REFERENCE-STATE DECOUPLING REGENERATION FOR RESTORING IMAGE EDITABILITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Protective perturbations can severely disrupt downstream image editing while remaining visually similar to clean images. Existing diffusion-based purification methods typically initialize generation from a noised protected image, causing image content and protective perturbations to enter the initial generative state together. Stronger noising may suppress the perturbations but also discard the content required for faithful editing. To better understand this trade-off, we distinguish two roles of the protected image: state initialization, which directly introduces protected pixels into the generative state, and reference conditioning, which provides image information to guide generation. We investigate how these distinct input roles affect editability restoration under protective perturbations. A diagnostic experiment shows that progressively reducing the protected image's contribution to the initial state improves editability restoration while maintaining comparable reconstruction fidelity, with independent initialization achieving the highest Recovery. Motivated by this finding, we propose Reference-State Decoupled Regeneration (RSDR), which retains the protected image as a fixed reference while initializing the evolving pixel state from independent Gaussian noise. Time-dependent reference features guide the state trajectory through a lightweight interface, enabling content-preserving regeneration without directly injecting protected pixels into the initial state. Experiments across four protection methods demonstrate consistent improvements in editability restoration over existing baselines. These results demonstrate the effectiveness of decoupling reference conditioning from state initialization for restoring image editability under protective perturbations. Code will be released upon acceptance.

1 INTRODUCTION

Diffusion-based image editing enables flexible manipulation of image content through text instructions but also raises concerns about unauthorized modification (Brooks et al., 2023; Couairon et al., 2023; Wang et al., 2023). To deter unauthorized editing, image protection methods such as PhotoGuard (Salman et al., 2023) and DiffusionGuard (Choi et al., 2025) introduce visually imperceptible perturbations that disrupt downstream editing models, causing edited outputs to become distorted or deviate from the intended edits. However, these methods are typically evaluated under fixed editing pipelines, and their protective effects may not persist when preprocessing is applied before editing. Restoring the editability of protected images therefore provides a means to assess the robustness of existing protection methods under such conditions. This requires reducing the influence of protective perturbations while retaining sufficient image content for downstream editing.

Existing methods attenuate protective perturbations through reconstruction-consistency optimization or diffusion-based regeneration (Cao et al., 2023; Zhao et al., 2024). A common design in diffusion-based purification initializes generation from a noised protected image (Nie et al., 2022). As shown in Figure 1(a), image content and protective perturbations enter the initial state together, so stronger noising may suppress perturbations at the cost of content loss. In contrast, the protected image can serve as a reference condition, guiding generation without directly determining the initial state, as shown in Figure 1(b). This raises a fundamental question: *Does the role of the protected image affect how effectively its editability can be restored?*

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.

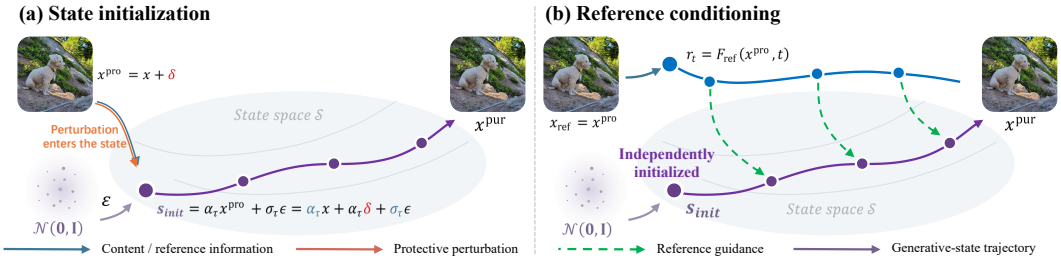


Figure 1: Two roles of the protected input in generative purification. (a) State initialization directly places both image content and the protective perturbation in the starting state. (b) Reference conditioning guides the evolution of an independently initialized state without directly incorporating the protected pixels into the starting state.

To investigate this question, we conduct a diagnostic experiment that varies the protected image's contribution to the initial generative state while retaining it as a reference condition. As shown in Figure 2, editability restoration consistently improves as this contribution decreases, with independent initialization achieving the highest Recovery. Meanwhile, reconstruction fidelity remains comparable across different initialization settings, indicating that reducing the protected image's contribution to the initial state does not necessarily compromise content preservation. Notably, the highest Recovery does not coincide with the highest PSNR, suggesting that improved reconstruction fidelity alone does not account for the observed gains in editability restoration. These findings suggest that the protected image can provide the content required for downstream editing through reference conditioning without directly determining the initial generative state. This observation provides empirical support for decoupling reference conditioning from state initialization.

Building on this insight, we propose Reference-State Decoupled Regeneration (RSDR), which separates reference conditioning from state initialization to restore the editability of protected images. Specifically, RSDR retains the protected image as a fixed reference while initializing the evolving pixel state from an independent Gaussian prior. Time-dependent reference features modulate the state dynamics through a lightweight interface, allowing image content to guide regeneration without directly injecting protected pixels into the initial state. RSDR builds upon a pretrained pixel-space generative backbone, with only the reference-to-state interface adapted using synthetic degradations, without requiring protection-specific training examples.

