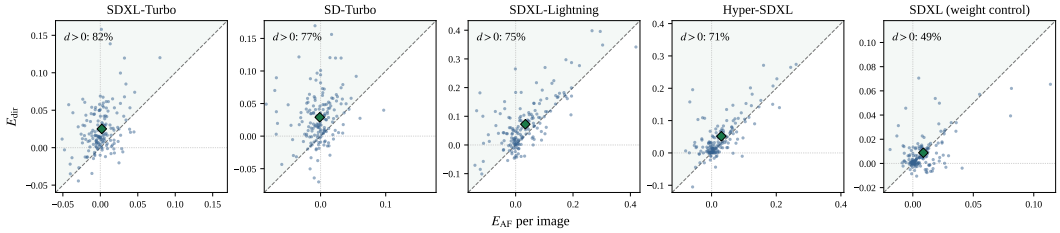

Figure C1: **The first-forward asymmetry holds image by image on the distilled backbones and vanishes with the original weights.** Each point is one image; the diamond marks the mean of Table C1, and $d > 0$ is the fraction of images above the diagonal, where the direct correction wins. Axes are clipped at the 0.5th and 99.5th percentiles and scaled per panel, as the backbones operate at different resolutions.

Table C1: **At the first forward, attention forcing is not significantly different from zero on the Turbo family, while the direct correction and the gap $d = E_{\text{direct}} - E_{\text{AF}}$ are not.** Means restate the first row of Table 3 at full precision. The p-values are paired permutation tests (Appendix B.1), whose resolution floor is 1.0×10^{-4}; the original-SDXL row is the weight control.

Backbone	E_{AF}	E_{direct}	d	$p(E_{\text{AF}}{=}0)$	$p(d{=}0)$
SDXL-Turbo	$+0.0020$	$+0.0251$	$+0.0231$	0.27	1.0×10^{-4}
SD-Turbo	-0.0013	$+0.0289$	$+0.0303$	0.52	1.0×10^{-4}
SDXL-Lightning	$+0.0339$	$+0.0723$	$+0.0384$	1.0×10^{-4}	1.0×10^{-4}
Hyper-SDXL	$+0.0302$	$+0.0516$	$+0.0214$	1.0×10^{-4}	1.0×10^{-4}
Original SDXL (4-step grid)	$+0.0087$	$+0.0088$	$+0.0001$	1.0×10^{-4}	0.91

C.2 Multi-Step Reference Intervention

Applying the identical matched single-forward protocol along the 50-step SDXL reference trajectory yields Table C2. Both interventions have terminal effects an order of magnitude smaller than on the native four-step trajectory, and no first-forward asymmetry comparable to Table C1 appears: a single early intervention, of either kind, is largely absorbed by the remaining trajectory. This is the reference behavior against which the few-step results should be read. Figure C2 places the two trajectories side by side.

C.3 Diagnostic Schedule Controls

Together with Tables C1 and C2, the two controls of this section complete a two-by-two design over the weights (distilled or original) and the schedule (four-step or dense): the intervention asymmetry appears only where distilled weights meet the coarse schedule. Both control configurations run outside the native inference regime of the respective model and are therefore diagnostic rather than a causal factorization. First, the original SDXL weights on the same four-step grid do not reproduce

Table C2: **On the reference trajectory both single-forward interventions stay an order of magnitude below the few-step effects.** Exact values at the eight probed noise levels; the corresponding curves are the right panel of Figure C2.

Forward index	Noise level σ	E_{AF}	E_{direct}
1	13.12	+0.0015	+0.0012
4	9.34	+0.0024	+0.0020
8	6.20	+0.0030	+0.0033
13	3.95	+0.0036	+0.0040
19	2.46	+0.0028	+0.0036
25	1.62	+0.0027	+0.0033
31	1.09	+0.0021	+0.0029
38	0.67	+0.0014	+0.0025

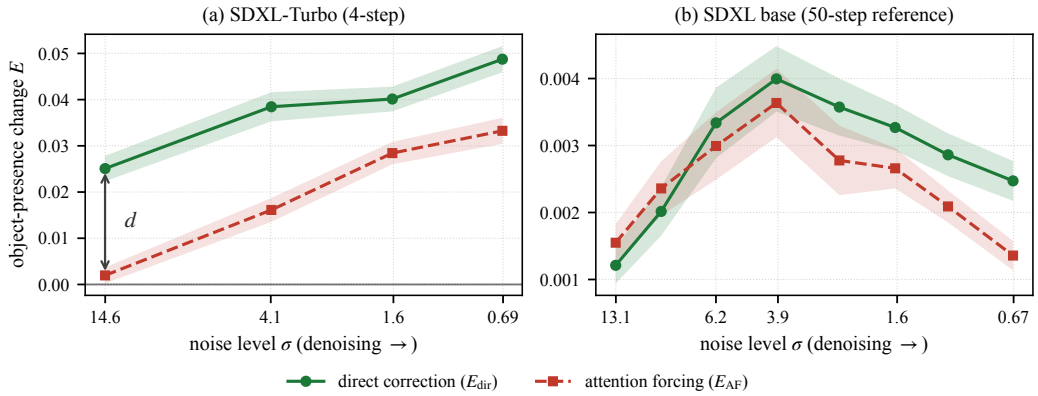

Figure C2: **On the reference trajectory both interventions are an order of magnitude weaker and the first-forward asymmetry vanishes.** Terminal intervention effect E of a single intervened forward at each noise level; bands are ± 1 standard error over $n = 163$ images. The left panel plots the SDXL-Turbo column of Table 3; the right panel plots Table C2, on a vertical scale an order of magnitude smaller.

the intervention gap: both interventions act moderately and symmetrically (last row of Table C1). Second, densifying the same SDXL-Turbo weights to 8 and 16 denoising evaluations shrinks the terminal effect of the first direct correction monotonically while the effect of attention forcing stays at its floor (Table C3): additional denoising opportunities reduce how much a single early correction survives to the output, and do not unlock an early effect of attention forcing.

