# RETHINKING IMAGE IMMUNIZATION FOR TRANSFERABLE AND PURIFICATION-ROBUST PROTECTION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Image immunization has been actively studied to protect images from unauthorized diffusion-based manipulation. For a practical deployment, however, image protection requires two key properties: black-box transferability to unknown generative models and robustness against perturbation purification, which have largely been studied along separate directions. In this paper, we present a unified immunization framework that jointly targets both requirements. Our controlled analysis identifies early VAE encoder representations as a promising common attack surface: they exhibit strong cross-model correspondence, and perturbations optimized at these stages better retain their induced latent shifts after purification. Based on these observations, we directly disrupt early VAE encoder representations, rather than relying on architecture-specific denoisers or explicit purification-aware designs. We further introduce a relation-disruption objective that disrupts the internal cross-channel structure of early representations. Extensive experiments demonstrate strong black-box transferability across heterogeneous diffusion-based generative models within multiple downstream manipulation tasks, and robustness against diverse purification techniques, with low computational overhead.

## 1 INTRODUCTION

Recent advances in diffusion models have enabled flexible and realistic image manipulation, while simultaneously raising serious risks of unauthorized manipulation and synthetic misinformation. To counter this, image immunization (Salman et al., 2023; Song et al., 2026; Bala et al., 2025) has emerged as a proactive defense against such threats, introducing imperceptible perturbations into private images to disrupt subsequent diffusion-based manipulation (Rombach et al., 2022; HaCohen et al., 2024; Labs, 2024).

For a practical deployment, image immunization must satisfy two key requirements: *black-box transferability* and *robustness to perturbation purification*. Transferability is particularly challenging because the target model is unknown and may differ substantially from the surrogate in model architecture (*e.g.*, U-Net and DiT), image-conditioning mechanism (*e.g.*, latent concatenation and visual-token conditioning), or downstream manipulation modality (*e.g.*, I2I and I2V). Most existing methods (Choi et al., 2025; Ozden et al., 2026; Shao et al., 2026; Wang et al., 2026) consider cross-model transferability, but architectural heterogeneity itself is rarely treated as an explicit design objective; transfer across substantially different architectures is often evaluated post hoc.

Purification poses a complementary challenge: common image processing (*e.g.*, JPEG compression (Guo et al., 2018)) and dedicated purification methods (Zhao et al., 2024; Diaz, 2025) can suppress protective perturbations before manipulation. Recent works (Kim et al., 2025; Bala et al., 2025; Yang et al., 2025b) have therefore explored either transferable protection or purification-robust perturbation design, but these two properties have largely been pursued as separate objectives. Together, these challenges suggest the need for a protection target that can remain effective across heterogeneous models while surviving purification.

Motivated by these challenges, we ask a central question: *Can the VAE representation hierarchy provide a common attack surface for both cross-model transferability and robustness to purification?* To isolate the effect of representation depth, we apply the same feature-discrepancy objective (Chen et al., 2024) at successive stages of the diffusion pipeline. Under this controlled setting, we find that



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.

targeting earlier VAE encoder stages yields stronger black-box transferability and better preserves protection after purification than targeting deeper representations. To understand this behavior, we analyze the representations from two complementary perspectives: cross-model correspondence and the survival of attack-induced effects under purification. Our analysis reveals that early VAE features exhibit stronger cross-model correspondence, while perturbations optimized at these stages better retain their induced effects after purification. Together, these observations motivate our use of early-stage VAE representations as a common attack surface for transferable and purification-robust immunization.

Building on this finding, we propose a unified image-immunization framework that jointly targets black-box transferability and purification robustness. Our framework directly disrupts the identified early-stage VAE representations, avoiding reliance on the final VAE latent (Chen et al., 2024; Wang et al., 2026; Shao et al., 2026) or diffusion denoiser (Jeon et al., 2025; Choi et al., 2025). To further strengthen the disruption of the selected early-stage representation, we introduce a *relation-disruption objective* that complements direct feature displacement by perturbing its internal cross-channel organization. Empirically, this complementary objective further improves transfer to unseen encoders.

We validate this design across heterogeneous generative models and manipulation pipelines, including image-to-image editing, inpainting, instruction-guided editing, and image-to-video generation. The evaluation spans both U-Net- and DiT-based diffusion models and multiple purification techniques. Across these settings, our method demonstrates strong black-box transferability while maintaining robust protection after purification. Furthermore, because optimization is confined to the early stages of the surrogate VAE encoder, our method maintains low computational overhead.

The main contributions can be summarized as follows:

- Through a controlled depth-wise study with a common discrepancy objective, we show that targeting earlier VAE encoder stages yields stronger black-box transferability and better retention after purification than targeting deeper representations, motivating early-stage features as a common attack surface.
- Based on this observation, we propose a simple yet effective image-immunization framework that directly disrupts early-stage VAE features and introduces a relation-disruption objective that perturbs the internal cross-channel organization of early representations.
- We demonstrate strong transferable and purification-robust protection across heterogeneous diffusion models, manipulation tasks, and purification techniques, while requiring only lightweight VAE-encoder optimization.

## 2 Related Work

**Image Immunization against Diffusion Models.** Practical image immunization requires *protection effectiveness*, *purification robustness*, and *transferability* across manipulation tasks (*e.g.*, I2I→I2V) and model architectures (*e.g.*, U-Net→DiT). However, prior works have largely addressed these objectives separately. Early approaches, including PhotoGuard (Salman et al., 2023), AdvDM (Liang et al., 2023), Glaze (Shan et al., 2023), and Mist (Liang & Wu, 2023), mainly considered general image editing or mimicry, while subsequent methods extended protection to instruction-guided editing (Chen et al., 2024), inpainting (Jeon et al., 2025; Choi et al., 2025), and image-to-video generation (Gui et al., 2025; Vu et al., 2026; Song et al., 2026). Robustness-oriented methods such as DCT-Shield (Bala et al., 2025) and BlurGuard (Kim et al., 2025) improve purification resistance through frequency-aware perturbation designs. Transferability-oriented methods instead pursue more model- and task-agnostic protection, including UniDef (Shao et al., 2026) and FaceArmor (Wang et al., 2026). To bridge these largely separate directions, we propose a unified framework that jointly targets purification robustness and transfer across manipulation tasks and U-Net/DiT architectures.

**Adversarial Attacks on Intermediate Features.** Traditional adversarial attacks on image classifiers (Goodfellow et al., 2015; Moosavi-Dezfooli et al., 2016; Dong et al., 2018; Lee et al., 2025; Liu et al., 2023) increasingly exploit intermediate representations to improve transferability. Methods such as Feature Disruptive Attack (FDA) (Ganeshan et al., 2019), Feature Importance-Aware Attack (FIA) (Wang et al., 2021), and Neuron Attribution-based Attack (NAA) (Zhang et al., 2022) manipulate intermediate features or important neurons, while later works including Semantics-aware

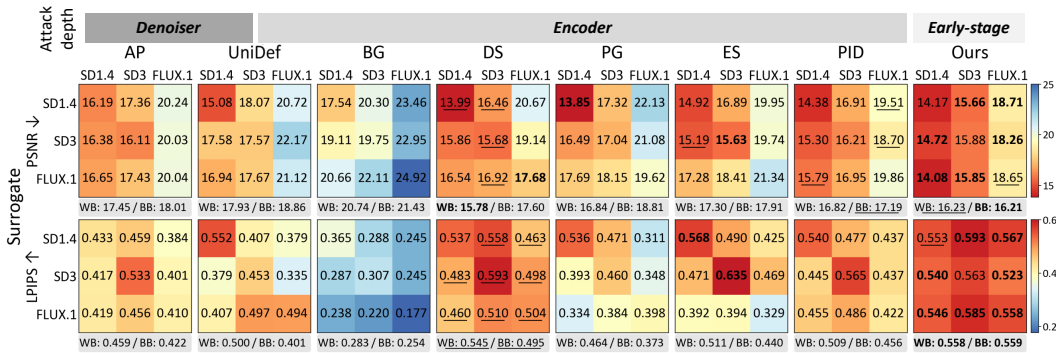



Figure 1: Transferability comparison on image-to-image task. Rows denote the surrogate model used for optimization, columns denote the target model used for evaluation, and diagonal and off-diagonal entries represent white- and black-box settings, respectively.

Multi-granularity Patchout (SMP) Yang et al. (2025a), Pixel2Feature Attack (P2FA) (Liu et al., 2025), and Low-Rank and Sparse Decomposition (LRS) (Pu et al.) further explore transferable feature structures across different depths. More recently, diffusion-model immunization methods have also utilized intermediate VAE encoder features (Ahn et al., 2025; Vu et al., 2026), but still rely primarily on the VAE output or latent representation.

## 3 Analysis of the Attack Surface

In this section, we first examine whether intermediate VAE representations provide a more suitable attack surface than downstream representations such as the final VAE latent and denoiser features. We then analyze the depth-wise behavior within the VAE encoder to better understand the observed transferability and purification-robustness trends.

### 3.1 Motivation for Revisiting the Attack Surface

We revisit the attack surface of image immunization from two practical perspectives: *cross-model transferability* and *robustness to purification*.

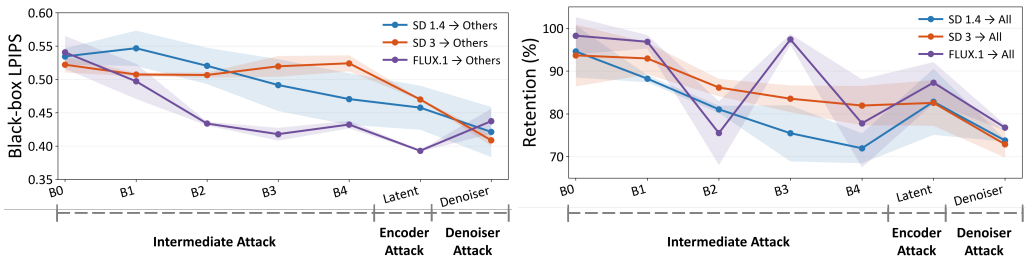
**Transferability.** Existing image-immunization methods optimize representations at different depths of the generative pipeline, ranging from final VAE representations to diffusion-denoiser features. However, the effect of attack depth on cross-model transferability remains unclear. In particular, even methods operating on the final VAE representation, such as PG, DS, and ES, exhibit substantial gaps between white-box and transferred protection, as shown in Fig. 1. This observation suggests that moving the attack target upstream of the denoiser does not by itself guarantee cross-model transfer, motivating us to examine whether representations at different depths of the VAE encoder provide more transferable attack surfaces.

**Purification Robustness.** Attack depth may also affect robustness to purification. Since purification modifies the input before encoding, the perturbation-induced effect observed at a given representation depends on how the purified input is transformed through the preceding encoder layers. As the target moves deeper in the encoder, this effect is shaped by a longer sequence of transformations, potentially making it more sensitive to purification. This motivates us to examine how purification robustness varies with representation depth.

### 3.2 Controlled Comparison of Attack Surfaces

To systematically examine the role of attack depth in both transferability and purification robustness, we conduct a controlled comparison across successive stages of the diffusion pipeline. Following the established use of discrepancy maximization between clean and immunized representations (Chen et al., 2024; Jeon et al., 2025), we isolate the effect of representation depth by applying the same objective to intermediate VAE encoder features, the final VAE latent, and denoiser Q/K/V features. We use the following representation-discrepancy objective:

$$\mathcal{L}_{\mathrm{Dis}}^{(k)} = -\frac{1}{|\mathcal{I}_k|} \sum_{i \in \mathcal{I}_k} \frac{\left\| \phi_k^{(i)}(x+\delta) - \phi_k^{(i)}(x) \right\|_2}{\left\| \phi_k^{(i)}(x) \right\|_2 + \epsilon}, \tag{1}$$




(a) Black-box transferability across attack surfaces. (b) Performance retention after purification.

Figure 2: Controlled comparison of attack surfaces across the diffusion pipeline. (a) Black-box LPIPS obtained by independently targeting intermediate VAE representations, the final VAE latent, and denoiser representations. (b) Fraction of protection retained after purification relative to the corresponding pre-purification performance. *SD 1.4 → Others* denotes SD 1.4 as the surrogate VAE, with results averaged over the other VAEs.

where $k$ indexes the representation depth along the diffusion pipeline, $\mathcal{I}_k$ the set of feature units selected at that depth, $x$ is the source image, $\delta$ represents the adversarial perturbation, and $\phi_k(\cdot)$ denotes the corresponding feature. For each $k$, we optimize only the representation at that depth, rather than using intermediate features as auxiliary supervision for a deeper objective (Ahn et al., 2025; Vu et al., 2026). This controlled setup allows us to compare attack surfaces while keeping the optimization objective fixed.

We conduct image-to-image editing experiments using SD1.4 (Rombach et al., 2022), SD3 (Esser et al., 2024), and FLUX.1 (Labs, 2024), with each model serving as the surrogate in turn, and denote an image optimized at depth $k$ of surrogate $s$ by $x^{(s)}_{\text{adv},k}$. For transferability, we evaluate each immunized image on the other two models and report LPIPS (Zhang et al., 2018) in Fig. 2a. For purification robustness, we apply Adverse Cleaner (Diaz, 2025) and GridPure (Zhao et al., 2024) and report the post-purification performance relative to the corresponding pre-purification performance across all three models in Fig. 2b.

Figs. 2a and 2b jointly show that intermediate VAE encoder representations generally provide stronger black-box transferability and greater protection retention after purification than the final VAE latent or denoiser representations, despite non-monotonic variations across individual depths. We also observe a similar overall trend with cosine and normalized-$\ell_2$ discrepancy objectives (Appendix C.1), suggesting that this behavior is not specific to the discrepancy formulation in Eq. (1). These results motivate a finer-grained analysis of the VAE encoder hierarchy to understand how these favorable properties vary across intermediate depths. Importantly, this trend does not imply a universal transferability advantage across loss function designs, but suggests the potential of earlier representations as a favorable surface for black-box transfer.

### 3.3 Depth-wise Analysis of Intermediate VAE Representations

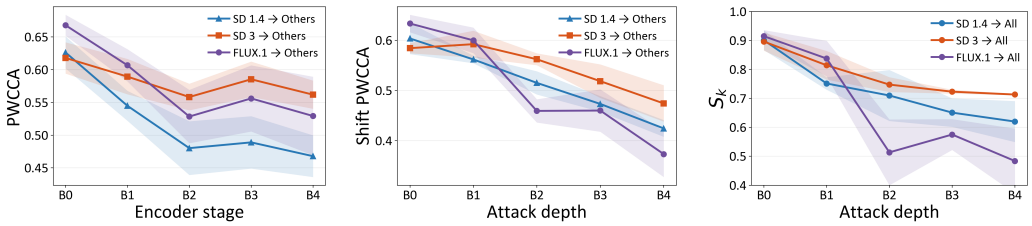
Having identified intermediate VAE representations as a promising attack surface, we next examine how their properties vary across encoder depth. Specifically, we analyze cross-VAE correspondence of both clean representations and perturbation-induced shifts, together with the survival of the induced latent shift after purification.

**Transferability across Encoder Depth.** We first examine how consistently intermediate representations are organized across different VAE encoders. To this end, we use PWCCA (Morcos et al., 2018), which measures correspondence between representation spaces through correlated linear combinations of feature dimensions without requiring one-to-one channel alignment.

Let $\mathcal{X}$ denote the evaluation set and $H^{(m)}_k(\mathcal{X})$ the representation matrix collected at the encoder block $k$ from model $m \in \{s, t\}$. We first measure *clean*-PWCCA as:

$$\mathcal{W}^{s\to t}_{k,\text{clean}} = \text{PWCCA}^{s\to t}\left(H^{(s)}_k(\mathcal{X}), H^{(t)}_k(\mathcal{X})\right). \tag{2}$$

A larger $\mathcal{W}^{s\to t}_{k,\text{clean}}$ indicates stronger correspondence between the representation spaces of the surrogate and target encoders at depth $k$.




(a) Cross-VAE clean-PWCCA. (b) Cross-VAE shift-PWCCA. (c) Latent shift survival.

Figure 3: Depth-wise analysis of intermediate VAE attack targets. (a) Cross-VAE representation correspondence measured by *clean*-PWCCA (Eq. (2)). (b) Cross-VAE correspondence of perturbation-induced representation shifts measured by *shift*-PWCCA (Eq. (3)). (c) Latent shift survival ($S_k$) after purification. *SD 1.4 → Others* and *SD 1.4 → All* denote SD 1.4 as the surrogate VAE, with results averaged over the other and all VAEs, respectively.