Our main contributions are summarized as follows:

- We revisit generative purification from the perspective of input roles, distinguishing state initialization from reference conditioning. Through a diagnostic experiment, we show that reducing the protected image's contribution to the initial state improves editability restoration while maintaining comparable reconstruction fidelity, motivating the decoupling of these two roles.
- We propose RSDR, which retains the protected image as a fixed reference while initializing the generative state from an independent Gaussian prior. A lightweight interface incorporates time-dependent reference features into the state dynamics, enabling content-guided regeneration without directly injecting protected pixels into the initial state.
- Extensive experiments across multiple protection methods and diverse downstream editing settings demonstrate that RSDR consistently and substantially improves editability restoration over existing baselines. Further analyses validate the effectiveness and generalizability of the proposed reference–state decoupled design.

2 Related Work

Protective perturbations for image editing. Protective perturbations disrupt different components of diffusion-based editing pipelines. PhotoGuard targets the image encoder and diffusion process (Salman et al., 2023), while DiffusionGuard perturbs early denoising stages and uses mask augmentation to improve robustness across editing regions(Choi et al., 2025). Other methods perturb

latent representations or attention mechanisms (Chen et al., 2024; Lo et al., 2024; Jeon et al., 2025), whereas DiffVax amortizes perturbation generation with a learned immunizer (Ozden et al., 2026). Together, these works illustrate diverse protection mechanisms against diffusion-based editing.

Purification of adversarial and protective perturbations. DiffPure applies forward noising and reverse diffusion for adversarial purification, primarily in image classification (Nie et al., 2022). For protective perturbations, IMPRESS uses reconstruction-consistency optimization (Cao et al., 2023), while GrIDPure and FreqPure use diffusion-based reconstruction with patch-wise and frequency-aware designs, respectively(Zhao et al., 2024; Ju et al., 2025). PDM-Pure exploits the empirical robustness of pixel-space diffusion models, whereas EditorClean uses heterogeneous reconstruction to exploit the limited transferability of protective perturbations across architectures (Xue & Chen, 2024; Zhao et al., 2026). Conversely, AntiPure studies protective perturbations designed to resist diffusion-based purification (Yang et al., 2025). Across these approaches, reference conditioning and state initialization are not independently controlled under a fixed generative backbone. We instead study these two roles through controlled interventions while holding the generative backbone fixed.

Conditional regeneration and pixel-space flow models. SDEdit initializes generation from a noised input, whereas DiffBIR uses a restored image as an external condition for diffusion-based regeneration (Meng et al., 2022; Lin et al., 2024). Together, they exemplify two distinct ways an input can participate in regeneration: as part of the evolving state or as an external condition. Flow-based generative models learn continuous transports between source and data distributions through velocity fields (Lipman et al., 2023; Liu et al., 2023; Ma et al., 2024). A recent pixel-space formulation further combines semantic and pixel-level generation through conditional modulation (Ma et al., 2026). We adopt its pretrained pixel-generation backbone and study how the protected input participates in reference conditioning and state initialization.

3 Problem Formulation and Input-Role Analysis

3.1 Editability Restoration under Protective Perturbations

Protective perturbations are designed to prevent unauthorized image editing by disrupting the behavior of downstream generative editors. Let $x \in \mathbb{R}^{3\times H\times W}$ denote a clean image and let its protected counterpart be $x^{\mathrm{pro}} = x + \delta, \|\delta\|_\infty \le \varepsilon_p$, where δ is a protective perturbation and ε_p is its perturbation budget. Although x^{pro} remains visually similar to x, the perturbation is optimized to cause editing failures such as visual distortion, subject corruption, or deviation from the editing instruction.

We consider editability restoration through a generative purification step applied before a fixed downstream editor. A purifier $\mathcal{P}$ maps x^{pro} to $x^{\mathrm{pur}} = \mathcal{P}(x^{\mathrm{pro}})$. At inference, the purifier has access only to x^{pro}, without observing its clean counterpart x, the downstream editing mask, or the text instruction.

Given a fixed editor $\mathcal{E}$, an editing mask m, a text instruction p, and editor randomness ω, we define the clean, protected, and purified editing results as

$$y^{\mathrm{clean}} = \mathcal{E}(x; m, p, \omega), \quad y^{\mathrm{pro}} = \mathcal{E}(x^{\mathrm{pro}}; m, p, \omega), \quad y^{\mathrm{pur}} = \mathcal{E}(x^{\mathrm{pur}}; m, p, \omega). \tag{1}$$

The objective of purification is to restore the editing behavior observed on the clean input. Specifically, y^{pur} should preserve the original subject content while correctly following the editing instruction. Meanwhile, x^{pur} should retain the perceptual and semantic content of x before editing. This setting evaluates whether protective perturbations remain effective after an adversarial editor applies a preprocessing-based countermeasure.

3.2 Two Roles of the Protected Input in Generative Purification

Two ways of incorporating an input image appear in related generative tasks. Diffusion-based purification can initialize reverse generation from a noised input (Nie et al., 2022), whereas conditional image restoration can use an observed image to guide generation from an independently sampled state (Lin et al., 2024). In the setting of protected-image purification, we distinguish two roles of the input: *state initialization* and *reference conditioning*.

State initialization. In diffusion-based purification, a protected image is noised and used to initialize reverse generation. At noising time τ, the starting state is $s_\tau = \alpha_\tau x^{\text{pro}} + \sigma_\tau \epsilon$, where $\epsilon \sim \mathcal{N}(0, I)$. As illustrated in Figure 1(a), both the image content and the protective perturbation contribute to this state. Increasing τ reduces their contributions together, so removing the perturbation by stronger noising may come at the cost of image content.

