RVDGUARD: ROBUST PROTECTION AGAINST UNAUTHORIZED IMAGE-TO-VIDEO GENERATION VIA FIRST-STEP VELOCITY DEFLECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Image-to-video (I2V) generation models can animate a single photo into a realistic video, raising serious concerns about privacy violations and deceptive content creation. Existing proactive protection methods add imperceptible perturbations to images before release so that videos generated from them are degraded. However, we find that existing I2V protections suffer from two robustness limitations. On the input side, their perturbations are largely neutralized by common resizing, cropping, and compression. On the output side, the generated videos often remain visually clean, so an attacker can restore them or regenerate new videos from unaffected frames. To address these limitations, we propose RVDGuard, a robust protection framework that requires only a single denoiser evaluation per transformation at each optimization step. For input robustness, RVDGuard applies Expectation over Transformation, passing original and protected images through shared transformation instances and averaging the objective over them. For output robustness, RVDGuard deflects the direction of the first-step velocity predicted by the I2V model, maximizing a per-temporal-slice cosine dissimilarity between the predictions for protected and original images under a shared initial noise. Because the first update sets the course of the entire sampling trajectory, this deflection propagates through all subsequent steps and corrupts every generated frame. Experiments on CogVideoX-5B-I2V, Wan2.2-TI2V-5B, and Wan2.1-I2V-14B show that the VBench-I2V scores of videos generated from images protected by RVDGuard drop by 25.7% on average relative to those generated from original images, compared with 14.1% for the strongest baseline, and that RVDGuard remains effective under common image transformations and transfers to an unseen I2V model.

1 INTRODUCTION

Image-to-video (I2V) generation has advanced rapidly in the past two years. Built on latent diffusion and flow matching with diffusion Transformer backbones (Rombach et al., 2022; Peebles & Xie, 2023; Lipman et al., 2023; Liu et al., 2023), modern I2V models can animate a single still image into a high-fidelity, temporally coherent video that follows a text prompt (Blattmann et al., 2023; Yang et al., 2025; Kong et al., 2024; Wang et al., 2025a). The same capability, however, drastically lowers the barrier for misuse: a single photo collected from social media is now sufficient to fabricate a realistic video of a person doing things that never happened, raising serious concerns about privacy, copyright, and deceptive content (Chesney & Citron, 2019). To counter this threat, a line of work studies proactive image protection, which adds an imperceptible adversarial perturbation to an image before release so that generative models conditioned on it produce degraded outputs, and these methods report substantial degradation of the generated videos (Pang et al., 2024; Gui et al., 2025; Vu et al., 2026; Long et al., 2026; Song et al., 2026).

Despite this progress, we find that existing I2V protections are brittle in realistic deployment, and their robustness breaks down at both ends of the generation pipeline. **(1) Input-side robustness.** Perturbations are typically optimized on a fixed pixel grid and evaluated on the exact protected image. In practice, however, images are routinely resized, cropped, and re-compressed when uploaded to social platforms, and the attacker must further resize or crop them to match the native resolution of the I2V model. Such benign transformations misalign the carefully crafted perturbation and can

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.

largely neutralize its effect, echoing observations in image protection (Cao et al., 2023; Hönig et al., 2025). **(2) Output-side robustness.** Even when the protection survives, the videos generated from protected images often remain visually clean, as shown in Figure 1: existing objectives mainly induce semantic deviations, such as identity drift, background mismatch, or temporal inconsistency, while individual frames stay sharp and realistic. Such outputs still carry usable content. An attacker can keep the unaffected frames, repair the video with restoration models (Wang et al., 2025c), or simply take a clean intermediate frame, which is re-synthesized by the decoder and no longer contains the adversarial perturbation, as a new conditioning image to regenerate the video. Protection is therefore only meaningful if it corrupts the entire generated video rather than a subset of frames or attributes.

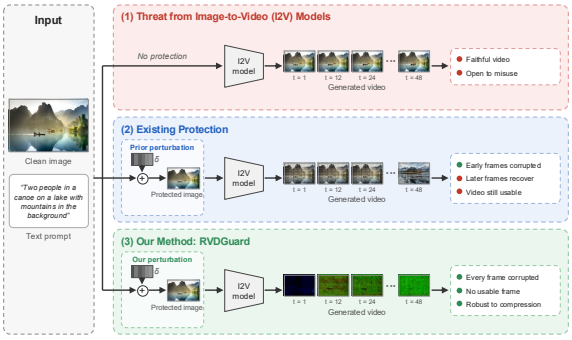


Figure 1: Illustration of the threat and protections. An I2V model faithfully animates the original image, videos generated from images with existing protections still contain usable frames, and RVDGuard corrupts every generated frame.

To address these limitations, we propose RVDGuard (**R**obust **V**elocity **D**eflection), a robust I2V protection framework that requires only a single denoiser evaluation per transformation at each iteration, as illustrated in Figure 2. To tackle the input-side limitation, RVDGuard adopts Expectation over Transformation (EoT) (Athalye et al., 2018): at each optimization step, the original and protected images are passed through the same sampled transformation instances, and the objective is averaged over these transformations, so that the perturbation remains effective under post-processing. To solve the output-side limitation, RVDGuard targets the first denoising step of the model's native sampler. Flow-matching I2V models predict a velocity that determines the direction of each latent update (Lipman et al., 2023; Liu et al., 2023), and the coarse layout and appearance of the generated content are largely settled in the earliest, noisiest steps (Choi et al., 2022). Given a shared initial noise, RVDGuard maximizes the cosine dissimilarity between the first-step velocities predicted for the protected and original images, computed separately for each latent temporal slice and averaged over time. Deflecting the direction of the very first update steers the sampling trajectory away from the clean one from its origin; this deviation is then propagated through all subsequent steps so that every generated frame, rather than a subset, is degraded.

We conduct extensive experiments on three representative open-source I2V models, CogVideoX-5B-I2V, Wan2.2-TI2V-5B, and Wan2.1-I2V-14B, using the VBench-I2V benchmark (Huang et al., 2026). Averaged over the three models and seven VBench-I2V dimensions, the scores of videos generated from images protected by RVDGuard drop by 25.7% relative to those generated from original images, compared with 14.1% for the strongest baseline Anti-Prompt (Song et al., 2026) and at most 13.5% for the other baselines. On Wan2.2-TI2V-5B, RVDGuard achieves the lowest values on all seven dimensions among all protections, reducing Subject Consistency from 0.963 to 0.661, whereas the best baseline only reaches 0.762. Under JPEG compression and resizing, the Subject Consistency of RVDGuard on Wan2.2-TI2V-5B stays at most 0.77, compared with at least 0.96 for videos generated from original images under the same transformations, and on the unseen DynamiCrafter model, the scores drop by 24.6% on average, compared with at most 8.1% for baseline protections.

Our main contributions are summarized as follows:

- We identify two robustness limitations of existing I2V image protection: an *input-side* limitation, where perturbations are neutralized by common resizing, cropping, and compression, and an *output-side* limitation, where generated videos remain visually clean and can be repaired or regenerated from intermediate frames.
- We propose RVDGuard, a robust protection framework that combines shared-instance EoT with a first-step velocity deflection objective. We further show empirically that optimizing

only the first denoising step provides strong protection at a fraction of the cost of joint multi-step optimization, whereas back-propagating through the full sampling trajectory largely fails.

- Extensive experiments on three state-of-the-art I2V models show that RVDGuard outperforms existing protection methods in average degradation, produces visibly corrupted videos, remains effective under common image transformations, and transfers to an unseen I2V model.