Clean representation correspondence alone, however, does not establish that an attack-induced change in the surrogate is reproduced in the target encoder. We therefore compute shift-PWCCA to assess the cross-encoder correspondence of perturbation-induced representation shifts. Let $\mathcal{X}_{\text{adv}}^{(s,k)}$ denote the immunized images optimized using surrogate $(s)$ at $k$-depth with Eq. (1). For each encoder $m$, we define *shift*-PWCCA as:

$$\mathcal{W}_{k,\Delta}^{s\to t} = \text{PWCCA}^{s\to t}\left(\Delta H_k^{(s)}, \Delta H_k^{(t)}\right), \quad \text{where } \Delta H_k^{(m)} = H_k^{(m)}\left(\mathcal{X}_{\text{adv}}^{(s,k)}\right) - H_k^{(m)}(\mathcal{X}). \tag{3}$$

A larger $\mathcal{W}_{k,\Delta}^{s\to t}$ indicates stronger correspondence between the feature changes induced in the surrogate and target encoders. Thus, Eq. (2) characterizes correspondence between the underlying representation spaces, whereas (3) measures how consistently the attack-induced changes are reflected across encoders.

We compute both measures at corresponding depths among the VAE encoders of SD1.4 (Rombach et al., 2022), SD3 (Esser et al., 2024), and FLUX.1 (Labs, 2024). As shown in Fig. 3a and Fig. 3b, both *clean*-PWCCA and *shift*-PWCCA are highest at early intermediate stages and generally decrease with encoder depth. Together with the black-box results in Fig. 2a, these consistent trends support early intermediate representations as favorable attack targets for cross-VAE transfer. Details of the PWCCA (Morcos et al., 2018) are provided in Sec. A.5.

**Purification Robustness across Encoder Depth.** We next examine how the effect induced at each intermediate depth survives input purification. Although the perturbation is optimized at an intermediate representation, its protection ultimately depends on how the resulting change propagates to the final VAE latent used by the downstream generative model.

Let $x_{\text{adv},k}^{(s)}$ denote an immunized image optimized at encoder depth $k$ of surrogate $s$. For an evaluation encoder $t$, we define *latent shift survival* as:

$$S_k^{P,(s\to t)} = \mathbb{E}_x\left[\frac{\left\langle E^{(t)}\left(P(x_{\text{adv},k}^{(s)})\right) - E^{(t)}(P(x)),\ E^{(t)}\left(x_{\text{adv},k}^{(s)}\right) - E^{(t)}(x)\right\rangle}{\left\|E^{(t)}\left(x_{\text{adv},k}^{(s)}\right) - E^{(t)}(x)\right\|_2^2 + \epsilon}\right], \tag{4}$$

where $E^{(t)}(\cdot)$ denotes the final latent of the evaluation encoder $t$, $P(\cdot)$ is a purification operator, and $\epsilon$ is a small constant for numerical stability. Applying $P(\cdot)$ to both the clean and immunized inputs controls for the latent change introduced by purification itself, allowing the metric to isolate the surviving perturbation-induced effect. The normalized projection captures both directional alignment and retained magnitude, with larger values indicating stronger survival.

We independently generate $x_{\text{adv},k}^{(s)}$ at each intermediate depth using Eq. (1) and compute $S_k^{P,(s\to t)}$ after JPEG compression (Guo et al., 2018) and GridPure (Zhao et al., 2024) across all surrogate–target encoder pairs. As shown in Fig. 3c, the result shows that latent shift survival is highest at early intermediate stages across both purification methods and evaluation encoders. Consistent with the protection retention in Fig. 2b, this supports early intermediate representations as favorable attack targets for purification-robust protection.

# 4 METHOD

Based on the above analysis, our method addresses two key design questions: *where to perturb* and *how to perturb*. We target an early VAE encoder representation and perturb it through two complementary objectives: *representation discrepancy* and *relational disruption*.

**Attack Target.** Motivated by the depth-wise analysis in Secs. 3.2 and 3.3, we focus on the early portion of the VAE encoder. We adopt the earliest block $k^\star$ (B0) as the optimization target and validate this choice under the full objective in Sec. 5.4.

**Representation-discrepancy Objective.** We use the representation-discrepancy objective defined in Eq. (1) at $k^\star$ and denote it by $\mathcal{L}_{\mathrm{Dis}} \equiv \mathcal{L}_{\mathrm{Dis}}^{(k^\star)}$. Minimizing $\mathcal{L}_{\mathrm{Dis}}$ therefore drives the immunized representation away from its clean counterpart, directly disrupting the selected representation.

**Relation-disruption Objective.** While minimizing $\mathcal{L}_{\mathrm{Dis}}$ maximizes the distance between clean and immunized representations, a large feature displacement does not necessarily imply a substantial change in how channels interact. For example, changes dominated by channel activation magnitudes can yield a large representation discrepancy while largely preserving the pairwise relations among channel activation maps. We therefore introduce a complementary relation-disruption objective, $\mathcal{L}_{\mathrm{Rel}}$, that explicitly perturbs these pairwise relations.

To focus this relational disruption on the feature transformation introduced by the selected encoder stage, we apply $\mathcal{L}_{\mathrm{Rel}}$ to the residual branches of its residual units, while $\mathcal{L}_{\mathrm{Dis}}$ operates on their output features. Specifically, for each residual unit $i \in \mathcal{I}_{k^\star}$ at the selected block $k^*$, let $s_i(x)$ and $a_i(x)$ denote the shortcut and residual-branch activations, respectively, with $h_i(x) = s_i(x) + a_i(x)$. We instantiate $\phi_{k^\star}^{(i)}(x) = h_i(x)$ in Eq. (1) and apply $\mathcal{L}_{\mathrm{Rel}}$ to $a_i(x)$, focusing relational disruption on the transformation introduced by the residual branch rather than the shortcut pathway. We also validate this design choice against the shortcut and post-addition activations in Appendix C.1.

For each $i \in \mathcal{I}_{k^\star}$, let $a_{i,p}(x) \in \mathbb{R}^{HW}$ denote the flattened spatial activation of channel $p$, and let $\hat{a}_{i,p}(x) = a_{i,p}(x)/\|a_{i,p}(x)\|_2$ denote its normalized form. We construct a channel-relation matrix $C^{(i)}(x) \in \mathbb{R}^{N\times N}$, where $N$ is the number of channels, and we define the relation-disruption loss as:

$$\mathcal{L}_{\mathrm{Rel}} = -\frac{1}{|\mathcal{I}_{k^*}|}\sum_{i\in\mathcal{I}_{k^*}} \frac{\left\|C^{(i)}(x_{\mathrm{adv}}) - C^{(i)}(x)\right\|_F}{\left\|C^{(i)}(x)\right\|_F}, \qquad C^{(i)}(x)_{pq} = \hat{a}_{i,p}(x)^\top \hat{a}_{i,q}(x). \tag{5}$$

Because each channel activation is normalized, $C^{(i)}(x)$ captures pairwise angular relations independently of activation magnitude. Thus, $\mathcal{L}_{\mathrm{Rel}}$ explicitly perturbs pairwise cross-channel relations, complementing the direct feature displacement induced by $\mathcal{L}_{\mathrm{Dis}}$. Importantly, $\mathcal{L}_{\mathrm{Rel}}$ is defined entirely within the surrogate representation and does not require cross-model channel correspondence; we do not assume that the relation matrices themselves are shared across encoders. Empirically, we observe that this relational objective $\mathcal{L}_{\mathrm{Rel}}$ improves average black-box transferability over $\mathcal{L}_{\mathrm{Dis}}$ alone, and their combination provides a further gain (Sec. 5.4).

**Overall Objective.** We combine $\mathcal{L}_{\mathrm{Dis}}$ and $\mathcal{L}_{\mathrm{Rel}}$ to disrupt both the selected representation and its cross-channel organization:

$$\delta^* = \arg\min_{\delta} \left[\mathcal{L}_{\mathrm{Dis}} + \lambda \mathcal{L}_{\mathrm{Rel}}\right], \qquad \text{s.t.} \quad \|\delta\|_\infty \le \eta_\infty, \tag{6}$$

where $\lambda$ controls the contribution of $\mathcal{L}_{\mathrm{Rel}}$, and $\eta_\infty$ denotes the perturbation budget $\ell_\infty$. We denote the constrained objective in Eq. (6) by $\mathcal{L}_{\mathrm{overall}}$.

**Optimization.** Starting from $x_{\mathrm{adv}}^{(0)} = x + \delta$, where $\delta \sim \mathcal{U}(-1/255, 1/255)$, we optimize the immunized image using projected gradient descent (Madry et al., 2018):

$$x_{\mathrm{adv}}^{(j+1)} = \Pi_{\mathcal{B}_{\ell_\infty}(x,\eta_\infty)}\left[x_{\mathrm{adv}}^{(j)} - \gamma\ \mathrm{sign}\left(\nabla_{x_{\mathrm{adv}}^{(j)}} \mathcal{L}_{\mathrm{overall}}\right)\right], \tag{7}$$

where $\gamma$ denotes the step size. Restricting both objectives to early encoder stages avoids backpropagation through the remaining encoder and denoiser components, thereby reducing computational cost compared with optimizing deeper representations. The overall algorithm is provided in Appendix B.

Table 1: Inpainting results. Lower PSNR/CLIP and higher LPIPS indicate stronger protection.

| Target | Surrogate | Metric | AP | UniDef | BG | DS | PG | ES | PID | Ours |
|---|---|---|---|---|---|---|---|---|---|---|
| SD-Inpaint | SD1.4 | PSNR ↓ | 17.68 | 18.32 | 22.81 | 17.34 | 18.11 | 17.02 | 17.47 | **16.95** |
| | | CLIP ↓ | **0.780** | 0.792 | 0.831 | 0.786 | 0.789 | 0.787 | 0.786 | 0.801 |
| | | LPIPS ↑ | 0.453 | 0.424 | 0.193 | 0.473 | 0.418 | 0.499 | 0.452 | **0.543** |
| | SD3 | PSNR ↓ | 18.17 | 20.11 | 22.90 | 17.86 | 19.37 | **17.02** | 18.32 | 17.36 |
| | | CLIP ↓ | 0.801 | 0.819 | 0.824 | 0.807 | 0.825 | 0.782 | **0.779** | 0.807 |
| | | LPIPS ↑ | 0.422 | 0.336 | 0.192 | 0.487 | 0.350 | **0.501** | 0.439 | **0.501** |
| | FLUX.1 | PSNR ↓ | 18.95 | 19.24 | 26.50 | 18.43 | 20.51 | 20.23 | 18.52 | **16.97** |
| | | CLIP ↓ | **0.803** | 0.823 | 0.831 | 0.817 | 0.836 | 0.826 | 0.809 | 0.805 |
| | | LPIPS ↑ | 0.386 | 0.382 | 0.112 | 0.454 | 0.296 | 0.334 | 0.420 | **0.536** |
| FLUX-Fill | SD1.4 | PSNR ↓ | 22.18 | 22.43 | 25.81 | 22.24 | 23.92 | 21.85 | 22.76 | **20.49** |
| | | CLIP ↓ | 0.815 | **0.801** | 0.863 | 0.827 | 0.844 | 0.804 | 0.823 | 0.807 |
| | | LPIPS ↑ | 0.406 | 0.371 | 0.193 | 0.451 | 0.306 | 0.446 | 0.398 | **0.556** |
| | SD3 | PSNR ↓ | 21.62 | 24.19 | 25.77 | 21.24 | 22.91 | 20.92 | 22.35 | **20.79** |
| | | CLIP ↓ | 0.827 | 0.850 | 0.855 | 0.814 | 0.840 | 0.811 | 0.828 | **0.808** |
| | | LPIPS ↑ | 0.422 | 0.315 | 0.180 | 0.507 | 0.328 | 0.492 | 0.423 | **0.529** |
| | FLUX.1 | PSNR ↓ | 21.92 | 23.56 | 28.39 | 22.27 | 23.26 | 23.29 | 22.67 | **20.29** |
| | | CLIP ↓ | 0.824 | 0.847 | 0.866 | 0.831 | 0.843 | 0.847 | 0.828 | **0.808** |
| | | LPIPS ↑ | 0.409 | 0.352 | 0.123 | 0.450 | 0.295 | 0.324 | 0.408 | **0.562** |

# 5 EXPERIMENTS

**Implementation Details.** We evaluate the baseline methods under an $\ell_\infty$ perturbation budget of $16/255$. For our method, we use $12/255$ to maintain image fidelity comparable to the competing methods, except BlurGuard, whose distinct perturbation construction makes direct distortion matching nontrivial. We use 50 optimization steps with a step size of $\gamma = 1/255$. For $\mathcal{L}_{\text{Dis}}$ and $\mathcal{L}_{\text{Rel}}$, we use all residual units in the selected encoder stage, which consists of two residual units. For our method, we select $k^\star = 0$ corresponding to $B_0$ for all VAE encoders, $\lambda$=1.25, 0.5, 1.25 for SD1.4 (Rombach et al., 2022), SD3 (Esser et al., 2024), and FLUX.1 (Labs, 2024), respectively. Our overall algorithm is provided in Appendix B.

**Dataset and Metrics.** For image-to-image (I2I) editing and instruction-guided editing, we evaluate on ImageNet-Edit from BlurGuard (Kim et al., 2025). We also use 42 image–editing prompt pairs from DiffusionGuard (Choi et al., 2025) for inpainting. We report three widely used evaluation metrics: PSNR, CLIP similarity (Radford et al., 2021), and LPIPS (Zhang et al., 2018), and following prior work (Wang et al., 2025), we adopt LPIPS as our primary metric for evaluating protection effectiveness. For image-to-video (I2V) generation, we randomly select 50 images from VBench (Huang et al., 2024) and follow the evaluation protocol of Anti-Prompt (Song et al., 2026).

**Baselines** We compare our method with representative image-immunization approaches, including PhotoGuard (PG) (Salman et al., 2023), Prompt-Independent Defense (PID) (Li et al., 2024), EditShield (ES) (Chen et al., 2024), and AdvPaint (AP) (Jeon et al., 2025). For purification robustness, we additionally consider DCT-Shield (DS) (Bala et al., 2025) and BlurGuard (BG) (Kim et al., 2025). For transferability, we additionally compare with UniDef (Shao et al., 2026), a task-agnostic image-immunization method.

**Surrogate and Target Models.** For surrogate models, we consider three representative diffusion models with different architectures: SD1.4 (Rombach et al., 2022), SD3 (Esser et al., 2024), and FLUX.1 (Labs, 2024). We evaluate black-box transferability across four manipulation tasks using both U-Net- and DiT-based target models. Specifically, we use SD1.4, SD3, and FLUX.1 for I2I editing; SD1.5 Inpainting and FLUX.1 Fill for inpainting; InstructPix2Pix (Brooks et al., 2023) and FLUX.1 Kontext (Labs et al., 2025) for context-based image editing; and DynamiCrafter (Xing et al., 2024) and LTX-Video (HaCohen et al., 2024) for I2V generation. This setup provides a stringent evaluation of black-box generalization across both denoiser architectures and manipulation tasks.