Reference conditioning. The protected image can instead guide generation as a reference, while the starting state is sampled independently. Specifically, $s_0 \sim \mathcal{N}(0, I)$ is independent of x^{pro}, and $r_t = F_{\text{ref}}(x^{\text{pro}}, t)$ conditions the evolving state at generation time t. As illustrated in Figure 1(b), the protected pixels do not directly enter the starting state. The reference representation may still convey the perturbation, so whether this role improves recovery requires empirical evaluation.

This distinction matters for protected images because purification must suppress the perturbation while retaining the content needed for editing. Stronger noising weakens the protected image's contribution to the initial state, but may also discard useful content. Reference conditioning keeps the protected pixels out of the initial state, although the reference may still convey the perturbation. This raises the question of how the role of the protected image affects editability restoration.

3.3 A Diagnostic Study of Input Roles

To investigate how the input roles identified above affect editability restoration, we vary the protected image's contribution to the initial generative state while retaining it as a reference condition. Specifically, we set $s_\eta = (1-\eta)x^{\text{pro}} + \eta\epsilon$, where $\epsilon \sim \mathcal{N}(0, I)$ and η controls the contribution of the protected image. Increasing η reduces this contribution, with $\eta = 1$ yielding an independently sampled state. A separate model is trained for each value of η using the same initialization scheme during training and inference. All models share the same backbone, training data, and training budget, and are evaluated using the same downstream editor.

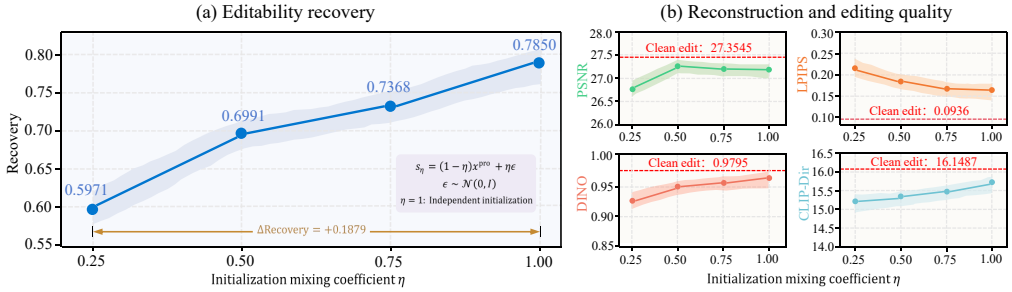


Figure 2: State-initialization analysis under protected-reference conditioning. Each value of η corresponds to a separately trained model with matched initialization during training and inference. (a) Editability recovery across initialization settings. (b) Reconstruction and editing quality. Dashed lines indicate clean-edit references, and shaded regions denote 95% confidence intervals.

As shown in Figure 2, Recovery increases from 0.5971 at $\eta = 0.25$ to 0.7850 at $\eta = 1$. Meanwhile, reconstruction fidelity remains stable, with PSNR ranging from 26.68 to 27.30 and LPIPS decreasing as η increases. PSNR peaks at $\eta = 0.50$, whereas independent initialization achieves the highest Recovery. DINO similarity and CLIP-Dir also improve with increasing η. These results indicate that reducing the protected image's contribution to the initial state improves editability restoration without compromising the reconstruction and editing metrics. These findings suggest that the protected image need not directly determine the initial generative state to preserve the content required for downstream editing when it remains available as a reference condition. This observation provides empirical support for decoupling reference conditioning from state initialization.

4 RSDR: Reference–State Decoupled Regeneration

Building on the input-role analysis in Section 3, we propose RSDR, a reference-conditioned pixel flow that separates reference conditioning from state initialization, as illustrated in Figure 3.

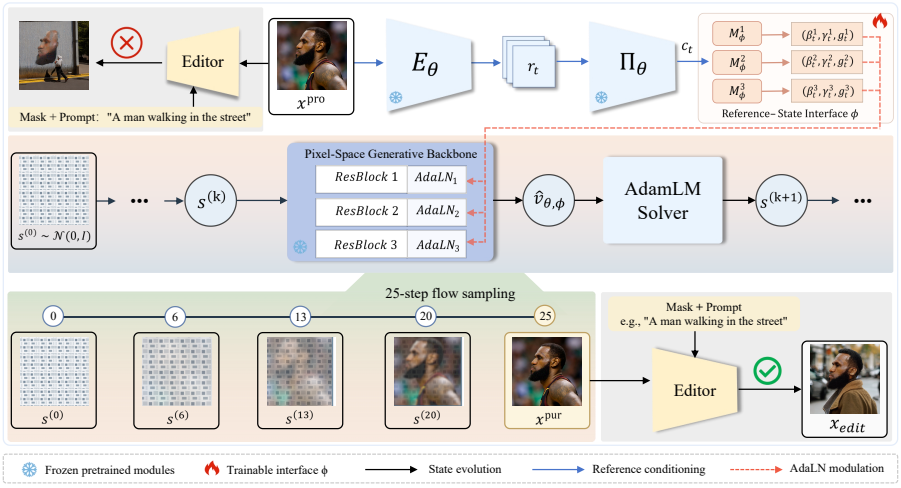


Figure 3: Overview of RSDR. The protected input is held fixed as the reference and produces a time-dependent condition, while the RGB state is initialized from independent Gaussian noise. Trainable modulation projections inject the reference condition into a frozen pixel-generation backbone during state evolution. The terminal state is returned as the purified image and subsequently passed to the downstream editor together with the editing mask and text instruction.