2 Related Work and Preliminaries

Image-to-Video Generation. Image-to-video (I2V) models generate videos conditioned on a reference image. Early video diffusion models extend U-Net-based latent diffusion (Rombach et al., 2022) with temporal layers, as in Stable Video Diffusion (Blattmann et al., 2023). Peebles & Xie (2023) introduce diffusion Transformers (DiTs), which process latent patches as token sequences and provide a scalable alternative to U-Net denoisers. Recent video models combine Transformer architectures with spatiotemporal compression and large-scale training. CogVideoX uses a three-dimensional variational autoencoder (VAE) and an expert Transformer to model video latents and integrate text features (Yang et al., 2025). Wan combines a causal video VAE with spatiotemporal attention and flow matching (Wang et al., 2025a). These advances enable realistic image animation while raising concerns about unauthorized image use, deceptive video creation, and privacy violations.

Image Protection against Generative Models. Early image protection methods optimize bounded input perturbations to resist unauthorized editing and content imitation. PhotoGuard, AdvDM, and MIST target image encoding or diffusion objectives (Salman et al., 2023; Liang et al., 2023; Liang & Wu, 2023). DiffusionGuard further focuses optimization on early denoising steps for efficient image editing protection (Choi et al., 2025). Video protection extends these objectives to image conditioning and temporal interactions. The prevention component of VGMShield disrupts encoder representations (Pang et al., 2024), while I2VGuard combines spatial, temporal, and diffusion objectives (Gui et al., 2025). Recent methods target more specific generation mechanisms. Anti-I2V attacks representations in the VAE and denoiser (Vu et al., 2026). Immune2V combines temporally balanced latent divergence with alignment to a collapse-inducing trajectory (Long et al., 2026). Anti-Prompt suppresses text-dependent interactions while strengthening visual pathways (Song et al., 2026). These methods are typically evaluated on the exact protected image and mainly induce semantic deviations in the generated videos, whereas RVDGuard targets robustness on both the input and output sides by deflecting the first-step velocity under shared transformations.

Conditional Video Generation. Starting from a Gaussian noise latent z_0, I2V models generate videos through conditional latent updates (Yang et al., 2025; Wang et al., 2025a):

$$v_k = v_\theta(z_k, t_k; c), \qquad z_{k+1} = \mathcal{S}_k(z_k, v_k). \tag{1}$$

Here, z_k and t_k denote the latent and model timestep at sampling step k, c contains the generation conditions, and θ denotes pretrained parameters. The native sampler $\mathcal{S}_k$ handles prediction conversion and any required sampling history. After N steps, the decoder produces the video $\widehat{V} = \mathcal{D}(z_N)$. Under the same initial noise, perturbing the input image can change the first velocity prediction and the resulting latent update, providing a direct signal for perturbation optimization.

3 Threat Model

Attacker. The attacker collects a publicly posted image without the owner's consent and uses an open-source I2V model to animate it into a fabricated video for purposes such as impersonation or misinformation. Before generation, the attacker may compress, resize, or crop the image, and after generation, the attacker may restore the video or regenerate a new video from any of its frames. Moreover, the attacker can choose among available I2V models.

Defender. The defender, either the image owner or a trusted third party acting on the owner's behalf, adds a bounded perturbation to the image before posting it. The protected image should

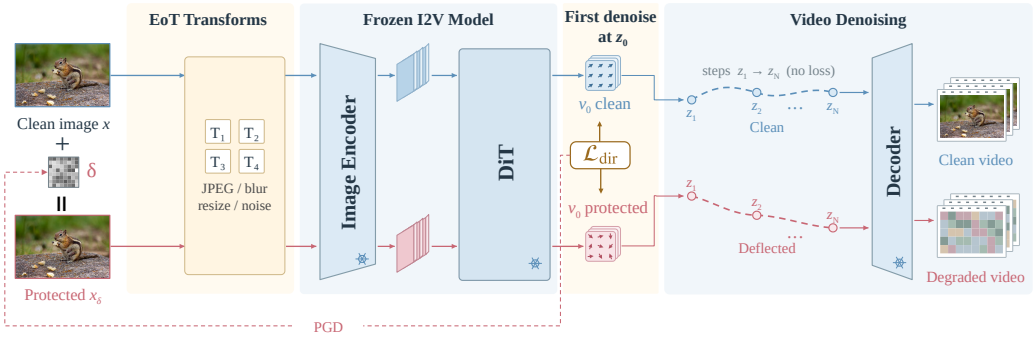


Figure 2: Overview of RVDGuard.

remain visually similar to the original image, while every video generated from it should be severely degraded, so that neither the video nor any of its frames serves as a faithful animation of the original image. The defender can use publicly available I2V models to optimize the perturbation but does not know which model the attacker will use. We therefore evaluate both the setting where the perturbation is optimized on the attacker's model and the setting where it is transferred to an unseen model in Section 5.5.

4 Methodology

Figure 2 gives an overview of RVDGuard. Given an input image, we optimize an imperceptible perturbation so that the velocity predicted at the first denoising step of an I2V model is deflected away from its clean counterpart, and we average this objective over common image transformations so that the protection survives post-processing. We describe each component below.

First-Step Velocity Deflection. I2V generation starts from Gaussian noise and progressively denoises it into a video conditioned on the input image; at each step, the model predicts a velocity that the sampler uses to update the noisy latent, as formalized in Equation (1). We perturb the input image so that the velocity predicted at the *first* step is deflected away from its clean counterpart, and let this deflection propagate through the remaining steps to degrade the generated video.

Let x be the original image and $x_\delta = x + \delta$ the protected image, with $\|\delta\|_\infty \le \epsilon$ and $x_\delta \in [0,1]^{H\times W\times 3}$. From the same initial noise latent z_0, the frozen denoiser predicts the first-step velocities $v_0 = v_\theta(z_0, t_0; x)$ and $v_0^\delta = v_\theta(z_0, t_0; x_\delta)$; the text prompt and sampler settings are shared and omitted from the notation. The velocity has the shape of the video latent, $v_0 \in \mathbb{R}^{C\times L\times h\times w}$, with C channels, L latent frames, and spatial size $h \times w$. We write $v_{0,b} \in \mathbb{R}^{C\times h\times w}$ for its b-th temporal slice and let $\mathcal{B}$ contain all slices except the first, which corresponds to the conditioning image. The deflection score on slice b is the cosine dissimilarity

$$s_{0,b}(\delta) = 1 - \cos\left(v_{0,b}^\delta,\, v_{0,b}\right), \tag{2}$$

which measures how far the perturbation rotates the first-step prediction, independent of its magnitude. We prefer it to an unnormalized or normalized ℓ_2 distance because directional deflection more consistently disrupts the downstream trajectory, as we show in Section 5.3.

Transformation-Averaged Optimization. Before reaching an I2V model, a protected image is often re-compressed, resized, or otherwise post-processed, which can weaken a perturbation optimized only on the raw image. Following EoT (Athalye et al., 2018), we average the deflection over a distribution $\mathcal{T}$ of common transformations: JPEG compression, Gaussian blur, resizing, additive Gaussian noise, and the identity. Each sampled $T \sim \mathcal{T}$ is applied to *both* images, i.e., the two velocities in Equation (2) are computed from $T(x_\delta)$ and $T(x)$, so that the pair differs only in the perturbation; comparing against the untransformed x would conflate the effect of δ with that of T itself. For JPEG, which is not differentiable, we use the real compression in the forward pass and a straight-through estimator (Bengio et al., 2013) in the backward pass, so that δ is optimized against actual compression artifacts rather than a differentiable approximation.

Overall Objective. Writing $s_{0,b}(\delta; T)$ for the score computed on the transformed pair, we maximize

$$\mathcal{L}_{\text{dir}}(\delta) = \mathbb{E}_{T\sim\mathcal{T}}\left[\frac{1}{|\mathcal{B}|}\sum_{b\in\mathcal{B}} s_{0,b}(\delta; T)\right] \tag{3}$$

over the feasible set of δ with projected gradient ascent (PGD) (Madry et al., 2018), estimating the expectation with K transformations sampled at every iteration. The initial noise z_0 is drawn once and kept fixed throughout the optimization, and gradients reach δ through the frozen image encoder and a single first-step denoiser evaluation per transformation, without unrolling the sampler. We use a uniform average over slices and transformations because emphasizing the weakest-responding transformations with CVaR yields weaker protection, while the aggregation over slices has little effect, as shown in Section 5.3. Transformation ranges and PGD hyperparameters are given in Appendix A.