## 5.1 TRANSFERABILITY ACROSS DIVERSE IMAGE MANIPULATION TASKS

We evaluate black-box transferability across I2I editing, inpainting, context-based editing, and I2V generation. Our method outperforms the baselines across black-box I2I targets in Fig. 1, although EditShield (ES) (Chen et al., 2024) performs slightly better in the gray-shaded white-box setting. Our method also achieves the strongest average protection for inpainting and context-based editing in Tabs 1 and 2, and transfers effectively to unseen I2V models in Tab 3. These results demonstrate

Table 2: Black-box transferability evaluation on instruction-guided image editing.

| Target | Surrogate | Metric | AP | UniDef | BG | DS | PG | ES | PID | Ours |
|---|---|---|---|---|---|---|---|---|---|---|
| IP2P | SD1.4 | PSNR ↓ | 16.13 | **12.80** | 19.71 | 14.01 | 15.97 | 13.52 | 16.28 | 15.82 |
| | | CLIP ↓ | 0.766 | **0.698** | 0.813 | 0.759 | 0.787 | 0.725 | 0.792 | 0.766 |
| | | LPIPS ↑ | 0.458 | 0.540 | 0.287 | 0.510 | 0.444 | **0.566** | 0.470 | 0.499 |
| | SD3 | PSNR ↓ | 17.68 | 19.10 | 21.19 | 16.28 | 18.17 | 16.58 | 17.35 | **16.19** |
| | | CLIP ↓ | 0.792 | 0.813 | 0.821 | **0.770** | 0.796 | 0.792 | 0.818 | 0.773 |
| | | LPIPS ↑ | 0.422 | 0.338 | 0.231 | 0.449 | 0.354 | **0.494** | 0.460 | 0.456 |
| | FLUX.1 | PSNR ↓ | 18.14 | 18.29 | 23.44 | 17.28 | 19.48 | 19.28 | 18.46 | **16.03** |
| | | CLIP ↓ | 0.797 | 0.808 | 0.822 | 0.783 | 0.813 | 0.809 | 0.824 | **0.763** |
| | | LPIPS ↑ | 0.380 | 0.369 | 0.150 | 0.404 | 0.291 | 0.293 | 0.390 | **0.496** |
| FLUX-Kontext | SD1.4 | PSNR ↓ | 20.52 | 18.80 | 21.87 | 18.69 | 20.31 | 19.67 | 19.77 | **18.13** |
| | | CLIP ↓ | 0.812 | **0.792** | 0.808 | 0.826 | 0.821 | 0.808 | 0.808 | 0.822 |
| | | LPIPS ↑ | 0.356 | 0.382 | 0.240 | 0.379 | 0.371 | 0.393 | 0.362 | **0.425** |
| | SD3 | PSNR ↓ | 19.52 | 20.61 | 22.46 | **18.60** | 19.68 | 19.07 | 19.43 | 18.79 |
| | | CLIP ↓ | **0.799** | 0.826 | 0.818 | 0.815 | 0.805 | 0.817 | 0.815 | 0.816 |
| | | LPIPS ↑ | 0.376 | 0.296 | 0.203 | 0.361 | 0.309 | **0.415** | 0.365 | 0.383 |
| | FLUX.1 | PSNR ↓ | 19.68 | 18.87 | 24.05 | 19.15 | 20.22 | 20.48 | 19.60 | **18.42** |
| | | CLIP ↓ | 0.810 | 0.819 | 0.822 | **0.798** | 0.807 | 0.815 | 0.816 | 0.819 |
| | | LPIPS ↑ | 0.354 | 0.371 | 0.146 | 0.345 | 0.263 | 0.279 | 0.330 | **0.436** |

Table 3: Transfer results to I2V models. Values are averaged over all metrics.

| Target | Surrogate | Anti-Prompt Protocol (Song et al., 2026) ↓ | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | AP | UniDef | BG | DS | PG | ES | PID | Ours |
| DynamiCrafter | SD1.4 | 3.23 | 3.18 | 3.98 | 3.01 | 3.15 | 3.20 | 3.18 | **2.92** |
| | SD3 | 3.72 | 3.39 | 4.06 | 3.24 | 3.41 | **2.51** | 2.71 | 2.98 |
| | FLUX.1 | 3.06 | 3.08 | 4.33 | 3.37 | 3.42 | 3.21 | 3.05 | **2.97** |
| LTX-Video | SD1.4 | 3.23 | 3.13 | 3.61 | 3.15 | 3.12 | 2.97 | 3.06 | **2.96** |
| | SD3 | 3.14 | 3.53 | 3.37 | 3.20 | 3.10 | 2.98 | 3.01 | **2.92** |
| | FLUX.1 | 3.07 | 3.02 | 3.84 | 3.04 | 3.34 | 3.35 | 3.31 | **2.82** |

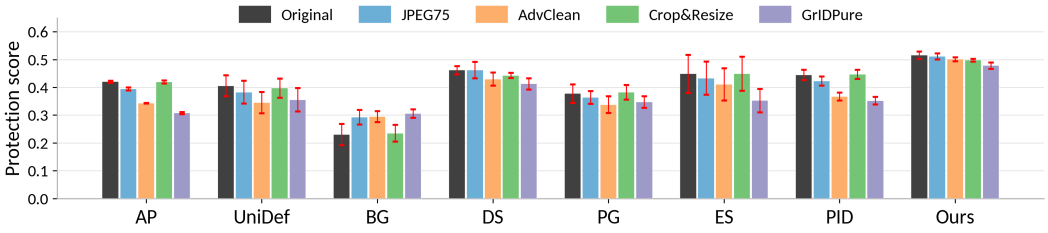



Figure 4: Summary of purification robustness across diverse manipulation tasks and surrogate models. Each bar denotes the mean protection performance across tasks/architectures, and error bars indicate the standard deviation across surrogate models.

consistent transferability across denoiser architectures and manipulation tasks. Qualitative results are provided in Appendix D.

### 5.2 Robustness against Perturbation Purification

For robustness evaluation, we use our primary metric LPIPS (Zhang et al., 2018) for I2I, inpainting, and instruction-guided editing, and the mean score from the Anti-Prompt (Song et al., 2026) protocol for I2V, normalized to $[0, 1]$ with higher values indicating stronger protection. As shown in Fig. 4, our method consistently achieves the highest average protection performance across diverse purification techniques while maintaining a low standard deviation across surrogate models, demonstrating stable cross-surrogate robustness under purification. These results highlight that our method provides transferable protection across diverse image manipulation tasks and model architectures while maintaining robustness against purification, advancing image immunization toward more practical real-world deployment.

### 5.3 Imperceptibility and Cost

Beyond transferability and purification robustness, practical image immunization requires imperceptible perturbations and low computational overhead. Following prior work (Bala et al., 2025), we evaluate distortion relative to the clean image using PSNR for pixel-level fidelity, DISTS (Ding et al., 2022) for deep feature-based perceptual similarity, and FSIM (Zhang et al., 2011) for low-

Table 4: Imperceptibility and cost comparison.

| Metric | | AP | UniDef | BG | DS | PG | ES | PID | Ours |
|---|---|---|---|---|---|---|---|---|---|
| Imperceptibility | PSNR ↑ | 28.62 | 27.89 | **31.64** | 27.52 | 27.78 | 27.25 | 27.97 | 27.40 |
| | DISTS ↓ | 0.204 | 0.249 | **0.087** | 0.262 | 0.327 | 0.304 | 0.239 | 0.273 |
| | FSIM ↑ | 0.959 | 0.925 | **0.985** | 0.978 | 0.954 | 0.902 | 0.944 | 0.964 |
| Cost | Time (sec) ↓ | 155.78 | 30.41 | 34.78 | 11.26 | 10.43 | 10.05 | 12.48 | **6.01** |
| | Memory (MiB) ↓ | 18012 | 9431 | 6973 | 5629 | 5472 | 6483 | 5475 | **3226** |




Figure 5: Visualizations of the immunized images. (Best viewed when zoomed in.)

Table 5: Ablation of the loss components.

| Target | Surrogate | $\mathcal{L}_{\text{Dis}}$ | $\mathcal{L}_{\text{Rel}}$ | $\mathcal{L}_{\text{Dis}} + \mathcal{L}_{\text{Rel}}$ |
|---|---|---|---|---|
| SD1.4 | SD1.4 | 0.527 | 0.545 | **0.553** |
| | SD3 | 0.534 | 0.507 | **0.540** |
| | FLUX.1 | 0.516 | 0.524 | **0.546** |
| SD3 | SD1.4 | 0.547 | 0.592 | **0.593** |
| | SD3 | 0.530 | 0.520 | **0.563** |
| | FLUX.1 | 0.565 | 0.571 | **0.585** |
| FLUX.1 | SD1.4 | 0.522 | **0.567** | **0.567** |
| | SD3 | 0.510 | 0.516 | **0.523** |
| | FLUX.1 | 0.541 | 0.557 | **0.558** |

Table 6: Ablation of the targeted intermediate block.

| Target | Surrogate | B0 | B1 | B2 | B3 | B4 |
|---|---|---|---|---|---|---|
| SD1.4 | SD1.4 | **0.553** | 0.513 | 0.514 | 0.463 | 0.513 |
| | SD3 | **0.540** | 0.519 | 0.524 | 0.486 | 0.466 |
| | FLUX.1 | **0.546** | 0.522 | 0.510 | 0.467 | 0.454 |
| SD3 | SD1.4 | **0.593** | 0.547 | 0.520 | 0.478 | 0.491 |
| | SD3 | 0.563 | **0.603** | 0.591 | 0.554 | 0.584 |
| | FLUX.1 | **0.585** | 0.564 | 0.532 | 0.476 | 0.479 |
| FLUX.1 | SD1.4 | **0.567** | 0.514 | 0.450 | 0.406 | 0.407 |
| | SD3 | 0.523 | **0.563** | 0.531 | 0.455 | 0.439 |
| | FLUX.1 | **0.558** | 0.539 | 0.530 | 0.445 | 0.408 |

level structural similarity. Tab. 4 shows that our method achieves image fidelity comparable to several baselines at the lowest computational cost. BlurGuard (BG) (Kim et al., 2025) preserves substantially higher input fidelity but provides weaker protection in our evaluated settings, reflecting a different protection–distortion trade-off. Fig. 5 also qualitatively confirms the preserved visual quality. Together with the white-box, black-box transfer and purification results in Figs. 1 and 4, our early-stage optimization achieves strong transferability and purification robustness, maintains comparable imperceptibility, and incurs the lowest computational cost.

### 5.4 Ablation Studies

In this section, we conduct ablation studies to further analyze the contribution of each component and design choice in our method. LPIPS (Zhang et al., 2018) is used as the primary metric for evaluating protection effectiveness.

**Loss Components.** We analyze the contribution of each loss component. As shown in Tab. 5, $\mathcal{L}_{\text{Rel}}$ generally achieves higher transferability than $\mathcal{L}_{\text{Dis}}$ alone, highlighting the benefit of disrupting inter-channel organization. Their combination further couples this relational distortion with a substantial clean–immunized representation shift induced by $\mathcal{L}_{\text{Dis}}$, yielding the best overall transferability.

**Selection of Early Block.** We further evaluate the choice of attack depth under $\mathcal{L}_{\text{overall}}$ in Eq. (6) by applying the objective to each intermediate encoder block. As shown in Tab. 6, B0 achieves the highest average black-box performance while retaining competitive white-box performance. Although B1 performs better with the SD3 surrogate for some targets, B0 performs more consistently across surrogate–target pairs. We therefore adopt B0 as our attack target, in line with our primary objective of black-box transferability.

## 6 Conclusion

In this paper, we developed a unified image immunization approach for transfer to unseen models and resilience against purification. Examining successive stages of diffusion pipelines revealed that early VAE encoder features remained more consistent across architectures and retained immunization effects better after purification. Based on this observation, we formulated early-stage representation-discrepancy and relation-disruption losses, enabling lightweight optimization using only the early portion of a surrogate encoder. Evaluations across multiple manipulation tasks, purification techniques, and diffusion architectures showed that the resulting protection transferred broadly and remained effective after purification, bringing image immunization closer to practical deployment.

### AI USE STATEMENT

In this work, we used generative AI tools to improve the readability of the paper, draft and revise portions of the text, discover related work, and assist with research ideation, method implementation, and experimental execution. All AI-assisted code was reviewed and tested by the authors, and all citations, mathematical formulations, and reported results were independently verified. Generative AI tools were not used to fabricate or modify experimental data or reported results. All methodological choices, analyses, and interpretations were critically assessed and finalized by the authors. We reviewed all AI-assisted content and take full responsibility for the final manuscript, including its text, code, and claims.

### ETHICS STATEMENT

This work investigates defensive image immunization against unauthorized manipulation by diffusion models. Our experiments use publicly available image datasets (Kim et al., 2025; Choi et al., 2025; Huang et al., 2024) and pretrained generative models (Rombach et al., 2022; Esser et al., 2024; Labs, 2024; Brooks et al., 2023; Labs et al., 2025; HaCohen et al., 2024; Xing et al., 2024) and do not involve human subjects or private personal data. Although adversarial perturbation techniques may have dual-use implications, our study is intended to support the protection of visual content and is evaluated only in controlled research settings.

### REPRODUCIBILITY STATEMENT

To support reproducibility, we provide detailed descriptions of the evaluated datasets, diffusion models, manipulation tasks, purification methods, optimization objectives, perturbation constraints, and hyperparameters in Sec. 5, and evaluation protocols in the main paper, and pseudo-code in Algorithm 1 on appendix. The appendix further includes random-seed stability results to demonstrate our reliability. We also include the executable code in our supplementary materials. Upon acceptance, we will publicly release the source code, optimization and evaluation scripts, and configuration files required to reproduce the reported experiments.

## REFERENCES

Hervé Abdi and Lynne J Williams. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4):433–459, 2010.

Namhyuk Ahn, KiYoon Yoo, Wonhyuk Ahn, Daesik Kim, and Seung-Hun Nam. Nearly zero-cost protection against mimicry by personalized diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 2025.

Vladimir Arkhipkin, Viacheslav Vasilev, Andrei Filatov, Igor Pavlov, Julia Agafonova, Nikolai Gerasimenko, Anna Averchenkova, Evelina Mironova, Anton Bukashkin, Konstantin Kulikov, et al. Kandinsky 3: Text-to-image synthesis for multifunctional generative framework. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (EMNLP Demo)*, 2024.

Aniruddha Bala, Rohit Chowdhury, Rohan Jaiswal, and Siddharth Roheda. Dct-shield: A robust frequency domain defense against malicious image editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

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 (CVPR)*, 2023.

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 (ECCV)*, 2024.

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 *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025.

Gonzalo Diaz. Adversecleanerextension, 2025. https://github.com/gogodr/AdverseCleanerExtension.

Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 44(5):2567–2581, 2022.

Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2024.

Aditya Ganeshan, Vivek BS, and R Venkatesh Babu. Fda: Feature disruptive attack. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.

Dongnan Gui, Xun Guo, Wengang Zhou, and Yan Lu. I2vguard: Safeguarding images against misuse in diffusion-based image-to-video models. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 2025.

Chuan Guo, Mayank Rana, Moustapha Cisse, and Laurens van der Maaten. Countering adversarial images using input transformations. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018.

Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, et al. Ltx-video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024.

Harold Hotelling. Relations between two sets of variates. In *Breakthroughs in statistics: methodology and distribution*, pp. 162–190. Springer, 1992.

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

Joonsung Jeon, Woo Jae Kim, Suhyeon Ha, Sooel Son, and Sung-Eui Yoon. Advpaint: Protecting images from inpainting manipulation via adversarial attention disruption. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025.

Jinsu Kim, Yunhun Nam, Minseon Kim, Sangpil Kim, and Jongheon Jeong. Blurguard: A simple approach for robustifying image protection against ai-powered editing. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025.

Flux Labs. black-forest-labs/flux, 2024. https://github.com/black-forest-labs/flux.

Chanhui Lee, Yeonghwan Song, and Jeany Son. Data-free universal adversarial perturbation with pseudo-semantic prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Ang Li, Yichuan Mo, Mingjie Li, and Yisen Wang. Pid: Prompt-independent data protection against latent diffusion models. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2024.

Chumeng Liang and Xiaoyu Wu. Mist: Towards improved adversarial examples for diffusion models. *arXiv preprint arXiv:2305.12683*, 2023.

Chumeng Liang, Xiaoyu Wu, Yang Hua, Jiaru Zhang, Yiming Xue, Tao Song, Zhengui Xue, Ruhui Ma, and Haibing Guan. Adversarial example does good: Preventing painting imitation from diffusion models via adversarial examples. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2023.

Renpu Liu, Hao Wu, Jiawei Zhang, Xin Cheng, Xiangyang Luo, Bin Ma, and Jinwei Wang. Pixel2feature attack (p2fa): Rethinking the perturbed space to enhance adversarial transferability. In *Forty-second International Conference on Machine Learning (ICML)*, 2025.

Xuannan Liu, Yaoyao Zhong, Yuhang Zhang, Lixiong Qin, and Weihong Deng. Enhancing generalization of universal adversarial perturbation through gradient aggregation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 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 (CVPR)*, 2024.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. 2018.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Ari Morcos, Maithra Raghu, and Samy Bengio. Insights on representational similarity in neural networks with canonical correlation. *Advances in Neural Information Processing Systems (NeurIPS)*, 31, 2018.

Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Anima Anandkumar. Diffusion models for adversarial purification. 2022.

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

Shuangshuang Pu, Wen Yang, Min Li, Guodong Liu, Chris Ding, and Di Ming. Low-rank and sparsity are all you need: Exploring robust hierarchical latent subspaces for transferable adversarial attack. In *Forty-third International Conference on Machine Learning (ICML)*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2021.