4.1 Reference–State Decoupled Formulation

Let x_{ref} denote the reference image and $s_t \in \mathbb{R}^{3\times H\times W}$ the evolving RGB state at generation time $t \in [0, 1]$. RSDR initializes the generative state from an independent Gaussian prior, $s_0 \sim \mathcal{N}(0, I)$, while keeping the reference image fixed throughout generation. At time t, the reference pathway produces $r_t = E_\theta(x_{\text{ref}}, t)$ and $c_t = \Pi_\theta(r_t)$, where E_θ extracts reference features and Π_θ maps them to the spatial condition used by the pixel network. Although the reference image remains fixed, its time-dependent condition c_t evolves with generation time. The reference condition modulates the velocity field governing state evolution:

$$\frac{ds_t}{dt} = \hat{v}_{\theta,\phi}(s_t, t \mid x_{\text{ref}}) = D_{\theta,\phi}(s_t, t, c_t), \qquad s_0 \sim \mathcal{N}(0, I). \tag{2}$$

Integrating the conditional dynamics from $t = 0$ to $t = 1$ yields the regenerated image $x^{\text{pur}} = s_1$. At inference, we set $x_{\text{ref}} = x^{\text{pro}}$, allowing the protected image to condition the state trajectory without directly determining its initialization.

4.2 Reference-Conditioned State Evolution

To guide state evolution without directly incorporating protected pixels into its initialization, RSDR introduces a time-dependent modulation interface between the reference and state pathways. We instantiate the conditional velocity field using a pretrained pixel-space generative backbone (Ma et al., 2026). The backbone operates on the evolving RGB state s_t, while the reference-derived condition c_t modulates three residual blocks at each generation time.

Let h_t^ℓ denote the state feature entering residual block $\ell \in \{1, 2, 3\}$. The corresponding interface projection M_ϕ^ℓ produces a shift, scale, and residual gate, $(\beta_t^\ell, \gamma_t^\ell, g_t^\ell) = M_\phi^\ell(c_t)$, which modulate the state features as

$$\tilde{h}_t^\ell = (1 + \gamma_t^\ell) \odot \text{LN}(h_t^\ell) + \beta_t^\ell, \tag{3}$$

$$h_t^{\ell+1} = h_t^\ell + g_t^\ell \odot F_\theta^\ell(\tilde{h}_t^\ell). \tag{4}$$

Here, LN denotes layer normalization, F_θ^ℓ is the frozen residual transformation, and M_ϕ^ℓ is the trainable interface projection. The shift β_t^ℓ and scale $1 + \gamma_t^\ell$ modulate the normalized state features, while g_t^ℓ controls the reference-conditioned residual update. As c_t is recomputed at each generation time, the fixed reference provides time-dependent guidance throughout the state trajectory. Thus, reference information influences generation through the conditional velocity field rather than directly determining the initial state.

4.3 Learning the Reference–State Interface

We learn the reference–state interface from clean images paired with synthetically degraded references. Given a clean image x, we construct $x_{\text{ref}} = \mathcal{A}(x; \xi)$, where $\xi \sim p_{\mathcal{A}}$ and $\mathcal{A}$ applies Gaussian noise, Gaussian blur, or their combination. We sample $\epsilon \sim \mathcal{N}(0, I)$ independently of (x, ξ) and $t \sim p_t$ to construct the flow-matching path:

$$s_t = (1-t)\epsilon + tx, \qquad v_t^* = x - \epsilon. \tag{5}$$

Although the reference and target are derived from the same clean image, the degraded reference is not used to initialize the generative state. This preserves the separation between reference conditioning and state initialization during training. The interface is trained using only synthetic degradations, without protection-specific examples.

We train the reference–state interface by matching the predicted velocity $\hat{v}_t = \hat{v}_{\theta,\phi}(s_t, t \mid x_{\text{ref}})$ with the target velocity v_t^* in both pixel and frequency domains. Specifically, $\mathcal{T}$ denotes RGB-to-YCbCr conversion followed by a non-overlapping 8×8 blockwise discrete cosine transform, and W is the normalized reciprocal of the JPEG luminance and chrominance quantization tables (Ma et al., 2026). The training objective is

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{x,\xi,\epsilon,t}\left[\text{mean}\big((\hat{v}_t - v_t^*)^2\big)\right], \tag{6}$$

$$\mathcal{L}_{\text{Freq}} = \mathbb{E}_{x,\xi,\epsilon,t}\left[\text{mean}\big(W \odot [\mathcal{T}(\hat{v}_t) - \mathcal{T}(v_t^*)]^2\big)\right], \tag{7}$$

$$\mathcal{L}_{\text{RSDR}} = \mathcal{L}_{\text{FM}} + \mathcal{L}_{\text{Freq}}, \tag{8}$$

where mean averages over spatial and channel elements. We optimize only the reference-to-state interface $\{M_\phi^\ell\}_{\ell=1}^3$, while keeping the reference encoder, condition projection, and pixel-space backbone frozen. This confines adaptation to the mapping from reference conditions to state modulation.