Why Optimize Only the First Step. As Equation (1) shows, the generated video aggregates the updates of all denoising steps, so a natural alternative is to optimize the perturbation over multiple steps. However, optimizing a later step k requires running the sampler for k steps to reach z_k, and each additional step adds a denoiser evaluation to every iteration. As shown in Appendix B, jointly optimizing the first ten steps further lowers I2V Subject and Subject Consistency at ten times the cost but degrades Aesthetic Quality and Imaging Quality less than optimizing the first step alone, and back-propagating through the unrolled sampler largely fails, with I2V Subject remaining at 0.900 compared with 0.659 for the first step alone. Moreover, injecting the velocity difference induced by the protected image into the first step alone already degrades the whole video, reducing I2V Subject on Wan2.2-TI2V-5B to 0.705. We therefore optimize only the first denoising step, which provides strong protection with a single denoiser evaluation per transformation.

5 Experiments

5.1 Experimental Setup

Models. We evaluate on three representative open-source I2V models, CogVideoX-5B-I2V (Yang et al., 2025), Wan2.2-TI2V-5B (Wang et al., 2025a), and Wan2.1-I2V-14B (Wang et al., 2025a), which differ in model scale and generation configuration. CogVideoX-5B-I2V and Wan2.2-TI2V-5B generate 49-frame videos, while Wan2.1-I2V-14B generates 17-frame videos, allowing us to examine protection across different frame counts and architectures. Video generation uses each model's official sampler with seed 42, which is also used to draw z_0 during optimization; Appendix G evaluates a generation seed never used during optimization. Resolutions, sampling steps, and guidance scales follow the official settings and are listed in Appendix A. We further use DynamiCrafter (Xing et al., 2024) as an unseen model to evaluate transferability in Section 5.5.

Datasets. We evaluate on the VBench Image-to-Video dataset (VBench-I2V) (Huang et al., 2026), which consists of 355 images in total. We use the official text prompts paired with each image in the dataset for evaluation, without additional text prompts, to ensure evaluation consistency.

Baselines and Implementation Details. We compare our method against several image protection methods designed for diffusion models, including PhotoGuard (Salman et al., 2023), AntiPrompt (Song et al., 2026), and Immune2V (Long et al., 2026), using their official implementations. Since Anti-I2V (Vu et al., 2026) and I2VGuard (Gui et al., 2025) do not provide public implementations, we reimplement them following the original papers. For a fair comparison, our method and all baselines adopt PGD-based optimization with unified hyperparameters, where the perturbation budget in RGB space is set to $\epsilon = 16/255$, the step size to $\alpha = 1/255$, and the number of iterations to 200. The only exception is the perturbation budget of Immune2V, for which we keep its native setting of $\epsilon = 64/255$ following the original paper, which favors Immune2V with a budget four times that of the other methods. Unless otherwise stated, each protection is optimized and evaluated on the same I2V model, and Section 5.5 evaluates the setting where the attacker's model is unknown. All experiments are conducted on four NVIDIA H100 GPUs.

Metrics. We measure video quality with the official VBench-I2V implementation across seven dimensions, namely I2V Subject, I2V Background, Subject Consistency, Background Consistency, Aesthetic Quality, Imaging Quality, and Motion Smoothness, where lower values indicate

Table 1: Quantitative comparisons with baseline protections against various I2V models. ↓ indicates that a lower value of the metric signifies poorer video quality and thus better protection.

Method	I2V Subject ↓	I2V Background ↓	Subject Consistency ↓	Background Consistency ↓	Aesthetic Quality ↓	Imaging Quality ↓	Motion Smoothness ↓
Wan2.1-I2V-14B							
Clean	0.975	0.980	0.962	0.962	0.657	0.699	0.979
PhotoGuard	0.897	0.921	0.916	0.941	0.439	0.638	0.972
Anti-Prompt	0.782	0.853	0.776	0.879	0.439	0.596	**0.954**
Immune2V	0.802	0.843	0.844	0.907	0.436	0.650	0.961
Anti-I2V	0.920	0.947	0.936	0.952	0.604	0.662	0.978
I2VGuard	0.752	0.813	0.780	0.902	0.412	**0.386**	0.981
RVDGuard	**0.685**	**0.730**	**0.548**	**0.774**	**0.359**	0.447	0.965
Wan2.2-TI2V-5B							
Clean	0.985	0.989	0.963	0.967	0.657	0.716	0.987
PhotoGuard	0.935	0.948	0.904	0.940	0.535	0.653	0.980
Anti-Prompt	0.915	0.931	0.762	0.870	0.483	0.580	0.977
Immune2V	0.844	0.875	0.867	0.925	0.454	0.674	0.978
Anti-I2V	0.967	0.972	0.939	0.956	0.613	0.664	0.979
I2VGuard	0.938	0.950	0.898	0.935	0.579	0.621	0.981
RVDGuard	**0.828**	**0.850**	**0.661**	**0.817**	**0.452**	**0.549**	**0.976**
CogVideoX-5B-I2V							
Clean	0.981	0.983	0.964	0.963	0.608	0.706	0.984
PhotoGuard	0.877	0.899	0.922	0.940	0.446	0.612	0.973
Anti-Prompt	0.865	0.895	0.801	0.876	0.477	0.535	**0.965**
Immune2V	**0.689**	**0.733**	0.906	0.928	0.341	0.679	0.974
Anti-I2V	0.919	0.946	0.938	0.950	0.596	0.646	0.976
I2VGuard	0.889	0.916	0.884	0.921	0.504	0.562	0.975
RVDGuard	0.739	0.772	**0.583**	**0.791**	**0.329**	**0.293**	0.991

stronger protection. We denote videos generated from original images as Clean. We further report LPIPS (Zhang et al., 2018) and PSNR between the protected image and the original image to measure imperceptibility. To summarize the seven dimensions, we also report the relative drop of each score with respect to Clean, averaged over dimensions and models.

5.2 Protection Performance

Overall Results. We compare the protection performance of RVDGuard with baseline protections against three I2V models, with the quantitative results shown in Table 1. RVDGuard outperforms baseline protections on almost all metrics. It achieves the lowest values on all seven metrics on Wan2.2-TI2V-5B, as well as the lowest Subject Consistency, Background Consistency, and Aesthetic Quality on all three I2V models. Compared with Anti-Prompt, the strongest baseline, RVDGuard further reduces Subject Consistency by 0.228, 0.101, and 0.218 on Wan2.1-I2V-14B, Wan2.2-TI2V-5B, and CogVideoX-5B-I2V, respectively, and further reduces Imaging Quality by 0.242 on CogVideoX-5B-I2V. These results indicate that RVDGuard does not merely make the generated videos deviate from the original image, but simultaneously destroys their temporal consistency and per-frame quality. Although Immune2V achieves lower I2V Subject and I2V Background on CogVideoX-5B-I2V, it uses a perturbation budget four times that of RVDGuard. The Motion Smoothness of RVDGuard on this model is slightly higher than that of clean image generation, since the generated frames degenerate into textures without trackable motion, with a VBench Dynamic Degree of 0.00. The qualitative comparisons in Figure 3 further show that videos generated from images protected by baselines remain clear and usable. PhotoGuard only introduces grid-like

Table 2: Imperceptibility of protected images.