Anton Razzhigaev, Arseniy Shakhmatov, Anastasia Maltseva, Vladimir Arkhipkin, Igor Pavlov, Ilya Ryabov, Angelina Kuts, Alexander Panchenko, Andrey Kuznetsov, and Denis Dimitrov. Kandinsky: an improved text-to-image synthesis with image prior and latent diffusion. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (EMNLP Demo)*, 2023.

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)*, 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 International Conference on Machine Learning (ICML)*, 2023.

Shawn Shan, Jenna Cryan, Emily Wenger, Haitao Zheng, Rana Hanocka, and Ben Y Zhao. Glaze: Protecting artists from style mimicry by {Text-to-Image} models. In *32nd USENIX Security Symposium (USENIX Security 23)*, 2023.

Mingwen Shao, Lingzhuang Meng, Xiang Lv, Mengyao Wu, Xinyuan Chen, Qiao Zhang, Chang Liu, Yuanjian Qiao, and Chao Dong. Unidef: Universal defense against unauthorized image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Hamid R Sheikh and Alan C Bovik. Image information and visual quality. *IEEE Transactions on image processing*, 15(2):430–444, 2006.

Yeonghwan Song, Chanhui Lee, Jinsoo Park, and Jeany Son. Anti-prompt: Image protection against text-guided image-to-video generation. In *European Conference on Computer Vision (ECCV)*, 2026.

Duc Vu, Anh Nguyen, Chi Tran, and Anh Tran. Anti-i2v: Safeguarding your photos from malicious image-to-video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 2026.

Hanhui Wang, Yihua Zhang, Ruizheng Bai, Yue Zhao, Sijia Liu, and Zhengzhong Tu. Edit away and my face will not stay: Personal biometric defense against malicious generative editing. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 2025.

Yiming Wang, Jiahao Chen, Qingming Li, Chunyi Zhou, Zhi Chen, Lingzhong Meng, Jinbao Li, and Shouling Ji. Facearmor: A universal facial image protection against diffusion-based manipulations. In *European Conference on Computer Vision (ECCV)*, 2026.

Zhibo Wang, Hengchang Guo, Zhifei Zhang, Wenxin Liu, Zhan Qin, and Kui Ren. Feature importance-aware transferable adversarial attacks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

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.

Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Wangbo Yu, Hanyuan Liu, Gongye Liu, Xintao Wang, Ying Shan, and Tien-Tsin Wong. Dynamicrafter: Animating open-domain images with video diffusion priors. In *European Conference on Computer Vision (ECCV)*, 2024.

Wen Yang, Guodong Liu, and Di Ming. Smp-attack: Boosting the transferability of feature importance-based adversarial attack with semantics-aware multi-granularity patchout. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025a.

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 (ICCV)*, 2025b.

Jianping Zhang, Weibin Wu, Jen-tse Huang, Yizhan Huang, Wenxuan Wang, Yuxin Su, and Michael R Lyu. Improving adversarial transferability via neuron attribution-based attacks. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Lin Zhang, Lei Zhang, Xuanqin Mou, and David Zhang. Fsim: A feature similarity index for image quality assessment. *IEEE transactions on Image Processing*, 20(8):2378–2386, 2011.

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/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

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 Computer Vision and Pattern Recognition Conference (CVPR)*, 2024.

# Rethinking Image Immunization for Transferable and Purification-Robust Protection

*- Appendix -*

**A Experimental Details** **15**
A.1 Target Models . . . 15
A.2 Baselines and Implementation Details . . . 15
A.3 Evaluation Metrics. . . . 16
A.4 Details of Purification Techniques . . . 16
A.5 PWCCA . . . 17

**B Algorithm** **17**

**C Additional Experiments** **18**
C.1 Ablation Studies . . . 18
C.2 Comparison with I2V-specific Immunization Methods . . . 20
C.3 Effect of Perturbation Budget . . . 20
C.4 Stability Evaluation . . . 21
C.5 Additional Evaluation Metrics for I2I Transferability . . . 21
C.6 Task-wise Purification Robustness Evaluation . . . 21
C.7 Transferability Evaluation across Different VAE . . . 21

**D Qualitative Results** **23**

**E Limitations and Discussion** **23**

**F Reproducibility** **32**

# A Experimental Details

## A.1 Target Models

The target pipelines differ in how the protected image enters the generation process. These conditioning mechanisms provide complementary tests of whether perturbations optimized through an image encoder remain effective after integration into another generative pipeline.

**Image-conditioned editing.** InstructPix2Pix (Brooks et al., 2023) concatenates the encoded source image with the noisy latent along the channel dimension, while the editing instruction is supplied through text conditioning. FLUX.1 Kontext (Labs et al., 2025) instead appends the context image's latent tokens to the target image tokens in the visual stream. The source representation is therefore accessed through different conditioning pathways in these editors, extending the comparison beyond a change in the denoiser backbone alone.

**Video generation.** DynamiCrafter (Xing et al., 2024) uses image-derived context embeddings together with the full image as a detail-preserving condition. LTX-Video (HaCohen et al., 2024) encodes the conditioning image with a causal video VAE and joins its latent tokens with the tokens to be generated, using token-specific timesteps to distinguish their roles. These mechanisms test transfer to video generation with conditioning structures that differ from the image-editing surrogates.

## A.2 Baselines and Implementation Details

Our comparisons include methods that modify latent representations, interfere with denoiser attention, or control the frequency content of protective noise. The following descriptions clarify the objectives and perturbation designs behind the baseline entries in the result tables.

**PhotoGuard (PG).** The encoder attack in PhotoGuard (Salman et al., 2023) uses a prescribed latent target to guide perturbation optimization. By reducing the distance between the protected image's encoding and this target, it steers the encoder away from its representation of the source content. The attack thus specifies a destination in latent space, whereas untargeted encoder attacks seek to increase the discrepancy from the clean encoding.

**EditShield (ES).** In EditShield (Chen et al., 2024), protection is obtained by separating the encoded representations of the original and perturbed inputs. This separation is intended to corrupt the image condition supplied to an instruction-guided editor. The original objective also penalizes perceptual distortion and averages over input transformations to account for their effect on the latent discrepancy. The representation being optimized is the final VAE encoding.

**Prompt-Independent Defense (PID).** PID (Li et al., 2024) formulates protection in terms of the distribution predicted by the visual encoder. Its published loss jointly modifies the latent mean and log variance; the logarithmic term accommodates the difference in scale between the two statistics. Since these quantities are computed from the image alone, the protection objective does not require the text that will later guide manipulation.

**AdvPaint (AP).** The protection objective of AdvPaint (Jeon et al., 2025) is defined on the internal attention representations of the diffusion denoiser. For inpainting, its original design perturbs self-attention queries, keys, and values together with cross-attention queries. These changes interfere with relationships within the image and with text conditioning, providing a denoiser-based counterpart to the encoder objectives considered here.

**DCT-Shield (DS).** In DCT-Shield (Bala et al., 2025), the optimization variables are changes to quantized DCT coefficients. The JPEG decoding operations map these coefficients back to an image, whose VAE latent norm supplies the original method's minimization objective. JPEG processing is therefore part of the perturbation construction, linking the latent objective to the representation used for compression.

**BlurGuard (BG).** BlurGuard (Kim et al., 2025) controls the spectrum of protective noise through Gaussian smoothing whose strength varies spatially. The blur parameters are learned to bring the protected image's power spectrum closer to that of the clean input, while an image-protection objective maintains the intended disruption. This design addresses spectral deviations that could otherwise make the perturbation easier to suppress through purification. Importantly, in our implementation, BlurGuard's distinct perturbation construction makes direct perceptual-distortion matching nontrivial. We therefore report its resulting fidelity metrics alongside protection performance, without claiming a distortion-matched comparison.

**UniDef.** A shared denoising objective forms the basis of task-agnostic protection in UniDef (Shao et al., 2026), with finite differences used to estimate gradients. Its inclusion allows the transfer results to compare perturbations derived from denoising behavior with those obtained by directly modifying encoder representations. The downstream tasks then test how well these distinct optimization targets generalize.

### A.3 Evaluation Metrics.

**PSNR.** PSNR evaluates pixel-level fidelity to the clean-reference editing result using mean squared error. A higher PSNR indicates greater pixel-wise similarity, whereas a lower value reflects stronger distortion induced by immunization.

**CLIP Similarity.** CLIP Similarity computes the cosine similarity between the CLIP (Radford et al., 2021) features of the source image and the edited output. A lower value indicates weaker semantic correspondence to the source image and thus stronger protection.

**LPIPS.** LPIPS (Zhang et al., 2018) measures perceptual distance from the clean reference editing result in a deep feature space. A higher LPIPS indicates a larger perceptual deviation and more effective immunization. Following FaceLock (Wang et al., 2025), we adopt LPIPS as our primary evaluation metric.

**Anti-Prompt Protocol.** We evaluate I2V immunization using the protocol of Anti-Prompt (Song et al., 2026). Qwen3-VL-8B scores generated videos on subject preservation, structural consistency, dynamic consistency, and artifact suppression, each on a 1-–5 scale. We report the mean of these four scores. A lower value indicates greater disruption of image-faithful video generation and thus stronger protection.

### A.4 Details of Purification Techniques

Before manipulation, the protected inputs are processed by one of four purification operations. Each probes a different source of perturbation sensitivity: coefficient quantization, local filtering, generative reconstruction, or spatial resampling.

**JPEG compression.** During JPEG encoding (Guo et al., 2018), image blocks are transformed into DCT coefficients and quantized. Coefficient differences smaller than the quantization intervals may disappear when the image is reconstructed. Consequently, compression can attenuate protective signals, particularly in frequency bands that are quantized more coarsely. This comparison assesses protection after an operation commonly encountered when images are stored or shared.

**Adverse Cleaner (AdvClean).** The filtering pipeline of Adverse Cleaner (Diaz, 2025) uses bilateral smoothing followed by guided refinement. The bilateral stage aggregates nearby pixels with similar appearance, while the guided stage uses local image structure to refine the output. Together, they seek to reduce noise without strongly blurring prominent edges, providing a direct test of resistance to image filtering.

**Crop&Resize.** This operation retains a subregion of the protected image and scales it back to the required dimensions. Cropping shifts the perturbation's spatial alignment, and interpolation modifies its pixel-level pattern during resizing. The corresponding rows measure how protection responds to changes in image framing and sampling.

**GridPure.** To accommodate high-resolution inputs, GridPure (Zhao et al., 2024) performs diffusion-based purification (Nie et al., 2022) on overlapping image patches. Each patch undergoes a short noising-and-denoising process, after which overlapping predictions are averaged to

reconstruct the full image. Repeated purification and blending aim to remove perturbations while limiting changes to the underlying structure. The patchwise procedure preserves the input resolution without requiring whole-image downsampling.

Sec. C.6 reports each operation separately so that the results expose purifier-specific behavior within individual surrogate–target pairs.

### A.5 PWCCA

Projection-weighted canonical correlation analysis (PWCCA) (Morcos et al., 2018) measures the similarity between two representation spaces without requiring one-to-one alignment between their feature dimensions. At encoder depth $k$, let $X_k^{(s)} \in \mathbb{R}^{c_k^{(s)} \times N}$ and $X_k^{(t)} \in \mathbb{R}^{c_k^{(t)} \times N}$ denote centered block representations from the surrogate and target encoders, where rows correspond to channels and columns correspond to the same $N$ observations. Although PWCCA is computed between corresponding blocks, its values are compared across depths with different channel dimensions (*e.g.*, *B0*: 128-channels and *B1*: 256-channels). We therefore project every block representation onto a fixed 128-dimensional PCA (Abdi & Williams, 2010) subspace before applying CCA (Hotelling, 1992), yielding $H_k^{(s)}, H_k^{(t)} \in \mathbb{R}^{128 \times N}$.

CCA identifies paired linear combinations of $H_k^{(s)}$ and $H_k^{(t)}$ with canonical correlations $\rho_{k,\ell}$. PWCCA aggregates these correlations using normalized projection weights $\alpha_{k,\ell}^{(s)}$ that reflect the importance of each canonical component to the surrogate representation:

$$\mathrm{PWCCA}_k^{s \to t} = \sum_{\ell=1}^{128} \alpha_{k,\ell}^{(s)} \rho_{k,\ell}, \qquad \sum_{\ell=1}^{128} \alpha_{k,\ell}^{(s)} = 1. \tag{8}$$

A larger value indicates stronger correspondence between surrogate-important feature combinations and the target representation. Because the projection weights are computed from the surrogate side, PWCCA is directional and generally satisfies

$$\mathrm{PWCCA}_k^{s \to t} \neq \mathrm{PWCCA}_k^{t \to s}. \tag{9}$$

We apply the same PCA preprocessing and PWCCA formulation to clean representations and perturbation-induced representation shifts in Sec. 3.3.

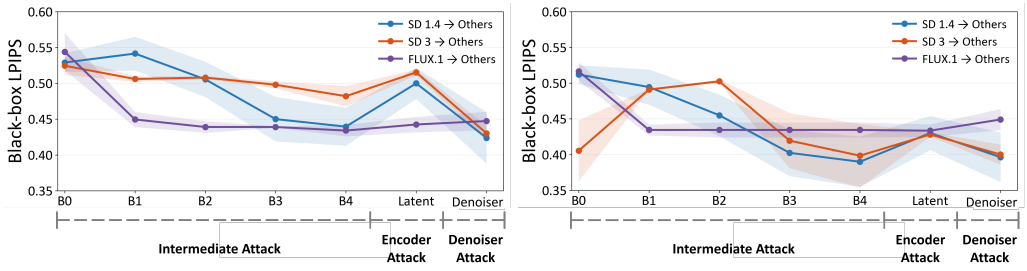
**Algorithm 1** Transferable and Purification-Robust Image Immunization

**Require:** Image $x$, surrogate encoder $E_s$, target stage $k$, Step size $\gamma$, iterations $T$ budgets $\eta_\infty$
**Ensure:** Immunized image $x_{\mathrm{adv}}$

$\{h_i^{\mathrm{clean}}, a_i^{\mathrm{clean}}\}_{i \in \mathcal{I}_k} \leftarrow E_s^{(k)}(x)$ *// Extract clean features*
$\delta \sim \mathcal{U}(-1/255, 1/255)$ *// Random initialization*
$\delta \leftarrow \mathrm{clip}_{[0,1]}(x + \delta) - x$
**for** $j = 0, \ldots, T-1$ **do**
  $x_{\mathrm{adv}} \leftarrow \mathrm{clip}_{[0,1]}(x + \delta)$ *// Construct immunized image*
  $\{h_i^{\mathrm{adv}}, a_i^{\mathrm{adv}}\}_{i \in \mathcal{I}_k} \leftarrow E_s^{(k)}(x_{\mathrm{adv}})$ *// Extract adversarial features*
  $\mathcal{L}_{\mathrm{overall}} \leftarrow \mathcal{L}_{\mathrm{Dis}} + \lambda \mathcal{L}_{\mathrm{Rel}}$ *// Compute overall loss*
  $\delta \leftarrow \delta - \gamma \ \mathrm{sign}\left(\nabla_\delta \mathcal{L}_{\mathrm{overall}}\right)$ *// Update perturbation*
  $\delta \leftarrow \mathrm{clip}_{[-\eta_\infty, \eta_\infty]}(\delta)$ *// Apply $\ell_\infty$ constraint*
  $\delta \leftarrow \mathrm{clip}_{[0,1]}(x + \delta) - x$
**end for**
**return** $x_{\mathrm{adv}} = \mathrm{clip}_{[0,1]}(x + \delta)$

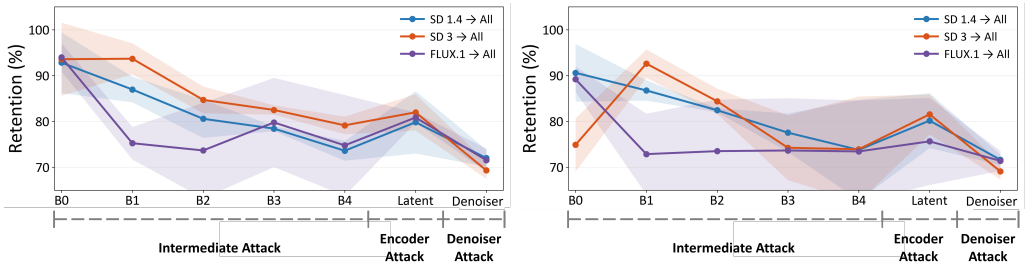
## B ALGORITHM

Algorithm 1 summarizes our immunization procedure. We first cache the clean output and additive features at the selected surrogate-encoder stage. At each iteration, we optimize the overall objective $\mathcal{L}_{\mathrm{Overall}}$ combined $\mathcal{L}_{\mathrm{Dis}}$ and $\mathcal{L}_{\mathrm{Rel}}$ using a sign-gradient update. The resulting perturbation is projected onto the $\ell_\infty$ and clipped to preserve the valid image range. Iterating this process yields the final immunized image $x_{\mathrm{adv}}$.