5 Experiments

5.1 Experimental Setup

Datasets and editing protocol. We use InpaintGuardBench (Choi et al., 2025), comprising 42 images of humans, animals, and objects, each paired with a SAM (Kirillov et al., 2023) mask and 10 editing prompts. This yields 420 editing instances per protection setting. We use Stable Diffusion Inpainting v1.5 (Rombach et al., 2022) at 512×512 resolution to edit the complement of the subject mask. All methods share the same images, masks, prompts, editor settings, and random seed.

Protection methods and baselines. We evaluate four protection methods: PhotoGuard-E and PhotoGuard-D (Salman et al., 2023), DiffusionGuard (Choi et al., 2025), and AdvPaint (Jeon et al., 2025). The first three use an ℓ_∞ perturbation budget of $16/255$, while AdvPaint follows OptimBox with a final 8-bit bound of $8/255$. We compare RSDR against JPEG compression, IMPRESS (Cao et al., 2023), GrIDPure (Zhao et al., 2024), and PDM-Pure (Xue & Chen, 2024).

Evaluation metrics. We assess subject preservation using subject-region PSNR, LPIPS (Zhang et al., 2018), and DINO (Caron et al., 2021) relative to the clean input. CLIP-Dir (Radford et al., 2021; Gal et al., 2022) measures image-text editing direction alignment. Let m_k, m_k^{clean}, and m_k^{prot} denote the metric means of the evaluated method, clean edits, and protected edits, respectively. We define normalized recovery as

$$\text{Recovery} = \frac{1}{|\mathcal{K}|} \sum_{k \in \mathcal{K}} \frac{m_k - m_k^{\text{prot}}}{m_k^{\text{clean}} - m_k^{\text{prot}}}, \tag{9}$$

where $\mathcal{K} = \{\text{PSNR}, \text{LPIPS}, \text{DINO}, \text{CLIP-Dir}\}$. Clean and protected edits yield Recovery scores of 1 and 0, respectively. Recovery is not clipped to $[0, 1]$ and is reported alongside the raw metrics.

Implementation details. We adapt a pretrained pixel-generation backbone (Ma et al., 2026) on 10,000 FFHQ (Karras et al., 2019) images, updating only the AdaLN modulation projections (9,504 parameters). All other parameters remain frozen. Training uses synthetic Gaussian noise and blur. At inference, RSDR generates from independent Gaussian noise in 25 sampling steps before passing the purified image to the downstream editor with the editing mask and text instruction. Implementation details and additional results are provided in Appendices A– G.

Table 1: Editability restoration under the unified SAM protocol. Clean and protected edits define empirical Recovery references and are excluded from ranking. CLIP-Dir is scaled by 100. Best and second-best purification results are shown in **bold** and underlined, respectively.

Protection	Method	PSNR↑	LPIPS↓	DINO↑	CLIP-Dir↑	Recovery↑
	Clean edit ref.	27.3545	0.0936	0.9795	16.1487	1.0000
PhotoGuard-E	Protected edit ref.	21.9495	0.3829	0.7155	13.6280	0.0000
	JPEG	23.5163	0.3316	0.8231	13.5248	0.2085
	GrIDPure	23.4651	0.2234	0.8670	14.8753	0.4751
	IMPRESS	21.1709	0.3731	0.7472	13.3396	-0.0262
	PDM-Pure	25.1600	0.2326	0.8077	15.1080	0.5124
	RSDR(ours)	**25.7431**	**0.1964**	**0.9416**	**15.5179**	**0.7381**
PhotoGuard-D	Protected edit ref.	19.1197	0.4222	0.6561	11.0149	0.0000
	JPEG	21.8393	0.3620	0.7684	12.8729	0.3057
	GrIDPure	24.2952	0.2196	0.8727	14.6294	0.6547
	IMPRESS	20.3859	0.3911	0.7074	12.1952	0.1592
	PDM-Pure	25.6126	0.2184	0.8218	**15.1891**	0.6836
	RSDR(ours)	**25.7187**	**0.2060**	**0.9185**	15.0685	**0.7651**
DiffusionGuard	Protected edit ref.	23.5649	0.3052	0.8249	14.8288	0.0000
	JPEG	24.3196	0.2766	0.8640	15.3675	0.2488
	GrIDPure	24.3033	0.1862	0.8696	15.3134	0.3534
	IMPRESS	24.1634	0.2944	0.8352	15.1378	0.1274
	PDM-Pure	25.5483	0.2164	0.8196	15.2105	0.2995
	RSDR(ours)	**26.3790**	**0.1760**	**0.9482**	**15.8510**	**0.7313**
AdvPaint	Protected edit ref.	23.4882	0.3974	0.5333	15.1148	0.0000
	JPEG	25.1767	0.3211	0.7972	14.9084	0.2699
	GrIDPure	24.3565	0.1993	0.8558	15.4967	0.4922
	IMPRESS	24.0034	0.3826	0.5775	14.8108	-0.0032
	PDM-Pure	25.6406	0.2189	0.8171	15.1567	0.4552
	RSDR(ours)	**27.2187**	**0.1741**	**0.9526**	**15.6323**	**0.7850**