Method	Wan2.1-I2V-14B		Wan2.2-TI2V-5B		CogVideoX-5B-I2V	
	PSNR ↑	LPIPS ↓	PSNR ↑	LPIPS ↓	PSNR ↑	LPIPS ↓
PhotoGuard	25.74	0.532	26.79	0.537	25.31	0.492
Anti-Prompt	29.03	0.366	**28.87**	0.513	**28.22**	0.447
Immune2V	15.44	0.696	17.03	0.687	14.89	0.702
Anti-I2V	27.27	0.474	28.57	0.440	27.04	0.459
I2VGuard	28.65	0.425	27.75	0.540	27.46	0.440
RVDGuard	**29.19**	**0.315**	26.87	**0.431**	27.06	**0.311**

textures, and I2VGuard only causes slight speckle artifacts. Anti-Prompt corrupts the background while the subject becomes clear again in later frames, and Immune2V corrupts the early frames while its later frames recover to clean content. In contrast, RVDGuard removes recognizable content from every frame. Even if an adversary restores a generated frame and uses it to regenerate the video, the regenerated video remains temporally inconsistent and deviates from the original image, with Subject Consistency of 0.762 and I2V Subject of 0.720 on CogVideoX-5B-I2V, whereas the Subject Consistency of baseline protections returns to above 0.95, as detailed in Appendix F.

Imperceptibility. We evaluate the visual quality of protected images relative to the original images with PSNR and LPIPS, with the results shown in Table 2. Under the same budget of $16/255$, RVDGuard achieves the lowest LPIPS on all three I2V models, namely 0.315, 0.431, and 0.311, indicating that its perturbation is the least perceptible. The PSNR of RVDGuard is comparable to that of baseline protections with the same budget, reaching the highest value of 29.19 dB on Wan2.1-I2V-14B and staying within 2.0 dB of the highest value on the other two models. In contrast, Immune2V uses a budget of $64/255$, resulting in a PSNR of at most 17.03 dB and an LPIPS above 0.68 on all three models, and its protected image exhibits obvious grid-like textures as shown in Figure 3. These results indicate that RVDGuard preserves the visual quality of protected images while achieving stronger protection.

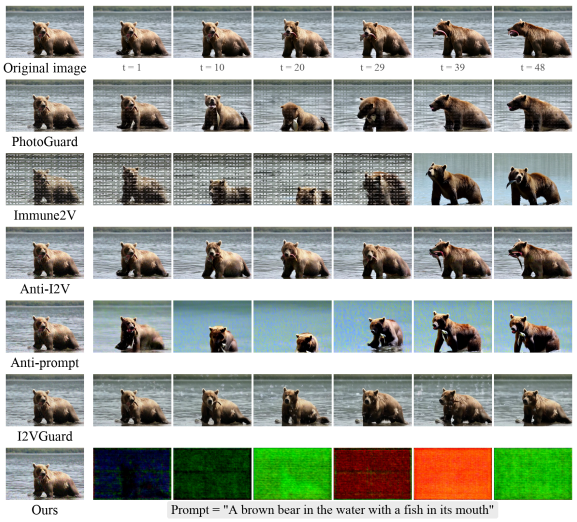


Figure 3: Qualitative comparisons with baseline protections on CogVideoX-5B-I2V.

5.3 Ablation Study

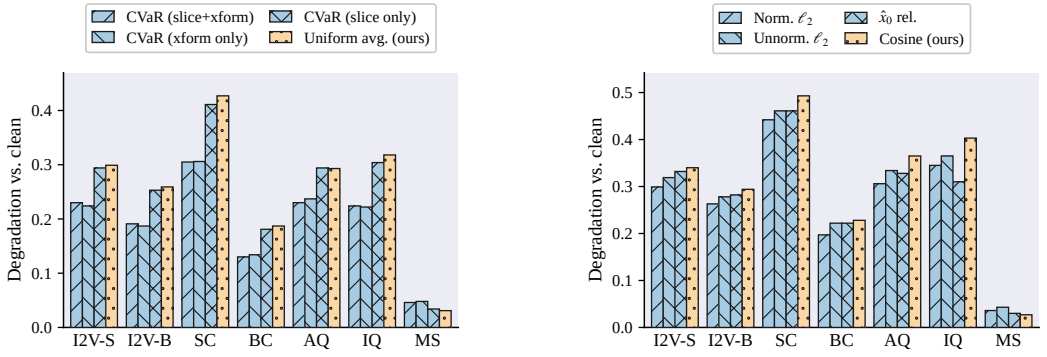
We conduct ablation studies on Wan2.2-TI2V-5B to justify three design choices in Section 4, namely the use of EoT, the aggregation strategy over temporal slices and input transformations, and the choice of deflection metric. The aggregation and metric ablations in Figure 4 report the degradation relative to Clean, so that higher bars indicate stronger protection, while the EoT ablation reports absolute scores.

Expectation over Transformation. Removing EoT makes the protection fragile to common transformations, as shown in Table 11 of Appendix E. Without EoT, Gaussian noise with $\sigma = 0.05$ and resizing with $\times 0.5$ raise Subject Consistency to 0.871 and 0.963, whereas with EoT they stay at 0.717 and 0.765.

Table 3: Quantitative results of RVDGuard against various transformations on Wan2.2-TI2V-5B.

Setting	Subject Consistency ↓	I2V Subject ↓	Aesthetic Quality ↓	Imaging Quality ↓
Transformation types used in optimization				
JPEG compression, $q = 50$	0.702	0.868	0.467	0.567
Gaussian noise, $\sigma = 0.05$	0.717	0.872	0.492	0.585
Resizing, $\times 0.5$	0.765	0.889	0.504	0.601
Gaussian blur, $\sigma = 1.0$	0.694	0.850	0.489	0.558
Transformation types unseen in optimization				
WebP compression, $q = 50$	0.681	0.892	0.494	0.640
Bit-depth reduction, 4-bit	0.661	0.832	0.453	0.549
Brightness, $\times 1.2$	0.717	0.870	0.453	0.577
Contrast, $\times 0.8$	0.696	0.853	0.476	0.596
Gamma correction, $\gamma = 0.8$	0.667	0.842	0.454	0.560
Center cropping, ratio 0.7	0.802	0.938	0.534	0.588
Median filtering, $k = 5$	0.821	0.937	0.555	0.618

Aggregation Strategy. We compare four ways of aggregating the deflection score in Equation (2) across temporal slices and transformations: emphasizing the weakest-responding slices and transformations with CVaR-style selection or uniformly averaging over each. Results are shown in Figure 4a. Uniform averaging over transformations consistently outperforms CVaR-style selection, while the aggregation strategy over temporal slices has comparatively little effect. We adopt uniform averaging over both as our default.



(a) Aggregation strategy.

(b) Deflection metric.

Figure 4: Ablation on two design choices, reported as degradation relative to the clean generation.

Deflection Metric. We compare the cosine-based deflection score in Equation (2) against three alternatives: normalized ℓ_2 distance, unnormalized ℓ_2 distance, and the corresponding relative deviation in the model's predicted clean latent $\hat{x}_0$. Results are shown in Figure 4b. The cosine-based formulation outperforms all ℓ_2-based alternatives on six of the seven dimensions, and we adopt it as our default deflection metric.

5.4 Protection Robustness

We evaluate the robustness of RVDGuard on Wan2.2-TI2V-5B and VBench-I2V against four transformation types used in optimization, namely JPEG compression, Gaussian noise, resizing, and Gaussian blur, as well as seven types unseen in optimization, namely WebP compression, bit-depth reduction, brightness, contrast, and gamma adjustment, center cropping, and median filtering. Each

Table 4: Quantitative results on transferability to DynamiCrafter.