(a) Black-box transferability of unit-normalized $\ell_2$ loss. (b) Black-box transferability of cosine direction loss.

Figure 6: Controlled comparison of attack surfaces across the diffusion pipeline with normalized-$\ell_2$ loss and cosine direction loss on transferability perspective. (a) Black-box LPIPS from independently optimizing the unit-normalized $\ell_2$ objective at each diffusion component. (b) The corresponding component-wise comparison using the cosine-direction objective.




(a) Retention after purification of unit-normalized $\ell_2$ loss. (b) Retention after purification of cosine direction loss.

Figure 7: Controlled comparison of attack surfaces across the diffusion pipeline with unit-normalized $\ell_2$ loss and cosine direction loss on purification robustness perspective. (a) Fraction of pre-purification protection retained after purification under the unit-normalized $\ell_2$ objective. (b) Corresponding retention fractions under the cosine-direction objective.

# C Additional Experiments

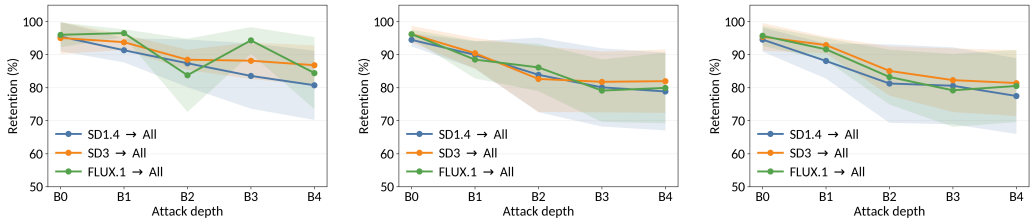
## C.1 Ablation Studies

**Discrepancy Maximization Loss Ablation.** As noted in Sec. 3.2, our controlled analysis examines whether intermediate VAE representations can serve as a promising attack surface for both black-box transferability and purification robustness. To further examine this tendency under different discrepancy formulations, we repeat the depth-wise comparison using the $\ell_2$ distance between unit-normalized features in Eq. (10) and the cosine-direction discrepancy in Eq. (11). We apply each objective to B0–B4, the final VAE latent, and denoiser Q/K/V features following Sec. 3.2.

$$\mathcal{L}^{(k)}_{\mathrm{norm}_{\ell_2}} = -\frac{1}{|\mathcal{I}_k|}\sum_{i\in\mathcal{I}_k}\left\|\frac{\phi^{(i)}_k(x+\delta)}{\|\phi^{(i)}_k(x+\delta)\|_2} - \frac{\phi^{(i)}_k(x)}{\|\phi^{(i)}_k(x)\|_2}\right\|_2, \tag{10}$$

$$\mathcal{L}^{(k)}_{\cos} = -\frac{1}{|\mathcal{I}_k|}\sum_{i\in\mathcal{I}_k}\left(1 - \frac{\phi^{(i)}_k(x+\delta)^\top \phi^{(i)}_k(x)}{\|\phi^{(i)}_k(x+\delta)\|_2\|\phi^{(i)}_k(x)\|_2}\right). \tag{11}$$

Figs. 6 and 7 show that, under both alternative losses, intermediate VAE stages generally provide favorable black-box transferability and protection retention after perturbation purification compared with deeper components, including the final VAE latent and denoiser Q/K/V features. As noted in Sec. 3.2, our analysis aims to identify a potential attack surface for both transferability and purification robustness; it does not imply that every loss applied to an intermediate stage guarantees both properties. For example, B0 with the SD3 (Esser et al., 2024) surrogate under the cosine-direction loss does not achieve the strongest transferability or retention. Nevertheless, B1 and B2 in the same




(a) Discrepancy maximization. (b) Relation-disruption. (c) Overall loss.

Figure 8: Depth-wise purification robustness ablations of intermediate VAE attack targets under different loss components: (a) discrepancy maximization loss $\mathcal{L}_{\text{Dis}}$, (b) relation-disruption loss $\mathcal{L}_{\text{Rel}}$, and (c) overall loss $\mathcal{L}_{\text{overall}}$. *SD1.4 → All* denotes SD1.4 as the surrogate VAE, with retention averaged over all three target VAEs. Lines and shaded regions indicate the mean and standard deviation across four purification methods, respectively.

setting outperform the final VAE latent and denoiser Q/K/V features on both measures, still supporting our motivation to investigate intermediate VAE representations as attack targets.

Table 7: Ablation of relation-disruption objectives evaluated by LPIPS (Zhang et al., 2018). Each variant uses only the indicated relation loss at B0. Best and second-best results are shown in bold and underlined, respectively.

| Target | Surrogate | $s$–$s$ | $a$–$s$ | $h$–$h$ | $a$–$a$ |
|---|---|---|---|---|---|
| SD1.4 | SD1.4 | 0.541 | <u>0.542</u> | 0.505 | **0.545** |
| | SD3 | 0.532 | <u>0.536</u> | **0.546** | 0.507 |
| | FLUX.1 | <u>0.540</u> | **0.542** | 0.499 | 0.524 |
| SD3 | SD1.4 | 0.560 | 0.554 | <u>0.562</u> | **0.592** |
| | SD3 | 0.507 | **0.541** | 0.510 | <u>0.520</u> |
| | FLUX.1 | 0.547 | 0.537 | <u>0.564</u> | **0.571** |
| FLUX.1 | SD1.4 | 0.527 | <u>0.536</u> | 0.517 | **0.567** |
| | SD3 | 0.503 | 0.507 | <u>0.511</u> | **0.516** |
| | FLUX.1 | 0.540 | <u>0.547</u> | 0.532 | **0.557** |

**Relation-disruption Loss Ablation.** We next isolate the relation-disruption objective and compare relations constructed within the shortcut path ($s$–$s$), between the residual and shortcut paths ($a$–$s$), within the block output ($h$–$h$), and within the residual path ($a$–$a$). Tab. 7 shows that, although the $a$–$a$ relation yields slightly lower white-box performance in one setting, it achieves the highest average black-box LPIPS across surrogate–target pairs. This aligns with our primary goal of transferability and supports our choice of the $a$–$a$ relation for the proposed loss.

**Robustness Ablation on Loss Components.** Our initial encoder depth-wise analysis in Fig. 2b identifies promising intermediate targets using a controlled discrepancy objective under Adverse Cleaner (Diaz, 2025) and GridPure (Zhao et al., 2024). However, the final method additionally incorporates relation disruption, which may alter the resulting perturbation structure and its response to purification. The robustness advantage observed in the initial analysis, therefore, may not directly carry over to the complete immunization objective.

We examine this compatibility by independently optimizing $\mathcal{L}_{\text{Dis}}$, $\mathcal{L}_{\text{Rel}}$, and $\mathcal{L}_{\text{overall}}$ at stages B0–B4 under identical settings. Retention is evaluated after JPEG (Guo et al., 2018), Adverse Cleaner (Diaz, 2025), Crop&Resize, and GridPure (Zhao et al., 2024). Fig. 8 reports the average across surrogate–target pairs and purification methods, with shaded regions indicating variation across purification methods.

All three variants achieve their highest retention at B0. Specifically, $\mathcal{L}_{\text{Dis}}$, $\mathcal{L}_{\text{Rel}}$, and $\mathcal{L}_{\text{overall}}$ retain approximately $95.7\%$, $95.6\%$, and $95.2\%$ of their pre-purification performance, respectively. The comparable retention of the overall objective confirms that incorporating $\mathcal{L}_{\text{Rel}}$ to improve transferability does not compromise the purification robustness of early-stage optimization. These results support B0 as the final attack depth within our objective design.

Table 8: Performance before and after purification for individual loss components across encoder depths. Before denotes the mean over all $3 \times 3$ surrogate–target pairs. After reports the mean of the cell-wise purification results, with the average sample standard deviation across JPEG, Adverse Cleaner, Crop&Resize, and GridPure.

| Objective | Purification | B0 | B1 | B2 | B3 | B4 |
|---|---|---|---|---|---|---|
| $\mathcal{L}_{\text{Dis}}$ | Before | 0.532 | 0.520 | 0.492 | 0.489 | 0.486 |
| | After | $0.509 \pm 0.024$ | $0.488 \pm 0.008$ | $0.427 \pm 0.031$ | $0.432 \pm 0.022$ | $0.409 \pm 0.038$ |
| $\mathcal{L}_{\text{Rel}}$ | Before | 0.544 | 0.531 | 0.520 | 0.486 | 0.479 |
| | After | $0.520 \pm 0.012$ | $0.476 \pm 0.026$ | $0.438 \pm 0.049$ | $0.390 \pm 0.048$ | $0.384 \pm 0.051$ |
| $\mathcal{L}_{\text{overall}}$ | Before | 0.559 | 0.543 | 0.523 | 0.470 | 0.471 |
| | After | $0.532 \pm 0.021$ | $0.493 \pm 0.019$ | $0.435 \pm 0.047$ | $0.379 \pm 0.049$ | $0.376 \pm 0.050$ |

Table 9: Comparison with I2V-specific immunization methods.

| Method | Surrogate | Anti-Prompt Protocol ↓ | Latency (sec) | Memory Usage (GiB) |
|---|---|---|---|---|
| I2VGuard | LTX-Video | 2.76 | 490.95 | 35.18 |
| Anti-Prompt | LTX-Video | 2.67 | 256.28 | 28.86 |
| Ours | SD1.4 | 2.96 | 6.01 | 3.15 |
| | SD3 | 2.92 | 6.01 | 3.15 |
| | FLUX.1 | 2.82 | 6.01 | 3.15 |
| | Average | 2.90 | 6.01 | 3.15 |

### C.2 Comparison with I2V-specific Immunization Methods

We compare our method with I2VGuard (Gui et al., 2025) and Anti-Prompt (Song et al., 2026) on LTX-Video (HaCohen et al., 2024), as Anti-Prompt is designed for DiT-based I2V models. Both baselines operate in a **white-box** setting using gradients from the target LTX-Video model, whereas our method accesses only the early stages of a surrogate VAE encoder and transfers the resulting perturbations to LTX-Video in a **black-box** setting.

As shown in Tab. 9, our method achieves comparable protection without access to the target model while requiring substantially lower latency and memory, further highlighting its effectiveness and computational efficiency for black-box I2V immunization.

Table 10: Trade-off between $\ell_\infty$ perturbation budget and performance.

| Target | Surrogate | Ours $\eta_\infty = 16/255$ | Ours $\eta_\infty = 12/255$ |
|---|---|---|---|
| SD1.4 | SD1.4 | 0.568 | 0.553 |
| | SD3 | 0.559 | 0.540 |
| | FLUX.1 | 0.556 | 0.546 |
| SD3 | SD1.4 | 0.606 | 0.593 |
| | SD3 | 0.581 | 0.563 |
| | FLUX.1 | 0.599 | 0.585 |
| FLUX.1 | SD1.4 | 0.587 | 0.567 |
| | SD3 | 0.550 | 0.523 |
| | FLUX.1 | 0.573 | 0.558 |

Table 11: Imperceptibility evaluation on surrogate models.

| Constraint | Surrogate | PSNR ↑ | FSIM ↑ | DISTS ↓ |
|---|---|---|---|---|
| Ours $\eta_\infty = 16/255$ | SD1.4 | 25.385 | 0.949 | 0.330 |
| | SD3 | 24.930 | 0.936 | 0.285 |
| | FLUX.1 | 25.526 | 0.955 | 0.304 |
| | Mean | 25.280 | 0.947 | 0.306 |
| Ours $\eta_\infty = 12/255$ | SD1.4 | 27.467 | 0.966 | 0.295 |
| | SD3 | 27.158 | 0.958 | 0.251 |
| | FLUX.1 | 27.571 | 0.969 | 0.274 |
| | Mean | 27.399 | 0.964 | 0.273 |

### C.3 Effect of Perturbation Budget

To match the imperceptibility of competing methods, we set $\eta_\infty = 12/255$ for our method in the main evaluation. To examine the protection–imperceptibility trade-off within our method, we evaluate it at $\eta_\infty = 12/255$ and $16/255$ and report both protection performance and imperceptibility.

As shown in Tabs 10 and 11, the *Ours*-$\eta_\infty = 12/255$ slightly reduces protection performance compared with *Ours*-$\eta_\infty = 16/255$ but provides imperceptibility comparable to competing methods in Tab. 4. Despite this reduction, our method maintains superior black-box transferability and purification robustness.

Table 12: I2I attack performance across five generation seeds. Each entry reports the mean ± standard deviation of LPIPS.

| Surrogate | Target | | |
|---|---|---|---|
| | SD1.4 | SD3 | FLUX.1 |
| SD1.4 | $0.553 \pm 0.016$ | $0.593 \pm 0.013$ | $0.567 \pm 0.014$ |
| SD3 | $0.540 \pm 0.011$ | $0.563 \pm 0.009$ | $0.523 \pm 0.015$ |
| FLUX.1 | $0.546 \pm 0.014$ | $0.585 \pm 0.013$ | $0.558 \pm 0.012$ |

### C.4 Stability Evaluation

To provide a comprehensive evaluation, our main tables compare a broad range of baselines (Salman et al., 2023; Kim et al., 2025; Bala et al., 2025; Li et al., 2024; Shao et al., 2026; Jeon et al., 2025; Chen et al., 2024) across multiple models and tasks. For readability, we report standard deviations separately in this section. We repeat each experiment across five generation seeds and report the mean and standard deviation. As shown in Tab. 12, our method exhibits consistently low variation across all surrogate–target pairs, demonstrating stable immunization performance across generation seeds.

### C.5 Additional Evaluation Metrics for I2I Transferability

Following FaceLock (Wang et al., 2025), we adopt LPIPS (Zhang et al., 2018) as the primary metric for the ablation and purification-robustness evaluations, while also reporting PSNR in our main evaluations. To verify that our conclusions are not specific to perceptual feature distance, we additionally evaluate, CLIP similarity (CLIP) (Radford et al., 2021) for measuring semantic correspondence between the source and the edited result, SSIM (Wang et al., 2004) for structural similarity and VIFp (Sheikh & Bovik, 2006) for visual information fidelity, both widely used in image immunization (Lo et al., 2024; Salman et al., 2023). Specifically, SSIM evaluates structural similarity in luminance, contrast, and texture, and VIFp quantifies preserved visual information through perceptual modeling, with lower values indicating stronger immunization for all three metrics. As shown in Table 13, our method consistently achieves strong black-box protection across all three metrics, demonstrating that its transferability extends beyond perceptual disruption to structural and information-level degradation.

### C.6 Task-wise Purification Robustness Evaluation

We provide separate purification robustness evaluations for image-to-image editing, inpainting, and instruction-guided editing tasks. We use LPIPS (Zhang et al., 2018) as the primary metric, where a higher value indicates stronger protection after purification.

The results in Tabs 16, 17, and 18 show that our method maintains strong overall protection after purification across image-to-image editing, inpainting, and instruction-guided editing, supporting its transferability and robustness to purification. Tab. 19 further evaluates purification robustness on unseen I2V models. Across two target models, three surrogate models, and four purification techniques, our method shows consistently competitive performance, achieving the best overall average and the strongest average protection on both DynamiCrafter (Xing et al., 2024) and LTX-Video (HaCohen et al., 2024). These results further support the robustness of our transferred protection across diverse I2V pipelines and purification operations.