Table C3: **Densifying the same Turbo weights shrinks the direct effect and does not unlock attention forcing.** First-forward effects at 4, 8, and 16 denoising evaluations; the native row anchors to Table C1.

Denoising evaluations	E_{AF}	$p(E_{\mathrm{AF}}{=}0)$	E_{direct}	$p(E_{\mathrm{direct}}{=}0)$
4	+0.0020	0.27	+0.0251	1.0×10^{-4}
8	+0.0018	0.21	+0.0122	1.0×10^{-4}
16	+0.0018	0.09	+0.0059	1.0×10^{-4}

On the 50-step reference we additionally fix the number of forced forwards and vary only their placement, so that any difference is attributable to timing rather than to the amount of intervention; the main text quotes the direction of this control, and Table C4 gives the configurations and numbers. The two eight-forward windows sit at the extremes of the schedule, $\sigma \in [13.1, 6.2]$ and $\sigma \in [0.44, 0.04]$, and the late window wins by 0.9 dB. Both fall far short of the sustained configu-

rations: extending attention forcing over the last 38 forwards recovers most of the reference quality, and only the full schedule reaches it. Timing matters at a fixed budget, but on the dense trajectory attention forcing needs sustained application to act at all.

Table C4: **At a fixed count, late attention forcing beats high-noise placement, and only the full schedule reaches the reference.** 50-step SDXL with the guidance of the reference configuration; the all-50 row is the 50-step attention-forcing reference of Table 5.

Placement	σ range	PSNR↑	SSIM↑	LPIPS↓
8 forwards, high-noise range	$13.1 \to 6.2$	11.57	0.255	0.399
8 forwards, late range	$0.44 \to 0.04$	12.47	0.284	0.353
Last 38 forwards	$3.95 \to 0.04$	17.17	0.429	0.221
All 50 forwards	$13.1 \to 0.04$	17.86	0.441	0.209

C.4 Retention Analysis

Retention. We measure how much a perturbation injected at the first forward survives to the final output. At forward k the predicted clean latent is pushed a fraction $\beta = 0.1$ toward the ground-truth background inside the mask, $\hat{x}_0' = (1-\beta)\,\hat{x}_0 + \beta\, z_{\mathrm{gt}}$, re-expressed as a noise prediction, and the trajectory then runs without further intervention. Retention is the ratio of the final to the immediate displacement between the injected and base runs, $s_k = \|\hat{x}_0^{\mathrm{final,inj}} - \hat{x}_0^{\mathrm{final,base}}\|_M \,/\, \|z^{\mathrm{inj}} - z^{\mathrm{base}}\|_M$, with masked RMS norms and the denominator measured at the input of forward $k{+}1$. The ratio is verified to be independent of the dose β, so it is the slope of the linear displacement response and applies to small perturbations along this direction. At the first forward the median over images is 1.38 on the native four-step SDXL-Turbo trajectory and 0.30 on the 50-step reference: the few-step trajectory preserves, and can even amplify, what the first forward establishes, while the dense trajectory attenuates it.

C.5 Residual-Frequency Analysis

We compare the high-band energy of the masked region against the ground-truth background. The band is the radial range $[0.45, 1.0)$ of Nyquist of the two-dimensional FFT power, computed on the grayscale bounding-box crop of the mask resized to 128^2 under a Hann window; the ground-truth crop is processed identically, and each image contributes the ratio of the two. Table C5 reports geometric means over images. Without intervention the masked region carries $8.3\times$ the high-band energy of the true background, the spectral signature of the object itself. Attention forcing alone reduces this only to $1.92\times$: the residual is object-shaped high-frequency structure. The direct correction alone lands at $0.30\times$, below parity, as expected from evidence B that averages background latents. ReState recovers to $0.51\times$: the object-shaped excess is eliminated, and the attention-constrained late forwards restore part of the missing texture. The remaining gap to parity is a smoother-than-real fill, not residual object structure, which is the opposite failure mode of attention forcing.

Table C5: **Attention forcing leaves object-shaped high-frequency structure; ReState eliminates it.** Masked high-band energy as a per-image ratio to the identically processed ground truth, geometric mean over OBER-Test (SDXL-Turbo).

Output	High-band ratio to ground truth
No intervention	8.34
Attention forcing only	1.92
Direct correction only	0.30
ReState	0.51

D METHOD ABLATIONS

Each ablation in this section varies one component of the default ReState configuration (Appendix B) on SDXL-Turbo and reports masked PSNR, SSIM, and LPIPS on OBER-Test. Tables report point estimates, and every statistical statement in the text is a two-sided paired permutation test against the default configuration under the protocol of Appendix B.1.

D.1 CORRECTION STRENGTH

Table D1 sweeps λ over a $32\times$ range with all other settings fixed. Removal quality is flat: masked PSNR varies by at most 0.14 dB across the sweep, and the default $\lambda = 8$ is within 0.01 dB of the optimum. The realized weights ω_t shift substantially with λ, so the insensitivity is a property of the correction itself rather than of a narrow operating point.