Method	I2V Subject ↓	I2V Background ↓	Subject Consistency ↓	Background Consistency ↓	Aesthetic Quality ↓	Imaging Quality ↓	Motion Smoothness ↓
Clean	0.977	0.981	0.966	0.964	0.667	0.697	0.974
PhotoGuard	0.883	0.904	0.887	0.920	0.591	0.664	0.914
Anti-Prompt	0.913	0.929	0.910	0.929	0.616	0.669	0.932
Immune2V	0.872	0.894	0.879	0.916	0.592	0.658	0.910
Anti-I2V	0.922	0.934	0.924	0.937	0.630	0.673	0.938
I2VGuard	0.882	0.904	0.883	0.917	0.589	0.661	0.917
RVDGuard	**0.682**	**0.764**	**0.765**	**0.831**	**0.465**	**0.405**	**0.848**

transformation is applied to the protected image before video generation, and Table 3 reports representative settings on four metrics of Table 1. Full parameter sweeps, including the same transformations applied to original images, are shown in Figure 5 of Appendix D; note that Gaussian noise with $\sigma = 0.05$ and resizing by $\times 0.5$ already exceed the ranges sampled during optimization. Under the four transformation types used in optimization, the Subject Consistency of RVDGuard stays at most 0.765, far below the values of at least 0.96 obtained from original images under the same transformations. Transformations of unseen types that only alter pixel values, namely WebP compression, bit-depth reduction, and color adjustments, are equally ineffective at removing the protection, with Subject Consistency staying between 0.661 and 0.717. Cropping and median filtering, which change the spatial layout of the perturbation, weaken the protection the most and raise Subject Consistency to 0.802 and 0.821, although the protection is still partially retained. Moreover, removing the protection requires transformation strengths well beyond common usage, such as JPEG quality below 30, Gaussian noise with $\sigma \geq 0.08$, resizing below $\times 0.35$, or Gaussian blur with $\sigma \geq 2.0$, and most of these settings also noticeably degrade the videos generated from original images, as shown in Figure 5. These results validate the robustness of RVDGuard against common transformations.

5.5 Protection Transferability

We evaluate the protective effect of RVDGuard on an I2V model unseen during optimization. Specifically, we jointly optimize the perturbation against the three I2V models in Section 5.2, where the gradient of $\mathcal{L}_{\text{dir}}$ with respect to the perturbation is computed on each surrogate model, normalized by its mean absolute value, and averaged with equal weights before each update, with more details provided in Appendix A. We then apply the resulting protected images to DynamiCrafter (Xing et al., 2024), which adopts a U-Net denoiser and thus differs substantially from the three DiT-based surrogate models. For baseline protections, we directly apply their perturbations optimized on Wan2.2-TI2V-5B to DynamiCrafter. The quantitative results are shown in Table 4. Baseline protections exhibit limited effectiveness after transfer, with the lowest Imaging Quality of 0.658, only 0.039 lower than 0.697 of Clean. In contrast, RVDGuard achieves the lowest values on all seven metrics, reducing I2V Subject from 0.977 to 0.682 and Imaging Quality to 0.405, far below 0.872 and 0.658 for Immune2V, the best-performing baseline. These results demonstrate that the protection of RVDGuard transfers to DynamiCrafter despite its different architecture.

6 Conclusion

We identify two robustness limitations of existing I2V protections, namely perturbations that are neutralized by common image transformations on the input side and generated videos that remain usable on the output side. To address them, we propose RVDGuard, which deflects the first-step velocity of the I2V model under transformation instances shared between the original and protected images. Experiments on three I2V models show that RVDGuard corrupts every generated frame, outperforms existing protections on almost all metrics, remains effective under common image transformations, and transfers to an unseen model. We hope this work encourages evaluating I2V protections under realistic input processing and output restoration.

AI USE STATEMENT

In this work, we used generative AI tools for implementing methods, developing the theoretical analysis, providing feedback on research methodology and experiments, and translation: the code of RVDGuard, the reimplementation of Anti-I2V and I2VGuard, and the evaluation and plotting scripts were written by Claude Opus 5 under the direction of the authors, who specified the method and the experimental protocol; the derivation in Appendix C and the objective formulation were drafted by the authors and refined with feedback from GPT-6 Astra; the method and all experiments were designed by the authors, and GPT-6 Astra was asked to review the design for missing analyses; and parts of the paper drafted by the authors in their native language were translated into English by Claude Opus 5. We have not used generative AI tools to generate synthetic datasets, propose or refine hypotheses, clean or reformat datasets, or interpret results beyond drafting the textual description of results already analyzed by the authors, and qualitative and thematic data analysis is not applicable to this work. Additionally, we used Claude Opus 5 to edit the paper for readability and to format tables; all figures, the literature review, and the references were produced by the authors without generative AI tools. We have reviewed all AI-assisted work: all AI-written code was reviewed and tested by two authors, all reported numbers were cross-checked against experiment logs, all mathematical statements were independently verified by the authors, and all translated or edited text was checked against the experimental results. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work aims to protect personal images from unauthorized animation by I2V models. The perturbations studied here are defensive and do not enable new forms of misuse. All experiments use the public VBench-I2V images and open-source models, and no human subjects were involved.

REPRODUCIBILITY STATEMENT

We describe the objective and optimization of RVDGuard in Section 4, the experimental setup in Section 5.1, and the implementation details, including the transformation ranges, sampler settings, and ensemble optimization, in Appendix A. All models and the VBench-I2V dataset are publicly available, and we will release our code upon publication.

REFERENCES

Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. In Jennifer G. Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pp. 284–293. PMLR, 2018. URL http://proceedings.mlr.press/v80/athalye18b.html.

Yoshua Bengio, Nicholas Léonard, and Aaron C. Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *CoRR*, abs/1308.3432, 2013. URL http://arxiv.org/abs/1308.3432.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets. *CoRR*, abs/2311.15127, 2023. doi: 10.48550/ARXIV.2311.15127. URL https://doi.org/10.48550/arXiv.2311.15127.

Bochuan Cao, Changjiang Li, Ting Wang, Jinyuan Jia, Bo Li, and Jinghui Chen. IMPRESS: evaluating the resilience of imperceptible perturbations against unauthorized data usage in diffusion-based generative AI. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/222dda29587fbc2979ca99fd5ed00735-Abstract-Conference.html.

Bobby Chesney and Danielle Citron. Deep fakes: A looming challenge for privacy, democracy, and national security. *Calif. L. Rev.*, 107:1753, 2019.

Jooyoung Choi, Jungbeom Lee, Chaehun Shin, Sungwon Kim, Hyunwoo Kim, and Sungroh Yoon. Perception prioritized training of diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 11462–11471. IEEE, 2022. doi: 10.1109/CVPR52688.2022.01118. URL https://doi.org/10.1109/CVPR52688.2022.01118.

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 *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=9OfKxKoYNw.

Dongnan Gui, Xun Guo, Wengang Zhou, and Yan Lu. I2vguard: Safeguarding images against misuse in diffusion-based image-to-video models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 12595–12604. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.01175. URL https://openaccess.thecvf.com/content/CVPR2025/html/Gui_I2VGuard_Safeguarding_Images_against_Misuse_in_Diffusion-based_Image-to-Video_Models_CVPR_2025_paper.html.

Robert Hönig, Javier Rando, Nicholas Carlini, and Florian Tramèr. Adversarial perturbations cannot reliably protect artists from generative AI. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=yfW1x7uBS5.

Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, Yaohui Wang, Xinyuan Chen, Ying-Cong Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Vbench++: Comprehensive and versatile benchmark suite for video generative models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 48(3):3268–3285, 2026. doi: 10.1109/TPAMI.2025.3633890. URL https://doi.org/10.1109/TPAMI.2025.3633890.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, Kathrina Wu, Qin Lin, Junkun Yuan, Yanxin Long, Aladdin Wang, Andong Wang, Changlin Li, Duojun Huang, Fang Yang, Hao Tan, Hongmei Wang, Jacob Song, Jiawang Bai, Jianbing Wu, Jinbao Xue, Joey Wang, Kai Wang, Mengyang Liu, Pengyu Li, Shuai Li, Weiyan Wang, Wenqing Yu, Xinchi Deng, Yang Li, Yi Chen, Yutao Cui, Yuanbo Peng, Zhentao Yu, Zhiyu He, Zhiyong Xu, Zixiang Zhou, Zunnan Xu, Yangyu Tao, Qinglin Lu, Songtao Liu, Daquan Zhou, Hongfa Wang, Yong Yang, Di Wang, Yuhong Liu, Jie Jiang, and Caesar Zhong. Hunyuanvideo: A systematic framework for large video generative models. *CoRR*, abs/2412.03603, 2024. doi: 10.48550/ARXIV.2412.03603. URL https://doi.org/10.48550/arXiv.2412.03603.