### C.7 Transferability Evaluation across Different VAE

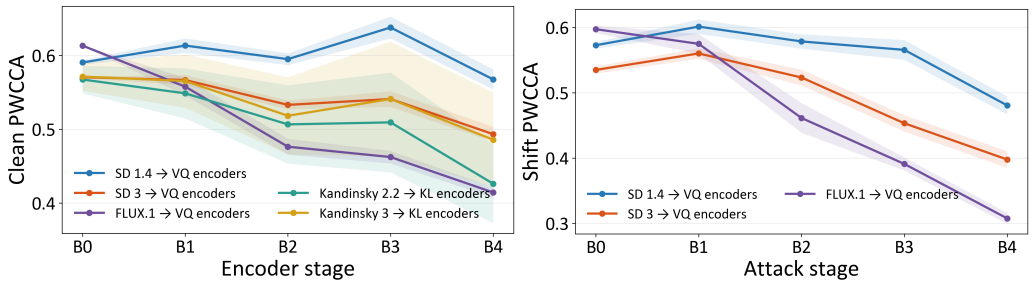
The VAEs of SD1.4, SD3, and FLUX.1 share the same hierarchical encoder backbone despite having different latent interfaces. Each encoder consists of an input convolution, four encoder down blocks, and a middle block. Each down block contains two residual units, and the first three blocks are followed by a $2\times$ spatial downsampler. Their channel widths are identically configured as $128$, $256$, $512$, and $512$. The middle block preserves the final spatial resolution and contains residual units and a self-attention layer. Consequently, for the common $512 \times 512$ input resolution used in our analysis, all three encoders produce spatially and channel-wise matched intermediate representations.

Table 13: I2I transferability measured with diverse metrics.

| Target | Surrogate | Metric | AP | UniDef | BG | DS | PG | ES | PID | Ours |
|---|---|---|---|---|---|---|---|---|---|---|
| SD1.4 | SD1.4 | PSNR ↓ | 16.19 | 15.08 | 17.54 | 13.99 | **13.85** | 14.92 | 14.38 | 14.17 |
| | | SSIM ↓ | 0.537 | 0.478 | 0.643 | 0.440 | 0.473 | **0.354** | 0.400 | 0.364 |
| | | VIFp ↓ | 0.174 | **0.087** | 0.233 | 0.119 | 0.131 | 0.112 | 0.130 | 0.123 |
| | | CLIP ↓ | 0.706 | 0.716 | 0.707 | 0.689 | **0.688** | 0.701 | 0.703 | 0.692 |
| | | LPIPS ↑ | 0.433 | 0.552 | 0.365 | 0.537 | 0.536 | **0.568** | 0.540 | 0.553 |
| | SD3 | PSNR ↓ | 16.38 | 17.58 | 19.11 | 15.86 | 16.49 | 15.19 | 15.30 | **14.72** |
| | | SSIM ↓ | 0.561 | 0.605 | 0.709 | 0.464 | 0.588 | 0.507 | 0.516 | **0.380** |
| | | VIFp ↓ | 0.186 | 0.215 | 0.287 | 0.146 | 0.213 | 0.157 | 0.180 | **0.131** |
| | | CLIP ↓ | 0.709 | 0.715 | 0.712 | 0.712 | 0.710 | 0.705 | **0.696** | 0.698 |
| | | LPIPS ↑ | 0.417 | 0.379 | 0.287 | 0.483 | 0.393 | 0.471 | 0.445 | **0.540** |
| | FLUX.1 | PSNR ↓ | 16.65 | 16.94 | 20.66 | 16.54 | 17.69 | 17.28 | 15.79 | **14.08** |
| | | SSIM ↓ | 0.559 | 0.574 | 0.758 | 0.497 | 0.652 | 0.578 | 0.500 | **0.369** |
| | | VIFp ↓ | 0.188 | 0.194 | 0.346 | 0.160 | 0.264 | 0.205 | 0.177 | **0.126** |
| | | CLIP ↓ | 0.707 | 0.714 | 0.713 | 0.712 | 0.712 | 0.711 | 0.703 | **0.688** |
| | | LPIPS ↑ | 0.419 | 0.407 | 0.238 | 0.460 | 0.334 | 0.392 | 0.455 | **0.546** |
| SD3 | SD1.4 | PSNR ↓ | 17.36 | 18.07 | 20.30 | 16.46 | 17.32 | 16.89 | 16.91 | **15.66** |
| | | SSIM ↓ | 0.560 | 0.640 | 0.764 | 0.413 | 0.570 | 0.508 | 0.538 | **0.369** |
| | | VIFp ↓ | 0.182 | 0.215 | 0.319 | **0.105** | 0.173 | 0.162 | 0.180 | 0.110 |
| | | CLIP ↓ | 0.723 | 0.733 | 0.746 | 0.726 | 0.722 | 0.720 | 0.729 | **0.718** |
| | | LPIPS ↑ | 0.459 | 0.407 | 0.288 | 0.558 | 0.471 | 0.490 | 0.477 | **0.593** |
| | SD3 | PSNR ↓ | 16.11 | 17.57 | 19.75 | 15.68 | 17.04 | **15.63** | 16.21 | 15.88 |
| | | SSIM ↓ | 0.469 | 0.572 | 0.739 | 0.361 | 0.585 | **0.336** | 0.446 | 0.405 |
| | | VIFp ↓ | 0.131 | 0.182 | 0.293 | 0.092 | 0.179 | **0.075** | 0.113 | 0.122 |
| | | CLIP ↓ | 0.719 | 0.735 | 0.747 | 0.711 | 0.731 | **0.692** | 0.712 | 0.726 |
| | | LPIPS ↑ | 0.533 | 0.453 | 0.307 | 0.593 | 0.460 | **0.635** | 0.565 | 0.563 |
| | FLUX.1 | PSNR ↓ | 17.43 | 17.67 | 22.11 | 16.92 | 18.15 | 18.41 | 16.95 | **15.85** |
| | | SSIM ↓ | 0.565 | 0.553 | 0.820 | 0.478 | 0.642 | 0.632 | 0.531 | **0.378** |
| | | VIFp ↓ | 0.184 | 0.156 | 0.390 | 0.143 | 0.247 | 0.229 | 0.168 | **0.111** |
| | | CLIP ↓ | 0.728 | 0.725 | 0.745 | **0.719** | 0.733 | 0.733 | 0.730 | 0.724 |
| | | LPIPS ↑ | 0.456 | 0.497 | 0.220 | 0.510 | 0.384 | 0.394 | 0.486 | **0.585** |
| FLUX.1 | SD1.4 | PSNR ↓ | 20.24 | 20.72 | 23.46 | 20.67 | 22.13 | 19.95 | 19.51 | **18.71** |
| | | SSIM ↓ | 0.634 | 0.671 | 0.806 | 0.549 | 0.741 | 0.580 | 0.585 | **0.407** |
| | | VIFp ↓ | 0.230 | 0.242 | 0.391 | 0.194 | 0.296 | 0.204 | 0.203 | **0.138** |
| | | CLIP ↓ | 0.810 | 0.815 | 0.832 | 0.818 | 0.826 | 0.807 | 0.793 | **0.792** |
| | | LPIPS ↑ | 0.384 | 0.379 | 0.245 | 0.463 | 0.311 | 0.425 | 0.437 | **0.567** |
| | SD3 | PSNR ↓ | 20.03 | 22.17 | 22.95 | 19.14 | 21.08 | 19.74 | 18.70 | **18.26** |
| | | SSIM ↓ | 0.612 | 0.708 | 0.791 | 0.459 | 0.703 | 0.533 | 0.568 | **0.446** |
| | | VIFp ↓ | 0.217 | 0.277 | 0.380 | **0.151** | 0.274 | 0.177 | 0.201 | 0.155 |
| | | CLIP ↓ | 0.807 | 0.826 | 0.833 | 0.803 | 0.822 | 0.796 | 0.787 | **0.780** |
| | | LPIPS ↑ | 0.401 | 0.335 | 0.245 | 0.498 | 0.348 | 0.469 | 0.437 | **0.523** |
| | FLUX.1 | PSNR ↓ | 20.04 | 21.12 | 24.92 | **17.68** | 19.62 | 21.34 | 19.86 | 18.65 |
| | | SSIM ↓ | 0.612 | 0.640 | 0.844 | 0.444 | 0.611 | 0.692 | 0.590 | **0.410** |
| | | VIFp ↓ | 0.213 | 0.193 | 0.454 | 0.147 | 0.239 | 0.271 | 0.224 | **0.141** |
| | | CLIP ↓ | 0.802 | 0.835 | 0.841 | **0.768** | 0.797 | 0.816 | 0.800 | 0.786 |
| | | LPIPS ↑ | 0.410 | 0.494 | 0.177 | 0.504 | 0.398 | 0.329 | 0.422 | **0.558** |

However, diffusion models are not limited to the KL-based encoders used by SD1.4 (Rombach et al., 2022), SD3 (Esser et al., 2024), and FLUX.1 (Labs, 2024). As summarized in Tab. 14, Kandinsky 2.2 (Razzhigaev et al., 2023) and Kandinsky 3 (Arkhipkin et al., 2024) instead employ structurally distinct MoVQ-based vector-quantized encoders with different channel dimensions. We therefore extend our black-box evaluation to these models. As shown in Tab. 15, our method consistently maintains the strongest overall transfer performance, especially in LPIPS (Zhang et al., 2018), across these heterogeneous encoder targets.

To verify that the observed transferability reflects shared representations rather than only downstream performance, we repeat the *clean*- and *shift*-PWCCA analyses from Sec. 3.3 across the KL and VQ encoders. As shown in Fig. 9a, *clean*-PWCCA generally favors early encoder stages, al-




(a) Clean-PWCCA between KL and VQ encoders. (b) Shift-PWCCA from KL to VQ encoders

Figure 9: Cross-family representation correspondence between KL and VQ encoders. (a) *Clean*-PWCCA measures correspondence between clean encoder representations. For each source encoder, results are averaged over target encoders from the opposite family. (b) *Shift*-PWCCA measures correspondence between perturbation-induced representation shifts, where attacks generated using SD1.4, SD3, and FLUX.1 are evaluated on Kandinsky 2.2 and Kandinsky 3. Shaded regions denote the population standard deviation across target encoders.

though SD1.4 exhibits a less consistent depth-wise trend. More importantly, *shift*-PWCCA in Fig. 9b exhibits an early-stage advantage in most cases, suggesting that perturbation-induced changes generally transfer more effectively at earlier representations. These results extend the early-stage transfer advantage beyond KL-based VAEs to VQ-based latent encoders. We emphasize, however, that this consistency across the evaluated architectures does not establish the same depth-wise behavior for all generative models or encoder designs.

Table 14: Model and latent-interface characteristics of the pipelines used in our controlled depth-wise analysis. Stage resolutions are measured at the final residual-unit output before the optional stage downsampler, using a $512 \times 512$ input. † denotes the latent scaling and shift factors applied at the VAE–diffusion-model interface; these factors do not affect the intermediate encoder features used in our blockwise analysis.

| Property | SD1.4 | SD3 | FLUX.1 | Kandinsky 2.2 | Kandinsky 3 |
|---|---|---|---|---|---|
| Latent autoencoder | AutoencoderKL | AutoencoderKL | AutoencoderKL | VQModel (MoVQ) | VQModel (MoVQ) |
| Latent channels | 4 | 16 | 16 | 4 | 4 |
| Latent scaling factor† | 0.18215 | 1.5305 | 0.3611 | 0.18215 | 0.18215 |
| Latent shift factor† | 0 | 0.0609 | 0.1159 | – | – |
| Denoiser backbone | U-Net | MMDiT | Flux Transformer | U-Net | Kandinsky3 U-Net |
| Generative formulation | Latent diffusion | Rectified flow | Rectified flow | Latent diffusion | Latent diffusion |
| Analyzed stages | B0–B3 + B4 (Mid) | | | | |
| Stage widths (B0–B4) | 128/256/512/512/512 | 128/256/512/512/512 | 128/256/512/512/512 | 128/256/256/512/512 | 256/512/512/1024/1024 |
| Spatial resolution (B0–B4) | $512^2/256^2/128^2/64^2/64^2$ | | | | |

## D Qualitative Results

We provide qualitative comparisons of manipulation outputs generated from immunized images across diverse surrogate models and manipulation tasks, visually demonstrating the protection effectiveness of our method in Figs. 10–12. For each example, we evaluate the same source image across the indicated target models, with the clean output as a reference.

Figs. 13–18 extend the comparison to I2V generation. Each figure shows the source image and prompt above the generated sequences. Rows correspond to protection methods, and the five columns show frames sampled in temporal order from the start to the end of each video. Our method produces more pronounced changes in object appearance, edited content, or scene composition than the baselines.

## E Limitations and Discussion

We presented the unified image-immunization framework that jointly addresses black-box transferability and purification robustness. Through controlled depth-wise analyses, we identified early VAE encoder representations as a common attack surface for both properties. Extensive experiments

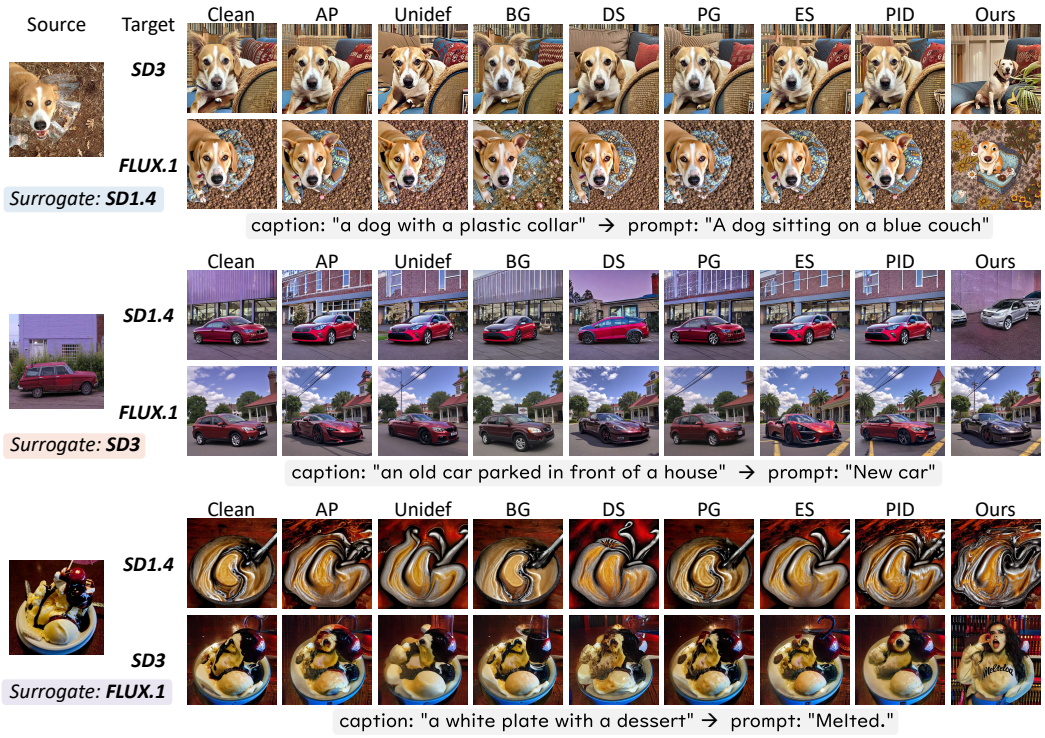



Figure 10: Qualitative image editing results across surrogate and target models. Each group shows the source image and outputs from two target models. Clean denotes the output from the unprotected image.

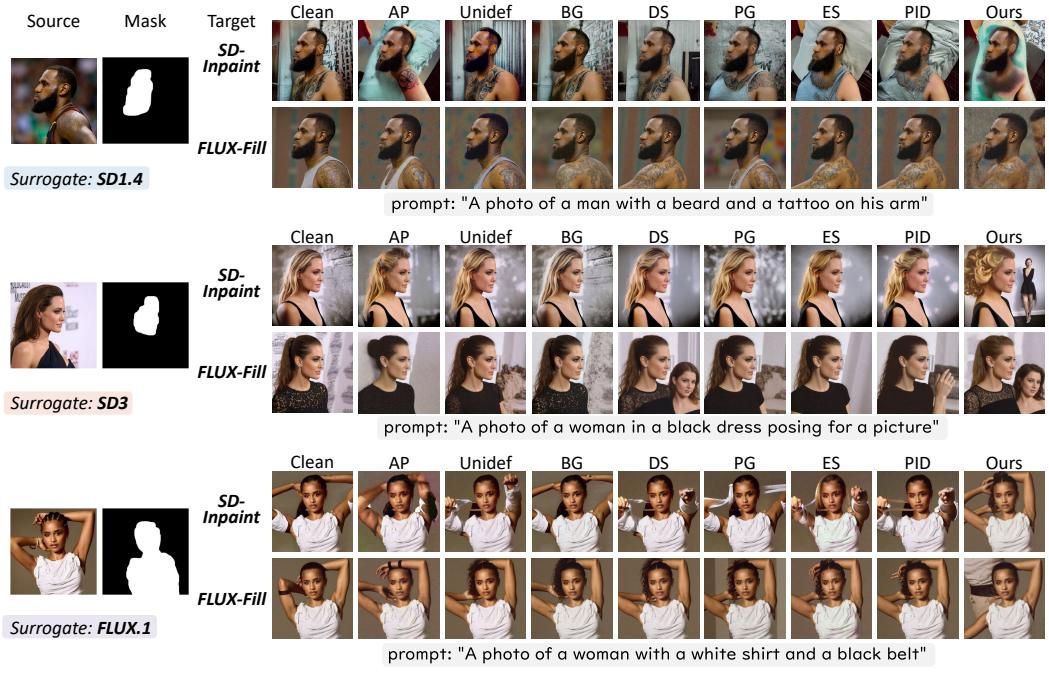



Figure 11: Qualitative inpainting results. The source image and editing mask are shown on the left, followed by outputs from SD-Inpaint and FLUX-Fill for each surrogate model.