(a) Editing setup **(b)** Downstream edited outputs

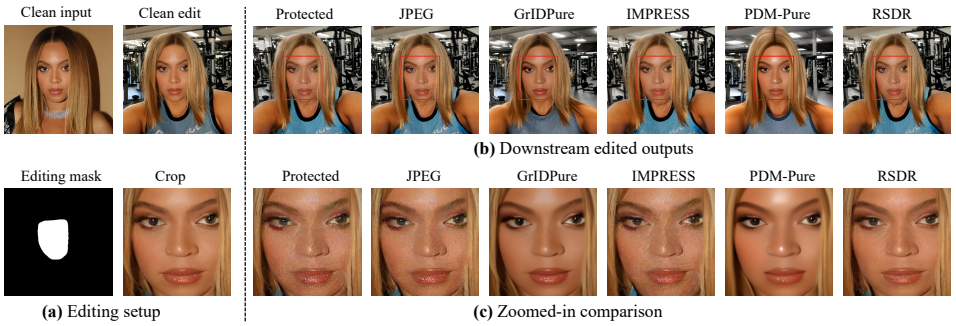
Figure 4: Qualitative comparison under the unified SAM editing protocol. (a) Editing setup and clean-edit reference. (b) Downstream edited outputs under four protection settings. The SAM subject mask is inverted to preserve the subject during editing.

5.2 Editability Restoration

We evaluate RSDR across four protection methods under the unified SAM editing protocol. As shown in Table 1, RSDR achieves the highest Recovery among all evaluated purification methods across the four protection settings, with an average score of 0.7549 and an average absolute gain of 0.2445 over the strongest baseline in each setting. It also achieves the highest PSNR and DINO similarity and the lowest LPIPS across all settings, indicating consistent improvements in subject preservation. RSDR obtains the highest CLIP-Dir in three of the four settings, with PDM-Pure achieving a slightly higher score under PhotoGuard-D. In contrast, existing baselines exhibit varying recovery performance across protection methods. Figure 4 shows a representative example. Given the same input, SAM subject mask, prompt "A cat in a hospital," and random seed, editing the unpurified protected inputs produces varying degrees of subject blur, texture artifacts, or subject–background entanglement. Different purification baselines alleviate some of these effects, but their behavior varies across protection settings. In this example, RSDR better preserves the appearance of the cat while generating a hospital scene consistent with the target prompt.

Table 2: Editability restoration under DiffusionGuard with four held-out masks. Overall metrics are computed over all 1,680 editing instances, with Recovery also reported per mask. CLIP-Dir is scaled by 100. Best and second-best purification results are **bold** and underlined, respectively.

	Overall editing metrics				Recovery by downstream mask				
Method	PSNR↑	LPIPS↓	DINO↑	CLIP-Dir↑	C-brush	Ellipse	Free.	Rect.	Overall↑
Clean edit ref.	20.3013	0.1528	0.9162	15.6355	–	–	–	–	1.0000
Protected edit ref.	18.9988	0.2870	0.8008	14.9204	–	–	–	–	0.0000
JPEG	19.2427	0.2656	0.8317	15.0792	0.2950	0.2004	0.1780	0.1564	0.2091
GrIDPure	19.0818	0.2104	0.8319	14.7706	0.1572	0.2762	0.2187	-0.1392	0.1736
IMPRESS	19.2146	0.2759	0.8140	14.8759	0.0540	0.1023	0.0789	0.0535	0.0751
PDM-Pure	19.5080	0.2285	0.7886	15.1011	0.1719	0.2903	0.2850	0.1065	0.2435
RSDR	**19.9568**	**0.2041**	**0.8944**	**15.3538**	**0.7857**	**0.6580**	**0.6614**	**0.6480**	**0.6926**



(a) Editing setup **(b)** Downstream edited outputs **(c)** Zoomed-in comparison

Figure 5: Qualitative comparison under DiffusionGuard with a held-out downstream mask. (a) Editing setup and clean-edit reference. (b) Downstream edits after different purification methods. (c) Corresponding facial crops. The protective perturbation is optimized using a SAM subject mask, while the downstream freehand mask is unseen during protection.

Table 3: Input fidelity, no-reference image quality, and downstream editability under DiffusionGuard. Protected serves as a reference and is excluded from ranking. Best and second-best purification results are **bold** and underlined, respectively.

	Reference-based fidelity				No-reference quality	Editability
Method	PSNR↑	SSIM↑	LPIPS↓	DINO↑	BRISQUE↓	Recovery↑
Protected	38.4897	0.9647	0.0783	0.9923	14.0679	0.0000
JPEG	**36.1899**	0.9417	0.1395	0.9893	21.6964	0.2488
GrIDPure	27.0981	0.8800	0.2449	0.9550	52.4078	0.3534
IMPRESS	36.0229	**0.9548**	**0.0898**	**0.9907**	14.6188	0.1274
PDM-Pure	28.0540	0.8418	0.3025	0.9090	55.3825	0.2995
RSDR(ours)	30.8798	0.7569	0.3225	0.9756	**12.6541**	**0.7313**

5.3 Editability Restoration Across Held-Out Downstream Masks

To evaluate editability restoration under different downstream editing regions, we retain the DiffusionGuard-protected images and their purified outputs while replacing the original SAM subject mask with four held-out masks from InpaintGuardBench: circle brush, ellipse, freehand, and rectangle. These masks are not used during perturbation generation, yielding 1,680 editing instances per purification method. As shown in Table 2, RSDR achieves an overall Recovery of 0.6926, outperforming the strongest baseline, PDM-Pure (0.2435), by 0.4491. It also achieves the best PSNR, LPIPS, DINO, and CLIP-Dir, along with the highest Recovery across all four masks. These results demonstrate consistent improvements in editability restoration across the evaluated downstream editing regions. Figure 5 further shows that RSDR better preserves subject structure and local facial details compared with existing purification methods.