Chumeng Liang and Xiaoyu Wu. Mist: Towards improved adversarial examples for diffusion models. *CoRR*, abs/2305.12683, 2023. doi: 10.48550/ARXIV.2305.12683. URL https://doi.org/10.48550/arXiv.2305.12683.

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 Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 20763–20786. PMLR, 2023. URL https://proceedings.mlr.press/v202/liang23g.html.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=XVjTT1nw5z.

Zeqian Long, Ozgur Kara, Haotian Xue, Yongxin Chen, and James M. Rehg. Immune2v: Image immunization against dual-stream image-to-video generation. *CoRR*, abs/2604.10837, 2026. doi: 10.48550/ARXIV.2604.10837. URL https://doi.org/10.48550/arXiv.2604.10837.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=rJzIBfZAb.

Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Animashree Anandkumar. Diffusion models for adversarial purification. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato (eds.), *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 16805–16827. PMLR, 2022. URL https://proceedings.mlr.press/v162/nie22a.html.

Yan Pang, Yang Zhang, and Tianhao Wang. Vgmshield: Mitigating misuse of video generative models. *CoRR*, abs/2402.13126, 2024. doi: 10.48550/ARXIV.2402.13126. URL https://doi.org/10.48550/arXiv.2402.13126.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pp. 4172–4182. IEEE, 2023. doi: 10.1109/ICCV51070.2023.00387. URL https://doi.org/10.1109/ICCV51070.2023.00387.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 10674–10685. IEEE, 2022. doi: 10.1109/CVPR52688.2022.01042. URL https://doi.org/10.1109/CVPR52688.2022.01042.

Hadi Salman, Alaa Khaddaj, Guillaume Leclerc, Andrew Ilyas, and Aleksander Madry. Raising the cost of malicious ai-powered image editing. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 29894–29918. PMLR, 2023. URL https://proceedings.mlr.press/v202/salman23a.html.

Yeonghwan Song, Chanhui Lee, Jinsoo Park, and Jeany Son. Anti-prompt: Image protection against text-guided image-to-video generation. *CoRR*, abs/2607.01499, 2026. doi: 10.48550/ARXIV.2607.01499. URL https://doi.org/10.48550/arXiv.2607.01499.

Duc Vu, Anh Nguyen, Chi Tran, and Anh Tran. Anti-i2v: Safeguarding your photos from malicious image-to-video generation. *CoRR*, abs/2603.24570, 2026. doi: 10.48550/ARXIV.2603.24570. URL https://doi.org/10.48550/arXiv.2603.24570.

Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Xiaofeng Meng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *CoRR*,

abs/2503.20314, 2025a. doi: 10.48550/ARXIV.2503.20314. URL https://doi.org/10.48550/arXiv.2503.20314.

Jianyi Wang, Shanchuan Lin, Zhijie Lin, Yuxi Ren, Meng Wei, Zongsheng Yue, Shangchen Zhou, Hao Chen, Yang Zhao, Ceyuan Yang, Xuefeng Xiao, Chen Change Loy, and Lu Jiang. Seedvr2: One-step video restoration via diffusion adversarial post-training. *CoRR*, abs/2506.05301, 2025b. doi: 10.48550/ARXIV.2506.05301. URL https://doi.org/10.48550/arXiv.2506.05301.

Jianyi Wang, Zhijie Lin, Meng Wei, Yang Zhao, Ceyuan Yang, Chen Change Loy, and Lu Jiang. Seedvr: Seeding infinity in diffusion transformer towards generic video restoration. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 2161–2172. Computer Vision Foundation / IEEE, 2025c. doi: 10.1109/CVPR52734.2025.00207. URL https://openaccess.thecvf.com/content/CVPR2025/html/Wang_SeedVR_Seeding_Infinity_in_Diffusion_Transformer_Towards_Generic_Video_Restoration_CVPR_2025_paper.html.

Jinbo Xing, Menghan Xia, Yong Zhang, Hao 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 Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XLVI*, volume 15104 of *Lecture Notes in Computer Science*, pp. 399–417. Springer, 2024. doi: 10.1007/978-3-031-72952-2_23. URL https://doi.org/10.1007/978-3-031-72952-2_23.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. Cogvideox: Text-to-video diffusion models with an expert transformer. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=LQzN6TRFg9.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 586–595. Computer Vision Foundation / IEEE Computer Society, 2018. doi: 10.1109/CVPR.2018.00068. URL http://openaccess.thecvf.com/content_cvpr_2018/html/Zhang_The_Unreasonable_Effectiveness_CVPR_2018_paper.html.

A Implementation Details

Generation Settings. Wan2.1-I2V-14B uses the 480×832 area bucket, which resolves to 832×464 for landscape images and 464×832 for portrait images, with 40 sampling steps and guidance scale 5. Wan2.2-TI2V-5B uses a 960×928 canvas from the 704×1280 area bucket, with 50 sampling steps and guidance scale 5. CogVideoX-5B-I2V uses a 720×480 canvas, with 50 sampling steps and guidance scale 6.

Optimization Settings. We use $\epsilon = 16/255$ and run PGD for 200 iterations with step size $\alpha = 1/255$. For every model, the initial noise z_0 is generated once with seed 42 and reused at all iterations. At each iteration, we sample $K = 4$ distinct transformations without replacement from the identity, JPEG compression, Gaussian blur, additive Gaussian noise, and resizing, and average the objective over them. The JPEG quality is an integer sampled uniformly from $\{50, \ldots, 90\}$, the Gaussian blur uses a kernel size of 7 with $\sigma \sim \mathcal{U}[0.4, 1.2]$, the additive Gaussian noise has $\sigma \sim \mathcal{U}[0.01, 0.04]$ for pixel values in $[0, 1]$, and the resizing downscales the image by a factor sampled from $\mathcal{U}[0.55, 0.85]$ with bilinear interpolation and upscales it back to the original resolution. The transformation distribution does not contain cropping.

Future Slices per Model. For all three models, the deflection is computed on the predicted video latent with its first temporal slice removed, and the models differ only in how the image enters the

network. Wan2.2-TI2V-5B conditions on the image by overwriting the first latent frame with the VAE latent of the image, so $\mathcal{B}$ excludes exactly this overwritten slot. Wan2.1-I2V-14B receives the image through a mask-plus-VAE conditioning input and CLIP features without overwriting the noisy latent, and the excluded slice is the one that the mask marks as the image frame, while both conditioning inputs remain in the network. CogVideoX-5B-I2V concatenates the image latent along the channel dimension, so the first frame of the noisy latent remains noise, and we remove the first frame of the predicted video latent rather than the concatenated image channels.

Single-Step Velocity Injection. At the injected step k, we evaluate the denoiser twice at the same latent z_k, once conditioned on the protected image and once on the original image. Let $\Delta v_0 = v_0^\delta - v_0$ denote the velocity difference at the first step and $L = \|\Delta v_0\|_2$ its magnitude. We pass

$$v_k + \frac{L}{\|v_k^\delta - v_k\|_2} \left(v_k^\delta - v_k\right) \tag{4}$$

to the sampler, so that the injected difference follows the direction induced by the protected image at step k and has the same magnitude as at the first step, which reduces to v_0^δ at $k = 0$. The latent z_k itself is not modified, and every other step is conditioned on the original image.