Table 15: Transferability to unseen Kandinsky I2I models. PSNR and LPIPS compare the clean and protected editing outputs, while CLIP compares the original input and protected output.

| Target | Surrogate | Metric | AP | UniDef | BG | DS | PG | ES | PID | Ours |
|---|---|---|---|---|---|---|---|---|---|---|
| Kandinsky 2.2 | SD1.4 | PSNR ↓ | 21.86 | 22.19 | 25.33 | 22.02 | 22.23 | 20.67 | 21.99 | **20.57** |
| | | CLIP ↓ | 0.687 | 0.688 | 0.687 | 0.688 | **0.687** | 0.689 | 0.689 | 0.689 |
| | | LPIPS ↑ | 0.300 | 0.251 | 0.152 | 0.294 | 0.259 | 0.368 | 0.304 | **0.382** |
| | SD3 | PSNR ↓ | 22.00 | 23.70 | 26.76 | 21.79 | 21.70 | **20.78** | 22.00 | 20.83 |
| | | CLIP ↓ | 0.686 | 0.690 | 0.690 | 0.687 | **0.685** | 0.686 | 0.688 | 0.688 |
| | | LPIPS ↑ | 0.295 | 0.241 | 0.123 | 0.319 | 0.267 | 0.362 | 0.291 | **0.367** |
| | FLUX.1 | PSNR ↓ | 22.44 | 23.17 | 29.54 | 22.43 | 23.34 | 24.21 | 22.99 | **20.58** |
| | | CLIP ↓ | 0.688 | 0.686 | 0.688 | 0.688 | **0.684** | 0.689 | 0.685 | 0.690 |
| | | LPIPS ↑ | 0.273 | 0.245 | 0.082 | 0.296 | 0.203 | 0.206 | 0.266 | **0.376** |
| Kandinsky 3 | SD1.4 | PSNR ↓ | 24.77 | 24.93 | 27.88 | 23.77 | 24.56 | 23.20 | 24.59 | **23.02** |
| | | CLIP ↓ | 0.889 | 0.881 | 0.910 | 0.880 | 0.885 | 0.874 | 0.883 | **0.873** |
| | | LPIPS ↑ | 0.367 | 0.345 | 0.171 | 0.423 | 0.393 | 0.470 | 0.402 | **0.498** |
| | SD3 | PSNR ↓ | 25.00 | 27.05 | 28.64 | 23.41 | 24.91 | **23.22** | 24.56 | 23.62 |
| | | CLIP ↓ | 0.889 | 0.908 | 0.913 | 0.886 | 0.899 | **0.875** | 0.888 | 0.883 |
| | | LPIPS ↑ | 0.365 | 0.282 | 0.123 | **0.470** | 0.304 | 0.462 | 0.384 | 0.463 |
| | FLUX.1 | PSNR ↓ | 26.03 | 26.90 | 32.43 | 24.21 | 25.84 | 27.64 | 25.84 | **23.26** |
| | | CLIP ↓ | 0.899 | 0.904 | 0.917 | 0.886 | 0.908 | 0.909 | 0.897 | **0.885** |
| | | LPIPS ↑ | 0.319 | 0.290 | 0.056 | 0.439 | 0.231 | 0.244 | 0.338 | **0.482** |

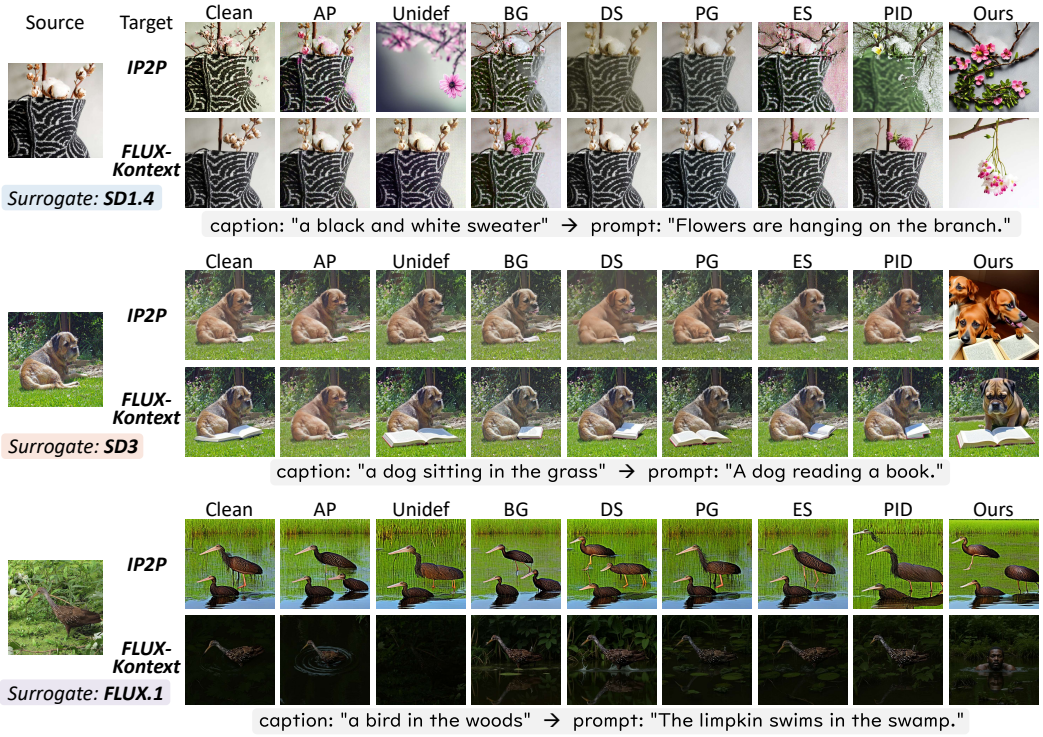



Figure 12: Qualitative instruction-guided editing results on IP2P and FLUX-Kontext.

covered U-Net- and DiT-based architectures, multiple manipulation tasks including image editing, inpainting, instruction-guided editing, and I2V generation, and diverse purification techniques.

Despite this broad evaluation, our experiments primarily considered publicly accessible models. The effectiveness of our method on commercial systems remains underexplored because their input preprocessing and model updates are generally undisclosed. Moreover, although our early-stage




Figure 13: Image-to-video qualitative results on DynamiCrafter using SD1.4 as the surrogate.

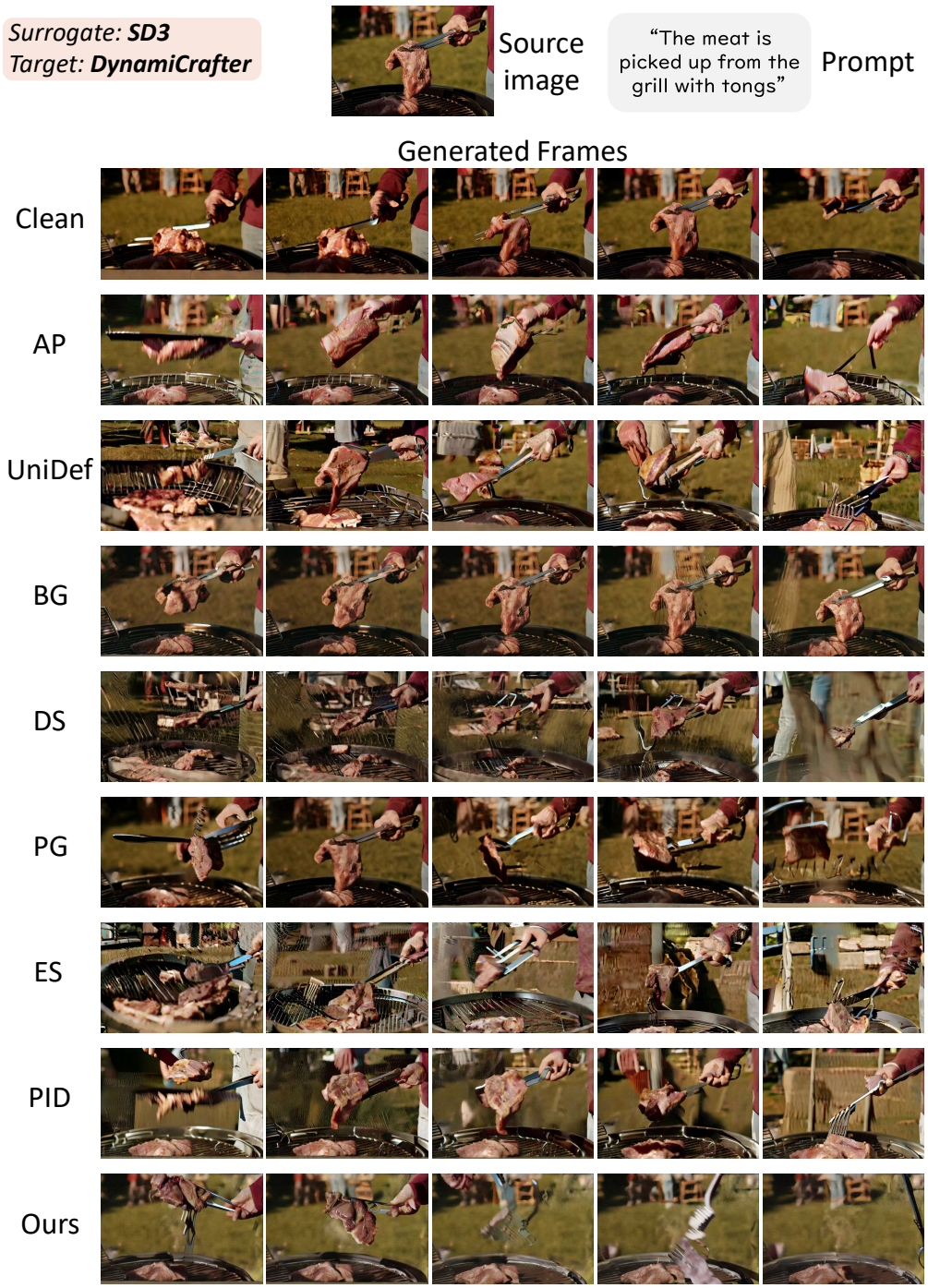



Figure 14: Image-to-video qualitative results on DynamiCrafter using SD3 as the surrogate.




Figure 15: Image-to-video qualitative results on DynamiCrafter using FLUX.1 as the surrogate.




Figure 16: Image-to-video qualitative results on LTX-Video using SD1.4 as the surrogate.

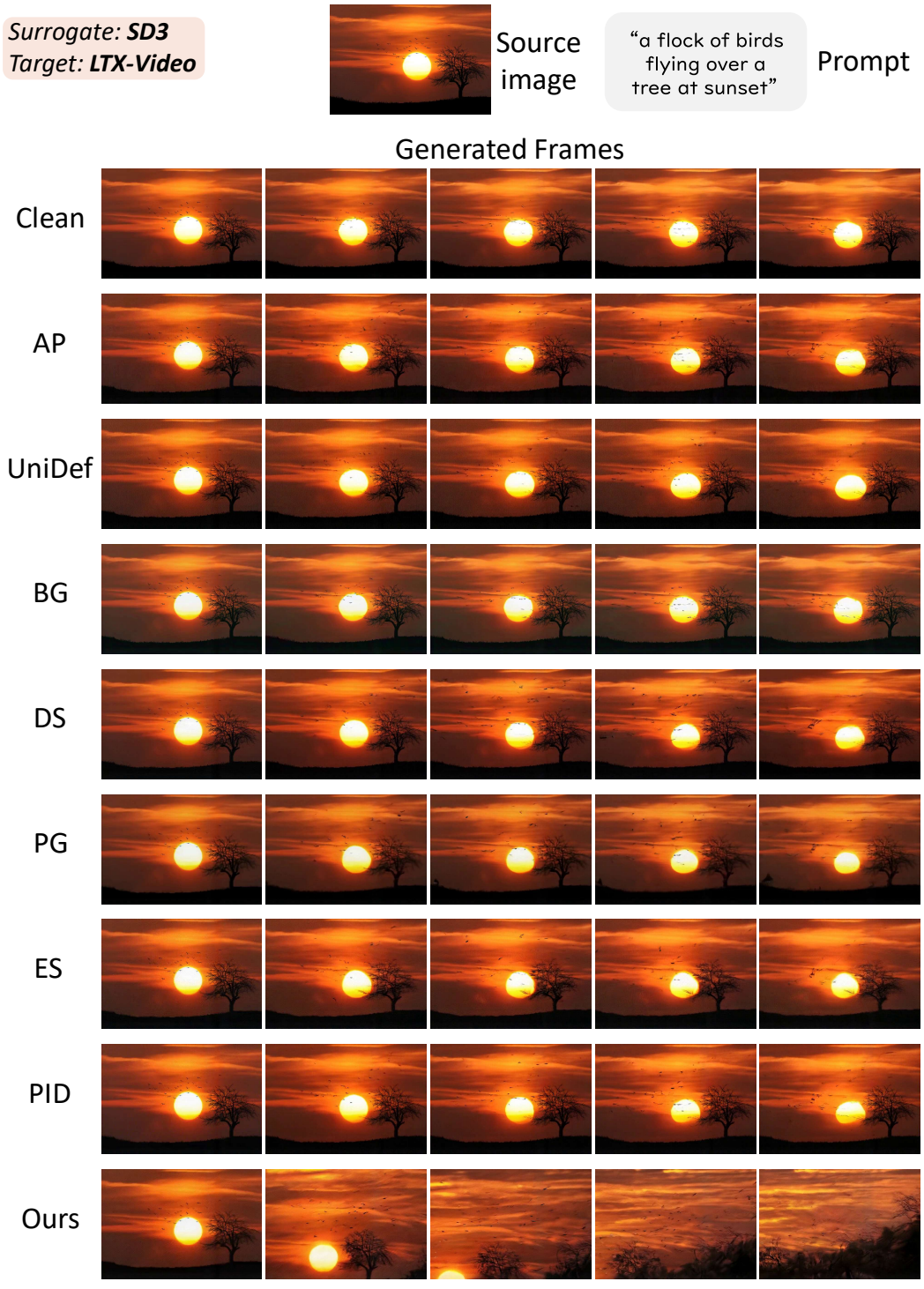



Figure 17: Image-to-video qualitative results on LTX-Video using SD3 as the surrogate.




Figure 18: Image-to-video qualitative results on LTX-Video using FLUX.1 as the surrogate.