This flatness is a statement about the strength profile within the structure that the closed form imposes, with strong corrections while noise remains and none at the final forward; that the structure itself is necessary is shown in Section D.2.

Table D1: **Removal quality is insensitive to the correction strength λ.** Realized weights $\omega_t = \lambda\sigma_{t-1}^2/(1+\lambda\sigma_{t-1}^2)$ at the first three forwards; $\omega_4 = 0$ for all λ by the closed form. The default $\lambda = 8$ is highlighted.

	Realized ω			OBER-Test		
λ	ω_1	ω_2	ω_3	PSNR↑	SSIM↑	LPIPS↓
1	0.943	0.722	0.325	18.61	0.487	0.194
2	0.971	0.839	0.490	18.70	0.492	0.193
4	0.985	0.912	0.658	18.74	0.496	0.194
8	0.993	0.954	0.794	18.75	0.498	0.194
16	0.996	0.977	0.885	18.73	0.498	0.195
32	0.998	0.988	0.939	18.72	0.498	0.195

D.2 CORRECTION TIMING AND SCHEDULE STRUCTURE

Table D2 restricts the direct correction to subsets of the four forwards, with object-excluding attention active at every forward in all rows. The windows are ordered primarily by the last corrected forward: windows ending at the third forward reach 18.47 to 18.75 dB on OBER-Test, windows ending at the second reach 18.18 to 18.35 dB, and correcting only the first forward reaches 16.85 dB. Within a fixed last forward, earlier corrections add a secondary refinement ($+0.27$ dB from $\{3\}$ to $\{2,3\}$, $+0.17$ dB from $\{2\}$ to $\{1,2\}$). This ordering follows from state propagation: every uncorrected forward re-predicts the masked content with whatever generative freedom the remaining noise allows, so the final result is governed by the freedom that remains after the last applied correction. The same absorption principle governs trajectory length in Appendix C.3, where additional denoising evaluations reduce how much a single early correction survives to the output. These windows measure deployed quality with the remaining schedule uncorrected; the persistence of a single matched perturbation is the separate measurement of Appendix C.4. The two measurements are consistent: limited absorption (Appendix C.4) means an early correction survives, and the ordering by the last corrected forward follows because every uncorrected forward re-predicts the masked content; the window ordering is therefore a consequence of the same absorption principle, not a counterexample to it.

Within this pattern the first-forward correction becomes redundant once the later forwards are corrected: the $\{2,3\}$ window is not significantly different from the full schedule on PSNR and LPIPS ($p = 0.54$ and $p = 0.90$), with a marginal SSIM difference (0.496 versus 0.498, $p = 0.002$). ReState nevertheless keeps the full schedule, because the closed form of Eq. 7 assigns every forward its weight without a window hyperparameter, and pruning to $\{2,3\}$ would introduce a discrete rule for no measurable gain.

The last row removes the boundary condition instead: keying ω to σ_t rather than σ_{t-1} leaves the final forward a correction of $\omega_4 \approx 0.79$, whose corrected prediction is the final output. Quality collapses perceptually, with LPIPS degrading from 0.194 to 0.284 and PSNR losing 1.49 dB (both $p = 1.0 \times 10^{-4}$). This row also answers why the schedule does not shift weight toward the later forwards although single late windows imprint best: a late correction locks in whatever it writes, including the smoothness of the averaged evidence B, so the correction should be strong while the prediction is still revisable and absent once it is not. The final forward is left uncorrected, and the σ_{t-1} keying supplies exactly this as a boundary condition of the objective rather than as an added rule.

Restoring that boundary by hand under σ_t keying, with an explicit final-forward skip, is statistically not significantly different from the default on every metric (18.71 versus 18.75 dB, $p = 0.23$; LPIPS 0.195 versus 0.194, $p = 0.08$). The keying is therefore not a performance choice: σ_{t-1} absorbs the final-forward exception into the closed form. A complementary sweep isolates the terminal weight: with ω_1–ω_3 fixed at their closed-form values, raising ω_4 from 0 to 0.2, 0.5, and 0.794 degrades masked LPIPS monotonically from 0.194 to 0.210, 0.251, and 0.283 (significant from $\omega_4 = 0.2$, $p = 1.0 \times 10^{-4}$), with PSNR and SSIM following from $\omega_4 \geq 0.5$. The benefit of the terminal release is therefore separable from the earlier weights: a terminal direct correction pulls the output toward the smoothness of the averaged evidence B exactly where fine texture is synthesized.

Table D2: **The correction is cumulative and the final forward is left uncorrected.** The direct correction is applied only at the listed forwards; object-excluding attention stays active at every forward in all rows. The last row keys ω to σ_t, so the final forward retains a correction of $\omega_4 \approx 0.79$ whose output is final.

Corrected forwards	PSNR↑	SSIM↑	LPIPS↓
None (attention forcing only)	15.50	0.348	0.269
$\{1\}$	16.85	0.411	0.237
$\{2\}$	18.18	0.469	0.203
$\{3\}$	18.47	0.482	0.200
$\{1, 2\}$	18.35	0.477	0.198
$\{2, 3\}$	18.74	0.496	0.194
$\{1, 2, 3\}$ (default)	18.75	0.498	0.194
$\{1, 2, 3, 4\}$ (ω keyed to σ_t)	17.26	0.466	0.284

D.3 Background Evidence Source

Replacing the retrieved B (Table D3) shows both directions of the dependence: a global background mean, which removes spatial correspondence, costs 0.59 dB, while the ground-truth background as an oracle target gains 3.57 dB. The correction mechanism is therefore not the bottleneck; the quality of the evidence is.