Ensemble Optimization for Transferability. For the transferability evaluation in Section 5.5, we optimize a single perturbation against Wan2.2-TI2V-5B, CogVideoX-5B-I2V, and Wan2.1-I2V-14B. The perturbation is initialized from the one optimized on Wan2.2-TI2V-5B for 200 iterations in Section 5.2, and is then further optimized against the three models for 100 iterations. At each iteration, we compute the gradient of $\mathcal{L}_{\text{dir}}$ with respect to δ on each model separately, normalize each gradient by its mean absolute value, and average the three normalized gradients with equal weights. We then update δ with sign momentum using the averaged gradient, with momentum factor $\mu = 0.9$, and project it back onto the ℓ_∞ ball of radius $\epsilon = 16/255$. When evaluating on DynamiCrafter, we use its 1024 version with the official sampling settings, and for all protections we bilinearly resize the perturbation to the input resolution of DynamiCrafter and add it to the resized original image, keeping the budget of $16/255$ for all protections except Immune2V, which keeps its native budget of $64/255$.

B Multi-Step Optimization Analysis

This section provides the experimental support for optimizing only the first denoising step, as discussed in Section 4. Unless otherwise noted, all experiments in this section are conducted on the first 100 images of VBench-I2V with the optimization settings of Appendix A.

Single-Step Injection. At a single denoising step k, we inject the velocity difference induced by the protected image with the same magnitude as at the first step, as described in Appendix A, while all other steps are left unchanged, and we sweep k across the sampling trajectory. Random direction replaces the injected difference at the first step with a random direction of the same magnitude, and Full trajectory applies the injection at every step. The results on Wan2.2-TI2V-5B and CogVideoX-5B-I2V are shown in Tables 5 and 6. On Wan2.2-TI2V-5B, injection at the first two steps strongly reduces I2V Subject and Subject Consistency, whereas injection at step 5 or later has no effect. On CogVideoX-5B-I2V, injection at later steps remains effective, although Subject Consistency and Aesthetic Quality are reduced the most by injection at the first step. The random direction barely changes the generated videos on either model, confirming that the degradation comes from the direction induced by the protected image rather than from the magnitude of the change. On both models, injection at the first step alone reduces I2V Subject, Subject Consistency, and Aesthetic Quality to a level comparable to injection at every step, which motivates optimizing only the first step.

Directional Conflict across Steps. For a latent z_k at step k of the clean sampling trajectory, we measure the relative deviation of the predicted clean latent

$$R_k(\delta) = \frac{\|\hat{x}_0(z_k; x_\delta) - \hat{x}_0(z_k; x)\|_2^2}{\|\hat{x}_0(z_k; x)\|_2^2}, \tag{5}$$

where $\hat{x}_0(z_k; x)$ is the clean latent predicted by the model at z_k conditioned on x, and the first latent frame is excluded. We independently optimize one perturbation that maximizes R_0 and another that

Table 5: Protection performance under single-step injection on Wan2.2-TI2V-5B.

Injection	I2V Subject	I2V Background	Subject Consistency	Background Consistency	Aesthetic Quality	Imaging Quality	Motion Smoothness
$k = 0$	0.705	0.714	0.541	0.711	0.318	0.762	0.968
$k = 1$	0.681	0.709	0.524	0.689	0.580	0.682	0.932
$k = 2$	0.914	0.881	0.861	0.827	0.614	0.761	0.965
$k = 5, 10, 20$	0.999	0.999	0.996	0.994	0.615	0.742	0.994
Random direction	0.999	0.999	0.998	0.997	0.610	0.741	0.996
Full trajectory	0.674	0.709	0.553	0.772	0.352	0.332	0.963

Table 6: Protection performance under single-step injection on CogVideoX-5B-I2V.

Injection	I2V Subject	I2V Background	Subject Consistency	Background Consistency	Aesthetic Quality	Imaging Quality	Motion Smoothness
$k = 0$	0.625	0.668	0.534	0.733	0.255	0.420	0.752
$k = 1$	0.615	0.648	0.586	0.720	0.350	0.370	0.783
$k = 2$	0.634	0.670	0.615	0.792	0.389	0.377	0.828
$k = 5$	0.437	0.550	0.609	0.892	0.416	0.595	0.770
$k = 10$	0.611	0.636	0.599	0.893	0.488	0.535	0.881
$k = 20$	0.594	0.662	0.654	0.872	0.538	0.485	0.937
Random direction	0.973	0.975	0.986	0.968	0.522	0.739	0.982
Full trajectory	0.590	0.642	0.545	0.760	0.256	0.417	0.750

maximizes R_5 under identical settings, and compare their directions in pixel space with the cosine similarity and the fraction of pixels with the same sign. As shown in Table 7, the two perturbations have a cosine similarity of 0.121 and a sign agreement of 0.520, close to 0.000 and 0.488 for a pair of random perturbations, indicating that perturbations optimized for different steps are nearly orthogonal. We further condition the model on the image protected by the step-5 perturbation at each of steps 0 to 5 of the clean trajectory and compute R_k, averaged over four initial noises. As shown in Table 8, R_k drops from 620 at step 5 to 1.4 at step 0, indicating that a perturbation optimized for a later step barely affects the first step.

Table 7: Directional similarity between step-specific perturbations.

Pair of perturbations	Cosine Similarity	Sign Agreement
Step 0 and step 5	0.121	0.520
Step 0 and first two steps	0.175	0.540
Random pair	0.000	0.488

Joint Multi-Step Optimization. We evaluate two strategies for optimizing the perturbation over multiple denoising steps on Wan2.1-I2V-14B with 17-frame videos, seed 42, and no input transformation, namely joint optimization over the first 10 steps with the latents fixed to the clean trajectory, and optimization by back-propagating through the fully unrolled 10-step trajectory. Both strategies are trained with 10 sampling steps and evaluated with 10 and 40 sampling steps. As shown in Table 9, joint optimization further lowers I2V Subject and Subject Consistency, but degrades Aesthetic Quality and Imaging Quality less than optimizing the first step alone, and requires ten denoiser evaluations per iteration instead of one. Back-propagating through the unrolled trajectory is much weaker, with I2V Subject of 0.900 and 0.943 under 10 and 40 sampling steps, compared with 0.659 and 0.666 for the first step alone. The protection obtained by optimizing the first step alone also carries over from 10 to 40 sampling steps with little change in I2V Subject and Subject Consistency.

First Step and First Two Steps. We compare optimizing only the first denoising step with jointly optimizing the first two steps on Wan2.2-TI2V-5B, where both variants are trained with $N = 50$ sampling steps and evaluated with different numbers of sampling steps N. As shown in Table 10, optimizing the first two steps yields lower Subject Consistency and much lower Imaging Quality

Table 8: Relative deviation R_k of the predicted clean latent caused by the step-5 perturbation at different steps.

Step k	0	1	2	3	4	5
R_k	1.4	5.8	252	613	665	620

Table 9: Protection performance under joint multi-step optimization on Wan2.1-I2V-14B.

Setting	Steps	I2V Subject	I2V Background	Subject Consistency	Background Consistency	Aesthetic Quality	Imaging Quality	Motion Smoothness
First step	10	0.659	0.683	0.492	0.719	0.262	0.250	0.966
First step	40	0.666	0.689	0.479	0.729	0.316	0.323	0.973
Joint, 10 steps	10	0.571	0.597	0.400	0.714	0.300	0.299	0.765
Joint, 10 steps	40	0.466	0.559	0.340	0.709	0.353	0.435	0.806
Unrolled, 10 steps	10	0.900	0.927	0.864	0.863	0.547	0.694	0.978
Unrolled, 10 steps	40	0.943	0.964	0.957	0.937	0.557	0.733	0.983

under all values of N, reaching 0.273 compared with 0.693 at $N = 50$, whereas optimizing the first step alone yields lower Motion Smoothness and requires half the denoiser evaluations per iteration.

Table 10: Protection performance of optimizing the first step and the first two steps on Wan2.2-TI2V-5B.