Table 16: Purification robustness results for I2I. Values are LPIPS after purification; higher values indicate stronger protection.

| Target | Surrogate | Purification | PG | BG | DS | ES | PID | AP | UniDef | Ours |
|---|---|---|---|---|---|---|---|---|---|---|
| SD1.4 | SD1.4 | JPEG | 0.406 | 0.379 | 0.452 | 0.525 | 0.448 | 0.382 | 0.496 | **0.549** |
| | | AdvClean | 0.464 | 0.378 | 0.510 | 0.514 | 0.390 | 0.339 | 0.485 | **0.538** |
| | | Crop&Resize | 0.466 | 0.354 | 0.496 | 0.501 | 0.486 | 0.412 | 0.416 | **0.516** |
| | | GridPure | 0.452 | 0.441 | 0.469 | 0.463 | 0.345 | 0.363 | 0.481 | **0.487** |
| | SD3 | JPEG | 0.363 | 0.338 | 0.530 | 0.452 | 0.385 | 0.368 | 0.301 | **0.532** |
| | | AdvClean | 0.372 | 0.351 | 0.465 | 0.407 | 0.319 | 0.317 | 0.290 | **0.535** |
| | | Crop&Resize | 0.369 | 0.290 | 0.474 | 0.454 | 0.416 | 0.418 | 0.356 | **0.507** |
| | | GridPure | 0.410 | 0.407 | 0.460 | 0.394 | 0.343 | 0.347 | 0.352 | **0.473** |
| | FLUX.1 | JPEG | 0.319 | 0.303 | 0.483 | 0.321 | 0.372 | 0.356 | 0.300 | **0.535** |
| | | AdvClean | 0.330 | 0.333 | 0.448 | 0.331 | 0.332 | 0.316 | 0.319 | **0.537** |
| | | Crop&Resize | 0.326 | 0.242 | 0.460 | 0.399 | 0.420 | 0.424 | 0.401 | **0.510** |
| | | GridPure | 0.379 | 0.385 | 0.416 | 0.368 | 0.353 | 0.357 | 0.340 | **0.494** |
| SD3 | SD1.4 | JPEG | 0.365 | 0.329 | 0.473 | 0.483 | 0.455 | 0.419 | 0.396 | **0.586** |
| | | AdvClean | 0.405 | 0.363 | 0.530 | 0.428 | 0.363 | 0.338 | 0.357 | **0.579** |
| | | Crop&Resize | 0.404 | 0.280 | 0.498 | 0.453 | 0.450 | 0.425 | 0.381 | **0.541** |
| | | GridPure | 0.385 | 0.387 | 0.484 | 0.382 | 0.343 | 0.329 | 0.385 | **0.565** |
| | SD3 | JPEG | 0.415 | 0.322 | **0.596** | 0.556 | 0.478 | 0.460 | 0.395 | 0.561 |
| | | AdvClean | 0.435 | 0.348 | 0.557 | **0.565** | 0.439 | 0.404 | 0.359 | 0.563 |
| | | Crop&Resize | 0.412 | 0.286 | 0.533 | **0.549** | 0.472 | 0.486 | 0.402 | 0.525 |
| | | GridPure | 0.440 | 0.371 | **0.541** | 0.496 | 0.379 | 0.346 | 0.361 | 0.539 |
| | FLUX.1 | JPEG | 0.360 | 0.282 | 0.519 | 0.368 | 0.433 | 0.417 | 0.403 | **0.579** |
| | | AdvClean | 0.333 | 0.337 | 0.473 | 0.332 | 0.374 | 0.354 | 0.464 | **0.578** |
| | | Crop&Resize | 0.377 | 0.218 | 0.476 | 0.385 | 0.453 | 0.429 | 0.456 | **0.542** |
| | | GridPure | 0.386 | 0.366 | 0.473 | 0.357 | 0.351 | 0.353 | 0.446 | **0.563** |
| FLUX.1 | SD1.4 | JPEG | 0.283 | 0.302 | 0.390 | 0.440 | 0.407 | 0.351 | 0.354 | **0.547** |
| | | AdvClean | 0.273 | 0.301 | 0.425 | 0.370 | 0.313 | 0.268 | 0.290 | **0.545** |
| | | Crop&Resize | 0.303 | 0.253 | 0.368 | 0.412 | 0.406 | 0.390 | 0.347 | **0.520** |
| | | GridPure | 0.262 | 0.336 | 0.341 | 0.308 | 0.261 | 0.250 | 0.355 | **0.505** |
| | SD3 | JPEG | 0.343 | 0.284 | **0.521** | 0.451 | 0.408 | 0.371 | 0.316 | 0.512 |
| | | AdvClean | 0.331 | 0.283 | 0.452 | 0.392 | 0.287 | 0.294 | 0.265 | **0.513** |
| | | Crop&Resize | 0.331 | 0.238 | 0.451 | 0.441 | 0.435 | 0.401 | 0.312 | **0.483** |
| | | GridPure | 0.370 | 0.311 | 0.429 | 0.350 | 0.289 | 0.257 | 0.274 | **0.461** |
| | FLUX.1 | JPEG | 0.328 | 0.240 | 0.481 | 0.313 | 0.345 | 0.375 | 0.327 | **0.546** |
| | | AdvClean | 0.282 | 0.256 | 0.441 | 0.273 | 0.311 | 0.311 | 0.425 | **0.546** |
| | | Crop&Resize | 0.329 | 0.177 | 0.431 | 0.340 | 0.388 | 0.404 | 0.359 | **0.517** |
| | | GridPure | 0.327 | 0.299 | 0.380 | 0.288 | 0.288 | 0.283 | 0.369 | **0.520** |

optimization requires fewer computational resources than existing encoder-based methods, it still consumes approximately 3 GiB of GPU memory and 6 seconds, which may limit deployment on resource-constrained devices. Future work will extend the evaluation to commercial and non-VAE pipelines while developing more memory-efficient optimization.

## F Reproducibility

To facilitate reproducibility, Sec. 5 documents the evaluated datasets, diffusion models, manipulation tasks, purification methods, optimization settings, perturbation constraints, and hyperparameters. The main paper describes the evaluation protocols, while Algorithm 1 in the appendix provides implementation pseudocode. The appendix also reports stability results across random seeds. Executable code is included in the supplementary materials, and upon acceptance, we will publicly release the complete source code, evaluation scripts, and configuration files required to reproduce our experiments.

Table 17: Purification robustness results for inpainting. Values are LPIPS after purification; higher values indicate stronger protection.

| Target | Surrogate | Purification | AP | UniDef | BG | DS | PG | ES | PID | Ours |
|---|---|---|---|---|---|---|---|---|---|---|
| SD-Inpaint | SD1.4 | JPEG | 0.388 | 0.432 | 0.231 | 0.431 | 0.352 | 0.454 | 0.439 | **0.537** |
| | | AdvClean | 0.393 | 0.386 | 0.377 | 0.473 | 0.419 | 0.454 | 0.433 | **0.526** |
| | | Crop&Resize | 0.449 | 0.456 | 0.178 | 0.462 | 0.458 | 0.501 | 0.508 | **0.522** |
| | | GridPure | 0.313 | 0.377 | 0.332 | 0.418 | 0.367 | 0.407 | 0.333 | **0.487** |
| | SD3 | JPEG | 0.388 | 0.321 | 0.238 | **0.519** | 0.355 | 0.470 | 0.425 | 0.493 |
| | | AdvClean | 0.386 | 0.303 | 0.385 | 0.460 | 0.368 | 0.448 | 0.390 | **0.496** |
| | | Crop&Resize | 0.437 | 0.362 | 0.201 | 0.490 | 0.392 | **0.527** | 0.513 | 0.497 |
| | | GridPure | 0.339 | 0.318 | 0.334 | 0.439 | 0.373 | 0.391 | 0.368 | **0.459** |
| | FLUX.1 | JPEG | 0.358 | 0.310 | 0.193 | 0.487 | 0.302 | 0.300 | 0.389 | **0.538** |
| | | AdvClean | 0.362 | 0.327 | 0.364 | 0.409 | 0.279 | 0.362 | 0.357 | **0.526** |
| | | Crop&Resize | 0.398 | 0.367 | 0.112 | 0.455 | 0.352 | 0.341 | 0.453 | **0.516** |
| | | GridPure | 0.307 | 0.321 | 0.316 | 0.383 | 0.312 | 0.311 | 0.302 | **0.506** |
| FLUX-Fill | SD1.4 | JPEG | 0.380 | 0.363 | 0.244 | 0.427 | 0.328 | 0.428 | 0.433 | **0.555** |
| | | AdvClean | 0.350 | 0.302 | 0.322 | 0.427 | 0.304 | 0.419 | 0.395 | **0.539** |
| | | Crop&Resize | 0.383 | 0.385 | 0.169 | 0.400 | 0.324 | 0.422 | 0.435 | **0.509** |
| | | GridPure | 0.238 | 0.308 | 0.286 | 0.349 | 0.247 | 0.280 | 0.265 | **0.478** |
| | SD3 | JPEG | 0.385 | 0.317 | 0.230 | **0.554** | 0.337 | 0.447 | 0.414 | 0.528 |
| | | AdvClean | 0.351 | 0.286 | 0.311 | 0.472 | 0.344 | 0.422 | 0.366 | **0.525** |
| | | Crop&Resize | 0.401 | 0.308 | 0.181 | 0.473 | 0.325 | 0.461 | 0.446 | **0.498** |
| | | GridPure | 0.252 | 0.238 | 0.258 | 0.426 | 0.314 | 0.334 | 0.291 | **0.496** |
| | FLUX.1 | JPEG | 0.376 | 0.295 | 0.194 | 0.485 | 0.303 | 0.291 | 0.379 | **0.556** |
| | | AdvClean | 0.340 | 0.320 | 0.285 | 0.417 | 0.273 | 0.303 | 0.346 | **0.545** |
| | | Crop&Resize | 0.377 | 0.322 | 0.113 | 0.422 | 0.296 | 0.310 | 0.406 | **0.510** |
| | | GridPure | 0.251 | 0.270 | 0.241 | 0.371 | 0.242 | 0.244 | 0.228 | **0.498** |

Table 18: Purification robustness results for instruction editing. Values are LPIPS after purification; higher values indicate stronger protection.

| Target | Surrogate | Purification | AP | UniDef | BG | DS | PG | ES | PID | Ours |
|---|---|---|---|---|---|---|---|---|---|---|
| IP2P | SD1.4 | JPEG | 0.393 | 0.474 | 0.318 | 0.416 | 0.381 | **0.492** | 0.421 | 0.488 |
| | | AdvClean | 0.361 | 0.459 | 0.313 | 0.468 | 0.389 | **0.506** | 0.379 | 0.494 |
| | | Crop&Resize | 0.430 | 0.447 | 0.261 | 0.430 | 0.425 | **0.507** | 0.450 | 0.483 |
| | | GridPure | 0.319 | 0.426 | 0.360 | 0.425 | 0.368 | 0.444 | 0.348 | **0.462** |
| | SD3 | JPEG | 0.365 | 0.298 | 0.264 | **0.484** | 0.341 | 0.451 | 0.409 | 0.453 |
| | | AdvClean | 0.330 | 0.289 | 0.271 | 0.447 | 0.348 | 0.427 | 0.356 | **0.462** |
| | | Crop&Resize | 0.408 | 0.314 | 0.226 | 0.450 | 0.341 | **0.482** | 0.444 | 0.463 |
| | | GridPure | 0.313 | 0.292 | 0.315 | 0.422 | 0.374 | 0.407 | 0.353 | **0.426** |
| | FLUX.1 | JPEG | 0.341 | 0.321 | 0.231 | 0.424 | 0.284 | 0.308 | 0.365 | **0.494** |
| | | AdvClean | 0.313 | 0.341 | 0.260 | 0.398 | 0.282 | 0.297 | 0.324 | **0.490** |
| | | Crop&Resize | 0.375 | 0.357 | 0.140 | 0.418 | 0.290 | 0.293 | 0.393 | **0.463** |
| | | GridPure | 0.300 | 0.341 | 0.296 | 0.373 | 0.321 | 0.296 | 0.320 | **0.476** |
| FLUX-Kontext | SD1.4 | JPEG | 0.315 | 0.370 | 0.278 | 0.348 | 0.325 | 0.383 | 0.351 | **0.427** |
| | | AdvClean | 0.313 | 0.345 | 0.263 | 0.356 | 0.336 | 0.359 | 0.323 | **0.419** |
| | | Crop&Resize | 0.345 | 0.367 | 0.238 | 0.365 | 0.361 | 0.377 | 0.353 | **0.405** |
| | | GridPure | 0.286 | 0.364 | 0.311 | 0.356 | 0.310 | 0.370 | 0.306 | **0.405** |
| | SD3 | JPEG | 0.344 | 0.284 | 0.244 | **0.397** | 0.318 | 0.391 | 0.350 | 0.394 |
| | | AdvClean | 0.321 | 0.263 | 0.236 | 0.349 | 0.308 | 0.379 | 0.313 | **0.390** |
| | | Crop&Resize | 0.359 | 0.269 | 0.208 | 0.346 | 0.316 | **0.394** | 0.359 | 0.359 |
| | | GridPure | 0.293 | 0.255 | 0.267 | 0.342 | 0.324 | 0.360 | 0.310 | **0.369** |
| | FLUX.1 | JPEG | 0.319 | 0.326 | 0.219 | 0.356 | 0.268 | 0.289 | 0.309 | **0.432** |
| | | AdvClean | 0.304 | 0.323 | 0.223 | 0.326 | 0.282 | 0.279 | 0.301 | **0.425** |
| | | Crop&Resize | 0.332 | 0.320 | 0.152 | 0.322 | 0.285 | 0.256 | 0.314 | **0.401** |
| | | GridPure | 0.280 | 0.329 | 0.264 | 0.340 | 0.288 | 0.266 | 0.280 | **0.433** |

Table 19: Purification robustness on unseen I2V models. Values are averaged over the Anti-Prompt metrics, and lower values indicate stronger protection.

| Target | Surrogate | Purification | AP | UniDef | BG | DS | PG | ES | PID | Ours |
|---|---|---|---|---|---|---|---|---|---|---|
| DynamiCrafter | SD1.4 | JPEG | 3.44 | 3.03 | 3.88 | 3.38 | 3.30 | **2.88** | 3.34 | 3.08 |
| | | AdvClean | 3.87 | 3.67 | 4.23 | **3.18** | 3.84 | 3.28 | 3.67 | **3.18** |
| | | Crop&Resize | 2.95 | 2.85 | 3.84 | 3.23 | 3.17 | **2.73** | 2.95 | 2.84 |
| | | GridPure | 3.80 | 3.17 | 4.49 | 3.23 | 3.49 | 3.76 | 3.20 | **3.12** |
| | SD3 | JPEG | 3.98 | 3.63 | 3.90 | 3.16 | 3.40 | **2.60** | 3.07 | 3.08 |
| | | AdvClean | 3.95 | 3.94 | 4.38 | 3.37 | 3.87 | **2.70** | 3.12 | 3.27 |
| | | Crop&Resize | 3.60 | 3.18 | 4.03 | 3.22 | 3.19 | **2.38** | 2.87 | 2.83 |
| | | GridPure | 3.81 | 3.80 | 4.38 | 3.19 | 3.83 | 3.38 | **2.99** | 3.12 |
| | FLUX.1 | JPEG | 3.12 | 3.39 | 4.36 | 3.36 | 3.74 | 3.46 | 3.20 | **3.10** |
| | | AdvClean | 3.49 | 3.55 | 4.54 | 3.73 | 4.04 | 3.49 | 3.54 | **3.20** |
| | | Crop&Resize | **2.71** | 2.85 | 4.08 | 3.44 | 3.27 | 3.00 | 2.95 | 2.93 |
| | | GridPure | 4.07 | 3.39 | 4.58 | 3.64 | 3.91 | 4.25 | 3.18 | **3.05** |
| LTX-Video | SD1.4 | JPEG | 2.79 | 2.77 | 2.88 | 2.85 | 2.88 | 3.02 | **2.69** | 2.87 |
| | | AdvClean | 3.40 | 3.15 | 3.53 | 3.28 | 3.34 | **2.99** | 3.30 | 3.01 |
| | | Crop&Resize | 3.14 | 2.95 | 3.50 | **2.83** | 2.94 | 2.86 | 2.92 | 2.97 |
| | | GridPure | 3.36 | 3.08 | 3.35 | 3.13 | 3.16 | 3.40 | 3.06 | **3.00** |
| | SD3 | JPEG | 2.76 | 2.86 | 2.86 | 3.00 | 2.83 | 2.92 | 2.92 | **2.73** |
| | | AdvClean | 3.31 | 3.71 | 3.49 | 3.24 | 3.30 | 2.93 | 3.22 | **2.79** |
| | | Crop&Resize | 3.10 | 3.48 | 3.40 | 3.16 | 3.09 | **2.72** | 3.05 | 2.76 |
| | | GridPure | 3.65 | 3.62 | 3.35 | 3.20 | 3.25 | 3.38 | 3.18 | **2.99** |
| | FLUX.1 | JPEG | 2.75 | 2.77 | 2.98 | 3.02 | 2.96 | 2.77 | 2.81 | **2.72** |
| | | AdvClean | 3.38 | 3.24 | 3.85 | 3.41 | 3.48 | 3.41 | 3.38 | **2.93** |
| | | Crop&Resize | 3.08 | **2.78** | 3.52 | 2.87 | 3.13 | 3.13 | 3.10 | **2.78** |
| | | GridPure | 3.44 | 3.15 | 3.51 | 3.23 | 3.48 | 3.50 | 3.26 | **3.06** |