Table D3: **Evidence quality, not the correction mechanism, is the remaining bottleneck.** Each row supplies a different evidence B to the identical correction.

Evidence B	PSNR↑	SSIM↑	LPIPS↓
Global background mean	18.16	0.481	0.214
Retrieved (default)	18.75	0.498	0.194
Ground-truth background (oracle)	22.32	0.588	0.133

D.4 Per-Forward Correction versus One-Shot Latent Initialization

The preceding ablations vary the evidence and the schedule; this one varies only how the same evidence enters sampling, and asks whether the gain of ReState is explained by obtaining a good proposal B alone. Both configurations share the first generation forward, the retrieved proposal

B, the seed, and the noise realization, and both run four denoiser evaluations with identical attention forcing and anchoring. ReState applies its per-forward correction as in the main text. The initialization alternative instead writes B into the trajectory once (Meng et al., 2022): at the noise level σ_{next} entering the second forward, the masked region is re-initialized as $z^{\text{init}} = M \odot (B + \sigma_{\text{next}}\,\epsilon) + (1 - M) \odot (z_{\text{src}} + \sigma_{\text{next}}\,\epsilon)$, after which the remaining three forwards run without any prediction correction. This differs from the $\{1\}$ window of Appendix D.2, which corrects the clean prediction and lets the sampler form the next latent; the initialization reconstructs the noisy latent itself, so the two are separate measurements. Masked LPIPS was declared the primary metric before the runs, since a one-shot fill can favor PSNR through smoothing. Table D4 reports the outcome: ReState is better on the primary metric on both benchmarks ($p = 1.0 \times 10^{-4}$; per-image win rate 62.0% on OBER-Test, 62.7% on RORD), and on masked PSNR and SSIM as well (win rates 69–90%). The gain of ReState is therefore not explained by the background proposal alone: at the same target and the same budget, applying the proposal as a per-forward prediction correction outperforms writing it into the latent once.

Table D4: **Per-forward correction versus one-shot latent initialization: the correction procedure itself adds value.** Both rows share the retrieved proposal B, the first generation forward, the noise realization, attention forcing, and anchoring, and use four denoiser evaluations. The initialization row writes B into the noisy latent once before the second forward; ReState applies it as a per-forward prediction correction. The last row reports ReState minus initialization; masked LPIPS is the pre-declared primary metric, and every gain is significant under the protocol of Appendix B.1 ($p = 1.0 \times 10^{-4}$).

Paired gains are computed from unrounded per-image values.

			OBER-Test			RORD		
Method	Same B	NFE	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
Initialization with B	✓	4	17.90	0.454	0.211	16.39	0.417	0.284
ReState (Ours)	✓	4	**18.75**	**0.498**	**0.194**	**17.21**	**0.451**	**0.273**
Paired gain			+0.85	+0.044	−0.017	+0.83	+0.033	−0.012

D.5 Retrieval Ratio

Figure D1 widens the retained fraction r over three orders of magnitude, up to the softmax-weighted average over the entire background. Below the default the bank is too small: $r = 0.001$ loses 0.33 dB ($p = 1.0 \times 10^{-4}$) and $r = 0.005$ still registers ($p = 0.010$). Above the default the sweep locates only a shallow numerical optimum: PSNR peaks around $r = 0.1$ to 0.25, at most 0.15 dB above the default and not significant on any metric ($p \geq 0.08$), while LPIPS is never significantly better than at the default (all $p \geq 0.15$) and is numerically worst at the full bank. We retain $r = 0.01$: the peak is not statistically distinguishable from it, and adopting a peak located on the evaluation benchmark would in any case amount to selecting r on the test set.

The sweep separates the two design choices of the retrieval. Even the full-bank softmax average stays at the level of the default on PSNR (18.80) while its LPIPS is the worst in the sweep, whereas removing the weighting entirely, the unweighted global mean of Table D3, costs 0.59 dB. The robustness to r is therefore a property of the similarity weighting rather than of a tuned retained fraction: across the entire range $r \in [0.005, 1]$, masked PSNR stays within 0.15 dB of the default.

D.6 Probe Feature Source

The background bank is built from features captured at the first generation forward ($\sigma = 14.6$, object-excluding attention active), at the output of the first up-block of the denoiser; the native resolution of this map at 512^2 is 32×32, and it is bilinearly resampled to the 64×64 correspondence grid of the latent. Table D5 rebuilds the bank from alternative feature sources and regenerates with the otherwise identical pipeline.