5.4 Input Fidelity and Editability Restoration

Generative purification must preserve image content while restoring downstream editability. We examine their relationship using PSNR, SSIM, LPIPS, and DINO to measure input fidelity, together with BRISQUE (Mittal et al., 2012) for no-reference image quality. Full results appear in Appendix D; we use DiffusionGuard as a representative case.

As shown in Table 3, the unpurified protected input remains highly similar to the clean image but has zero Recovery. IMPRESS preserves high input fidelity yet achieves limited editability restoration. In contrast, RSDR achieves the highest Recovery and lowest BRISQUE despite lower pixel and structural fidelity. Across 20 protection–purification combinations, the Spearman correlations between Recovery and PSNR, SSIM, $-$LPIPS, DINO, and $-$BRISQUE are -0.185, -0.242, -0.036, 0.060, and 0.009, respectively. Figure 6 illustrates the weak association between DINO similarity and Recovery. These results indicate that input fidelity alone is insufficient to assess editability restoration, motivating their use as complementary evaluation dimensions.

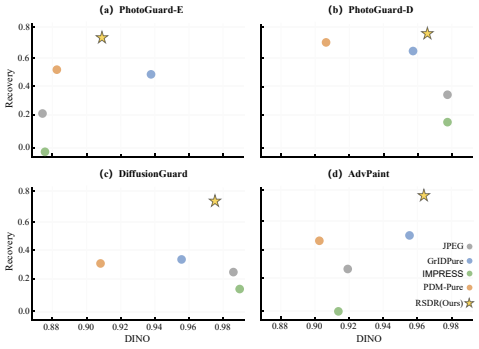


Figure 6: Input fidelity (DINO) versus editability restoration (Recovery) across four protection settings. Stars denote RSDR. The pooled Spearman correlation is $\rho = 0.060$.

Table 4: Input-role configurations and macro-averaged editing metrics across four protection settings. $\epsilon \sim \mathcal{N}(0, I)$ denotes independent Gaussian initialization, and 0 denotes a zero reference. Group D is RSDR. CLIP-Dir is scaled by 100. Recovery uses clean- and protected-edit references.

Group	State	Reference	PSNR↑	LPIPS↓	DINO↑	CLIP-Dir↑	Recovery↑
A	ϵ	0	10.2983	0.5881	0.2140	**16.8245**	-0.9804
B	x^{pro}	0	22.2918	0.3778	0.6839	13.6829	0.0140
C	x^{pro}	x^{pro}	22.5119	0.3647	0.7114	13.7792	0.0666
D (RSDR)	ϵ	x^{pro}	**26.2649**	**0.1881**	**0.9402**	15.5174	**0.7549**

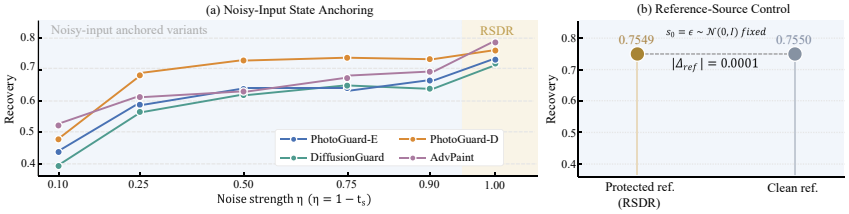


Figure 7: Ablation of reference–state decoupling. (a) Recovery across four protection settings with input anchoring, where $\eta = 1$ denotes independent initialization. (b) Recovery with protected versus clean references under independent initialization, averaged across four protection settings.

5.5 Ablation Studies

(1) Input-role configurations. Table 4 compares four configurations: independent initialization without reference (***A***), protected-image initialization without (***B***) or with reference (***C***), and RSDR (***D***). ***A*** achieves high CLIP-Dir but poor subject preservation, while ***B*** and ***C*** yield limited Recovery. RSDR achieves a macro-averaged Recovery of 0.7549, demonstrating the benefit of independent initialization with protected-reference conditioning. **(2) Noisy-input state anchoring.** Unlike the diagnostic study in Section 3.3, which trains separate models for different initialization settings, we fix the trained RSDR model and vary the protected image's contribution to the initial state only at inference. Figure 7(a) shows that Recovery generally increases as input anchoring weakens, with independent Gaussian initialization achieving the highest Recovery across all four protection settings. **(3) Reference-source control.** With the trained RSDR model and independent Gaussian initialization fixed, we compare protected and clean references. As shown in Figure 7(b), their macro-averaged Recovery scores are 0.7549 and 0.7550, respectively, indicating that replacing the protected reference with its clean counterpart provides negligible improvement under the evaluated protocol. See Appendix B for full ablation results.

6 Conclusion

In this paper, we investigate how the input roles of protected images affect editability restoration and propose RSDR to separate reference conditioning from state initialization. Extensive experiments demonstrate that RSDR consistently improves editability restoration while preserving image content across different protection methods and editing settings. Future work will explore extending this framework to diverse generative architectures and more challenging editing scenarios.