Every variant harvests its features in a separate pass of the diagnostic harness, so the generation trajectory itself is identical across rows and stays at the native four evaluations. For the suppression-off and noise-level rows this pass is required by construction: in the default the probe and the first

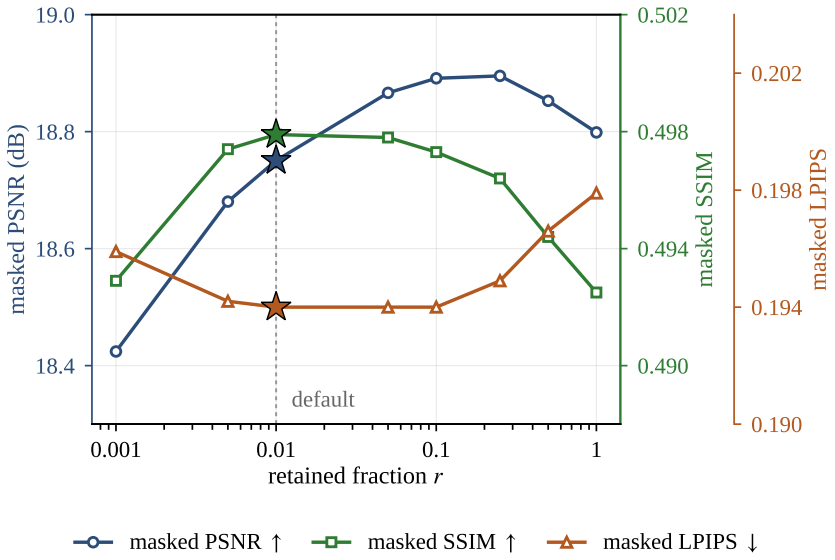

Figure D1: **Removal quality is insensitive to the retained fraction r across three orders of magnitude.** Masked metrics on OBER-Test with the bank widened up to the similarity-weighted average over the entire background ($r = 1$); the stars mark the default $r = 0.01$. Exact test statistics are given in the text.

generation forward are the same pass, run under suppression at $\sigma = 14.6$, so changing the probe condition inside that forward would change the generation as well. The layer rows need no such pass in principle and could share the generation forward exactly as the default does.

The shared probe is robust to the attention intervention under which it runs. Harvesting the same features without object-excluding attention changes nothing measurable (18.72 versus 18.75 dB, $p \geq 0.46$ on every metric). Keeping the suppression active therefore costs nothing, and it requires no separate pass.

The noise level of the probe input, in contrast, is functional rather than incidental. Harvesting from a nearly clean input, $\mathcal{E}(x_{\text{src}}) + \sigma_{\min}\epsilon$ at $\sigma_{\min} = 0.693$, degrades sharply to 17.44 dB and LPIPS 0.214 (all $p = 1.0 \times 10^{-4}$), while features from the second forward ($\sigma = 4.08$) remain at par ($p \geq 0.33$). This pattern is consistent with heavy noise acting as content suppression on the probe input.

At high σ the source object is largely erased from the input, so masked-position features are driven by surrounding context, whereas at low σ the object remains present and reaches the masked features through convolutional receptive fields that attention suppression does not gate, and the retrieval then favors background patches resembling the object. We advance this as the consistent reading of the pattern rather than as an isolated causal claim.

The layer rows separate semantic depth from resolution. Mid-block features at native 16×16 are not significantly different from the default ($p \geq 0.62$ on every metric), while the encoder-side block at the same 32×32 native resolution as the default degrades strongly (17.96 dB, LPIPS 0.216, $p \leq 0.002$), and the decoder block with native 64×64 resolution also loses (18.42 dB, LPIPS 0.207; $p \leq 0.002$ on SSIM and LPIPS, $p = 0.14$ on PSNR). Correspondence quality therefore tracks where the features sit in the denoiser rather than their native resolution, and the bilinear resampling of the default is not the limiting factor: the block that needs no resampling performs worse than the resampled default.

The bank values are fixed to clean source latents, so the only quantity that can change during sampling is the correspondence itself. Re-computing the similarity weights at every generation forward from its own shared features, with no auxiliary pass and the evidence space unchanged, raises quality slightly (19.02 versus 18.75 dB, LPIPS 0.190 versus 0.194), while the retrieved positions change substantially (mean top-k overlap of 0.39 to 0.43 against the first-forward matching).

The gain is consistent with the content-suppression reading above: features from the later forwards are harvested from a trajectory whose masked content has already been corrected toward background, so the probe input is object-free by prior correction rather than by noise. The default retains the single shared harvest as the minimal-intervention configuration, and we read the per-step update as the most immediate route toward the evidence headroom of Table D3.

Table D5: **The shared noisy probe is the operative choice: near-clean inputs contaminate the bank, while attention suppression and the feature layer admit slack.** Each row rebuilds the background bank from a different feature source and regenerates with the identical four-forward pipeline; every variant harvests its features in a separate diagnostic pass, so the deployed method stays at the native count.

Feature source	PSNR↑	SSIM↑	LPIPS↓
Shared first forward (default)	18.75	0.498	0.194
No attention suppression	18.72	0.497	0.194
Near-clean input ($\sigma = 0.693$)	17.44	0.474	0.214
Second forward ($\sigma = 4.08$)	18.87	0.499	0.193
Encoder block (native 32^2)	17.96	0.475	0.216
Mid block (native 16^2)	18.75	0.499	0.194
Decoder block (native 64^2)	18.42	0.482	0.207

E Additional Results

E.1 Full Quantitative Tables

Table E1 collects the full quantitative comparison on both benchmarks: the training-free block repeats the SDXL-Turbo rows of Table 2, and the lower block adds the training- and optimization-based methods, each run from its released checkpoint under the official default inference settings (Appendix B). SDXL-Inpainting and FLUX.1-Fill operate and are evaluated at 1024^2, and all remaining methods at 512^2. Latency is the median per-image end-to-end wall-clock on OBER-Test and VRAM is the peak allocation, both measured on a single A100 80GB for every row.

OSOR (Zhou et al., 2026) fine-tunes SDXL-Inpainting and FLUX.1-Fill on CORNE, a set of 280K removal pairs curated from single-edit instruction triplets, and additionally refines the FLUX variant for 1K steps on the captured subset of OBER. This refinement is described only for the FLUX variant; because both variants share the same training pipeline, we conservatively treat the same refinement as possible for the SDXL-Inpainting release as well, so out-of-distribution evaluation on OBER-Test is not guaranteed for either variant, and OSOR is excluded from the matched comparison of Table E1.

For reference, its SDXL-Inpainting variant reaches 23.19 dB masked PSNR on OBER-Test and 18.77 dB on RORD at 0.40 s and 7.2 GiB under the identical protocol. The training data of every training-based comparison is classified in Table B1.

E.2 Full Removal on Additional Distilled Backbones

Table E2 extends the full removal comparison of Table 2 to SDXL-Lightning and Hyper-SDXL under the identical frozen configuration: the same $\lambda = 8$, $r = 0.01$, probe layer, and schedule keying, with no per-backbone tuning. On both backbones and both benchmarks ReState improves over DesignEdit, the strongest training-free baseline of Table 2, on every metric, by $+1.46$ to $+2.39$ dB masked PSNR, and over the no-intervention baseline by 6.9 to 7.3 dB; every difference is significant at $p = 1.0 \times 10^{-4}$ under the protocol of Appendix B.1.

The margins over attention forcing are smaller than on the Turbo family ($+3.25$ dB on SDXL-Turbo), in the direction the intervention analysis predicts. On Lightning and Hyper-SDXL the first-forward terminal effect of attention forcing is already positive ($+0.034$ and $+0.030$, Table C1) rather than not significantly different from zero as on the Turbo family, so attention forcing recovers

Table E1: **Without any training, ReState remains competitive with the training-based methods at the lowest latency.** Latency and VRAM are measured as described in the text; the training-free rows repeat the SDXL-Turbo block of Table 2.

	OBER-Test			RORD				
Method	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	Lat. (s)	VRAM
Training/Optimization-based methods								
SDXL-Inpainting	11.65	0.365	0.391	10.57	0.348	0.403	3.75	9.0GB
FLUX.1-Fill	11.04	0.376	0.405	12.33	0.415	0.366	15.9	33.8GB
PowerPaint ECCV'24	14.89	0.463	0.255	14.91	0.438	0.282	3.86	5.1GB
CLIPAway NeurIPS'24	15.71	0.426	0.246	14.84	0.406	0.268	7.83	9.8GB
EntityErasure CVPR'25	15.70	0.496	0.222	15.70	0.460	0.247	3.21	12.5GB
SmartEraser CVPR'25	18.67	0.510	0.170	17.81	0.463	0.193	6.12	7.7GB
EraseLoRA ECCV'26	18.78	0.481	0.175	16.77	0.433	0.237	198s	33.5GB
Training-free methods								
SDXL-Turbo (4 steps)	9.86	0.133	0.435	7.55	0.074	0.516	0.247	7.1GB
+ Attentive Eraser AAAI'25	7.09	0.046	0.514	7.35	0.065	0.538	0.338	7.1GB
+ DesignEdit AAAI'25	15.50	0.348	0.269	14.34	0.326	0.334	0.281	7.1GB
+ AdaEraser ICML'26	11.71	0.205	0.372	9.47	0.122	0.462	1.145	7.1GB
+ **ReState (Ours)**	18.75	0.498	0.194	17.21	0.451	0.273	0.373	7.1GB

part of the removal on its own and the relative advantage of the direct correction narrows. On both benchmarks the ordering of the four ReState margins follows the ordering of that first-forward effect across the four distilled backbones; we read this as consistency with the diagnosis rather than as a causal prediction test.

The removal pipeline is unchanged; the two backbones operate and are evaluated at 1024^2 (Appendix B.1), so absolute metric levels are not comparable to the 512^2 backbones and comparisons are within each backbone, as in Table 3. Attentive Eraser and AdaEraser are omitted from this extension: both already fall below DesignEdit or the no-intervention baseline on the Turbo backbones of Table 2.

Table E2: **The single frozen configuration transfers: ReState improves every masked metric on two further distilled backbones.** Full removal at 1024^2 with the identical hyperparameters as Table 2 and no per-backbone tuning; rows with + apply the method on the backbone of their block. Bold is best per column within a block; test statistics are given in the text.

	OBER-Test			RORD		
Method	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
SDXL-Lightning (4 steps)	10.51	0.393	0.424	9.60	0.318	0.445
+ DesignEdit AAAI'25	15.95	0.517	0.296	15.08	0.475	0.325
+ **ReState (Ours)**	**17.51**	**0.578**	**0.271**	**16.54**	**0.528**	**0.308**
Hyper-SDXL (4 steps)	10.21	0.346	0.434	9.21	0.273	0.455
+ DesignEdit AAAI'25	15.20	0.487	0.315	14.14	0.434	0.358
+ **ReState (Ours)**	**17.49**	**0.574**	**0.273**	**16.53**	**0.527**	**0.315**

E.3 Multi-Step Baseline Comparison

Table E3 runs the attention-forcing baselines of Table 2 at their native 50-step count, under the identical reference configuration as Table C4 (same guidance, seed, and background anchoring); the DesignEdit row is the 50-step reference of Table 5. The methods that degrade under the four-step port operate correctly at their native step count: Attentive Eraser reaches 18.79 dB against 7.09 dB in Table 2, and AdaEraser 15.66 against 11.71 dB. The few-step failures of Section 4 are therefore consistent with a regime effect rather than a port artifact, and the recovery of Attentive Eraser is

exactly the calibration argument of the main text: its SARG amplification is calibrated for 50-step sampling and acts as intended there.