AI USE STATEMENT

Generative AI tools were used to assist with language polishing, manuscript organization, and drafting. All technical content, experimental results, and scientific conclusions were reviewed and verified by the authors, who take full responsibility for the final manuscript.

REFERENCES

Tim Brooks, Aleksander Holynski, and Alexei A. Efros. InstructPix2Pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18392–18402, 2023.

Bochuan Cao, Changjiang Li, Ting Wang, Jinyuan Jia, Bo Li, and Jinghui Chen. IMPRESS: Evaluating the resilience of imperceptible perturbations against unauthorized data usage in diffusion-based generative AI. In *Advances in Neural Information Processing Systems*, volume 36, pp. 10657–10677, 2023.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9650–9660, 2021.

Ruoxi Chen, Haibo Jin, Yixin Liu, Jinyin Chen, Haohan Wang, and Lichao Sun. EditShield: Protecting unauthorized image editing by instruction-guided diffusion models. In *European Conference on Computer Vision*, pp. 126–142, 2024.

William June Suk Choi, Kyungmin Lee, Jongheon Jeong, Saining Xie, Jinwoo Shin, and Kimin Lee. DiffusionGuard: A robust defense against malicious diffusion-based image editing. In *International Conference on Learning Representations*, 2025.

Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. DiffEdit: Diffusion-based semantic image editing with mask guidance. In *International Conference on Learning Representations*, 2023.

Rinon Gal, Or Patashnik, Haggai Maron, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. StyleGAN-NADA: CLIP-guided domain adaptation of image generators. *ACM Transactions on Graphics*, 41(4):141:1–141:13, 2022. doi: 10.1145/3528223.3530164.

Joonsung Jeon, Woo Jae Kim, Suhyeon Ha, Sooel Son, and Sung-eui Yoon. AdvPaint: Protecting images from inpainting manipulation via adversarial attention disruption. In *International Conference on Learning Representations*, 2025.

Yan Ju, Hongfei Xue, and Siwei Lyu. FreqPure: A high-frequency preservation diffusion-based purification method for protective perturbation removal. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 1544–1553, 2025.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4401–4410, 2019.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4015–4026, 2023.

Xinqi Lin, Jingwen He, Ziyan Chen, Zhaoyang Lyu, Bo Dai, Fanghua Yu, Yu Qiao, Wanli Ouyang, and Chao Dong. DiffBIR: Toward blind image restoration with generative diffusion prior. In *Computer Vision – ECCV 2024*, pp. 430–448. Springer, 2024. doi: 10.1007/978-3-031-73202-7_25.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations*, 2023.

Ling Lo, Cheng Yu Yeo, Hong-Han Shuai, and Wen-Huang Cheng. Distraction is all you need: Memory-efficient image immunization against diffusion-based image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24462–24471, 2024.

Nanye Ma, Mark Goldstein, Michael S. Albergo, Nicholas M. Boffi, Eric Vanden-Eijnden, and Saining Xie. SiT: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In *European Conference on Computer Vision*, pp. 23–40, 2024.

Zehong Ma, Longhui Wei, Shuai Wang, Shiliang Zhang, and Qi Tian. DeCo: Frequency-decoupled pixel diffusion for end-to-end image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 43600–43610, 2026.

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.

Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21(12):4695–4708, 2012. doi: 10.1109/TIP.2012.2214050.

Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Animashree Anandkumar. Diffusion models for adversarial purification. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pp. 16805–16827. PMLR, 2022.

Tarik Can Ozden, Ozgur Kara, Oguzhan Akcin, Kerem Zaman, Shashank Srivastava, Sandeep P. Chinchali, and James M. Rehg. DiffVax: Optimization-free image immunization against diffusion-based editing. In *International Conference on Learning Representations*, 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 8748–8763. PMLR, 2021.

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*, pp. 10684–10695, 2022.

Hadi Salman, Alaa Khaddaj, Guillaume Leclerc, Andrew Ilyas, and Aleksander Madry. Raising the cost of malicious AI-powered image editing. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 29894–29918. PMLR, 2023.

Su Wang, Chitwan Saharia, Ceslee Montgomery, Jordi Pont-Tuset, Shai Noy, Stefano Pellegrini, Yasumasa Onoe, Sarah Laszlo, David J. Fleet, Radu Soricut, Jason Baldridge, Mohammad Norouzi, Peter Anderson, and William Chan. Imagen Editor and EditBench: Advancing and evaluating text-guided image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18359–18369, 2023.

Haotian Xue and Yongxin Chen. Pixel is a barrier: Diffusion models are more adversarially robust than we think. *arXiv preprint arXiv:2404.13320*, 2024.

Wenkui Yang, Jie Cao, Junxian Duan, and Ran He. Towards robust defense against customization via protective perturbation resistant to diffusion-based purification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19290–19300, 2025.

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*, pp. 586–595, 2018.

Qichen Zhao, Shengfang Zhai, Xinjian Bai, Qingni Shen, Qiqi Lin, Yansong Gao, and Zhonghai Wu. Purify once, edit freely: Breaking image protections under model mismatch. *arXiv preprint arXiv:2603.13028*, 2026.

Zhengyue Zhao, Jinhao Duan, Kaidi Xu, Chenan Wang, Rui Zhang, Zidong Du, Qi Guo, and Xing Hu. Can protective perturbation safeguard personal data from being exploited by Stable Diffusion? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24398–24407, 2024.