Against this 50-step field, ReState at its native four steps (18.75 dB, SSIM 0.498, LPIPS 0.194) exceeds every 50-step attention configuration on SSIM and LPIPS, and sits within 0.04 dB of the best attention-configuration PSNR (Attentive Eraser), at $6.7\times$ lower latency (Table 5).

Transferring ReState unchanged to the 50-step schedule under-performs (17.57 dB, LPIPS 0.231): the σ_{t-1}-keyed weight keeps the correction strong across the many low-noise forwards, so the smooth evidence B is imprinted repeatedly in the range where fine texture is synthesized.

Restricting the correction to transitions with $\sigma_{t-1} \geq \theta$ recovers quality: over $\theta = \infty, 4.08, 1.61, 0.693, 0.1, 0$ masked PSNR moves as $17.86, 18.82, 18.83, 18.21, 17.57, 17.57$, a plateau at $\theta = 4.08$ to 1.61 that recovers past the naive transfer to the level of the strongest 50-step attention configuration. We report the sweep in full; the windowed row of the table is the plateau configuration and is a diagnostic, not part of the method. Notably, its LPIPS (0.181) improves on the four-step default (0.194), indicating that part of the under-textured residual of Appendix C.5 reflects the limited free-synthesis budget of the coarse schedule rather than the correction itself.

This behavior refines, and is predicted by, the schedule principle of Appendix D.2: what must be preserved is uncorrected synthesis freedom before terminal rendering. On the native coarse schedule the closed-form boundary $\omega \to 0$ provides exactly this, because a single low-noise forward remains; on a dense schedule the boundary condition and the freedom requirement separate, and an explicit window substitutes for it. This is one more reason the operating point of the method is the native few-step schedule.

Table E3: **Attention forcing recovers in its 50-step regime, and 4-step ReState matches the strongest 50-step configuration at $6.7\times$ lower latency.** OBER-Test under the 50-step reference configuration; the 4-step port numbers of the same methods are in Table 2, and the correction-window sweep is given in the text.

Configuration at 50 steps	PSNR↑	SSIM↑	LPIPS↓
+ DesignEdit AAAI'25 (reference)	17.86	0.441	0.209
+ Attentive Eraser AAAI'25	18.79	0.462	0.225
+ AdaEraser ICML'26	15.66	0.399	0.274
+ ReState, every forward corrected	17.57	0.479	0.231
+ ReState, corrections at $\sigma_{t-1} \geq 1.61$	18.83	0.485	0.181
ReState on SDXL-Turbo (4 steps; Table 2)	18.75	0.498	0.194

E.4 QUALITATIVE RESULTS ON OBER

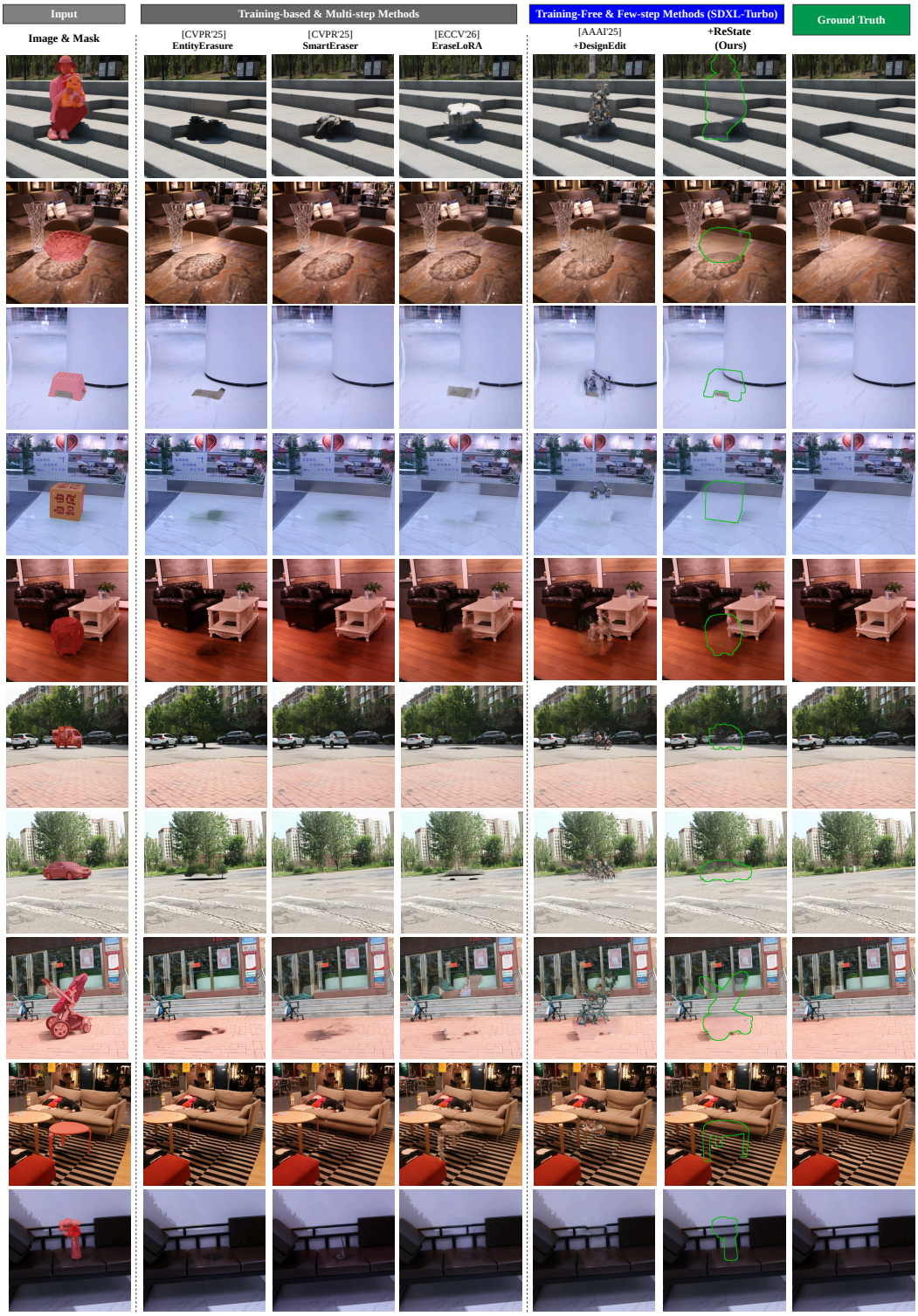

Figure E1: **Qualitative results on OBER-Test.**

E.5 QUALITATIVE RESULTS ON RORD

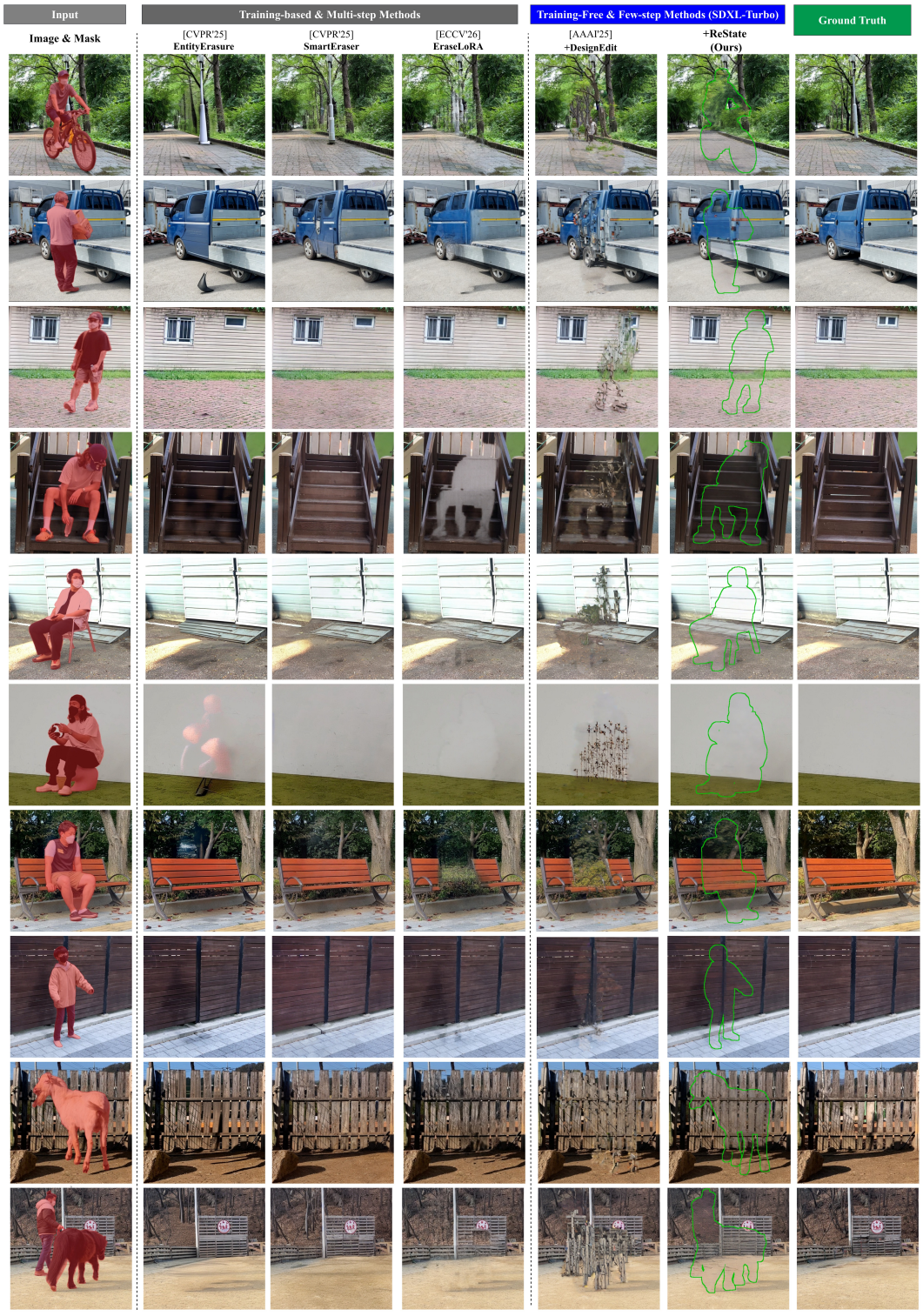

Figure E2: **Qualitative results on RORD.**