N	Setting	I2V Subject	I2V Background	Subject Consistency	Background Consistency	Aesthetic Quality	Imaging Quality	Motion Smoothness
10	First step	0.734	0.747	0.618	0.756	0.444	0.565	0.883
10	First two steps	0.657	0.718	0.512	0.751	0.373	0.302	0.979
20	First step	0.663	0.704	0.538	0.754	0.328	0.724	0.860
20	First two steps	0.673	0.738	0.510	0.758	0.360	0.266	0.981
40	First step	0.673	0.711	0.522	0.754	0.319	0.615	0.904
40	First two steps	0.680	0.741	0.510	0.750	0.315	0.269	0.979
50	First step	0.716	0.749	0.594	0.765	0.428	0.693	0.894
50	First two steps	0.687	0.736	0.525	0.757	0.333	0.273	0.976

C Sensitivity Decomposition of the Sampling Trajectory

We derive how the final latent z_N depends on the input perturbation δ, which shows how the protective effect can in principle be distributed across denoising steps. Specializing Equation (1) to an Euler-type update, the sampling recursion reads

$$z_{k+1} = z_k + h_k\, v_\theta(z_k, t_k; x_\delta), \qquad k = 0, \dots, N-1, \tag{6}$$

where $x_\delta = x + \delta$ is the conditioning image and h_k is the step size determined by the sampler. Differentiating with respect to δ gives

$$\frac{\partial z_{k+1}}{\partial \delta} = \Big(I + h_k \frac{\partial v_k}{\partial z_k}\Big)\frac{\partial z_k}{\partial \delta} + h_k \frac{\partial v_k}{\partial x_\delta}. \tag{7}$$

Since the initial noise z_0 does not depend on δ, unrolling this recursion from $k = 0$ to $k = N-1$ gives

$$\frac{\partial z_N}{\partial \delta} = \sum_{k=0}^{N-1} \underbrace{\Big[\prod_{j=k+1}^{N-1} \Big(I + h_j \frac{\partial v_j}{\partial z_j}\Big)\Big]}_{M_k}\; h_k\; \underbrace{\frac{\partial v_k}{\partial x_\delta}}_{J_k}, \tag{8}$$

where $J_k = \partial v_k / \partial x_\delta$ is the Jacobian of the velocity prediction at step k with respect to the conditioning image, and M_k is the linearized propagation operator that maps a perturbation introduced at step k to its effect on the final latent, with the product ordered so that later steps are on the left.

For Wan2.2-TI2V-5B, which overwrites the first latent slice with the image latent, we absorb this dependence into the conditioning argument of v_θ, so the identity holds unchanged.

Equation (8) expresses the total sensitivity of the generation to δ as a sum of per-step contributions $M_k h_k J_k$, each depending on how much the input perturbation locally deflects the velocity prediction at step k, captured by J_k, and how this deflection is subsequently amplified or attenuated by the remaining sampling trajectory, captured by M_k. A joint multi-step objective optimizes δ against a weighted sum of these per-step terms. As shown in Table 7, perturbations optimized for different steps are nearly orthogonal in pixel space, so a joint objective has to split the same budget among them, which is consistent with the limited gain of joint multi-step optimization relative to its cost in Table 9.

D Robustness under Different Transformation Strengths

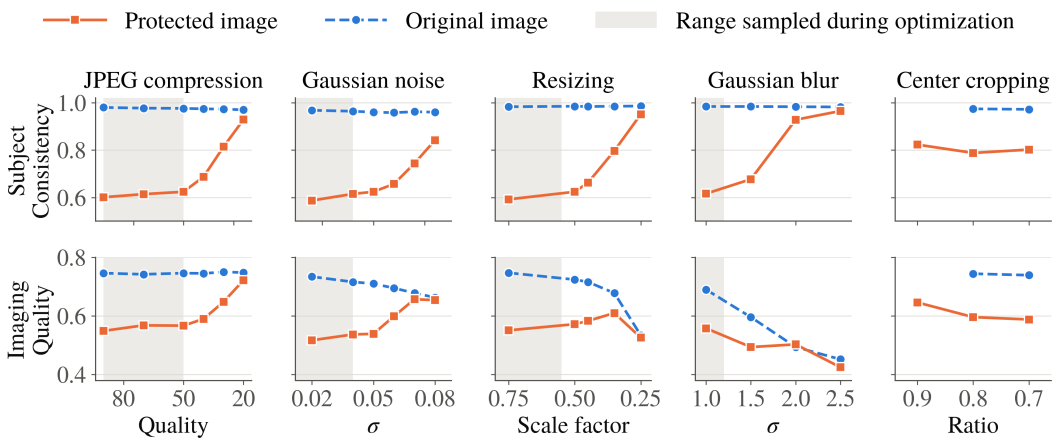


Figure 5: Protection robustness of RVDGuard against various transformations under different parameter settings. Each transformation is applied with the same parameter to the protected image and to the original image. Shaded regions indicate the parameter ranges sampled during optimization.

Figure 5 shows the results of RVDGuard on Wan2.2-TI2V-5B under all parameter settings of the four transformation types used in optimization and of center cropping, where each transformation is applied with the same parameter to both the protected image and the original image. Most of the evaluated settings lie beyond the parameter ranges sampled during optimization, which cover JPEG quality from 50 to 90, Gaussian noise with σ from 0.01 to 0.04, resizing with scale from 0.55 to 0.85, and Gaussian blur with σ from 0.4 to 1.2. The protective effect is largely preserved at moderate strengths and is gradually reduced at stronger ones. For JPEG compression with quality below 30, Gaussian noise with $\sigma \geq 0.08$, resizing below $\times 0.35$, and Gaussian blur with $\sigma \geq 2.0$, the Subject Consistency of videos generated from the protected image approaches that of videos generated from the original image under the same transformation. Except for JPEG compression, these settings also noticeably degrade the videos generated from original images, whose Imaging Quality drops to 0.663, 0.534, and 0.494 for Gaussian noise with $\sigma = 0.08$, resizing with $\times 0.25$, and Gaussian blur with $\sigma = 2.0$, respectively. Center cropping, which is not included in the transformation distribution, raises the Subject Consistency of RVDGuard to about 0.8 at all three ratios, even at the mildest ratio of 0.9, while the videos generated from cropped original images remain almost unaffected. This indicates that cropping weakens the protection by shifting the perturbation relative to the pixel grid rather than through its strength, and the protection is only partially retained, with Subject Consistency still below 0.83 compared with about 0.97 for the original image.

E Ablation on Expectation over Transformation

To examine the effect of EoT, we optimize the perturbation with the same settings but without any transformation during optimization, and apply the same transformations as in Table 3 to the

resulting protected images. As shown in Table 11, removing EoT makes the protection fragile to common transformations. Without EoT, Gaussian noise with $\sigma = 0.05$ and resizing with $\times 0.5$ raise Subject Consistency to 0.871 and 0.963 and I2V Subject to 0.908 and 0.980, which largely removes the protection, while JPEG compression with $q = 50$ is comparatively mild and raises Subject Consistency to 0.747. With EoT, Subject Consistency stays between 0.702 and 0.765 under the three transformations. EoT does not help against center cropping, which is not included in the transformation distribution, with Subject Consistency of 0.802 compared with 0.795 without EoT.

Table 11: Ablation on EoT under various transformations.

	Subject Consistency ↓		I2V Subject ↓		Imaging Quality ↓	
Transformation	w/o EoT	w/ EoT	w/o EoT	w/ EoT	w/o EoT	w/ EoT
Gaussian noise, $\sigma = 0.05$	0.871	**0.717**	0.908	**0.872**	0.670	**0.585**
JPEG compression, $q = 50$	0.747	**0.702**	0.891	**0.868**	0.599	**0.567**
Resizing, $\times 0.5$	0.963	**0.765**	0.980	**0.889**	0.716	**0.601**
Center cropping, ratio 0.7	**0.795**	0.802	**0.937**	0.938	**0.584**	0.588

F Robustness against Frame Regeneration

Beyond generating videos directly from protected images, an adversary may exploit any frame of the generated video that still has acceptable quality. We therefore evaluate a frame regeneration attack on CogVideoX-5B-I2V, in which the adversary selects the frame with the highest DINO similarity to the original image, excluding the first frame, from the video generated with the protected image, restores it with SeedVR2-3B (Wang et al., 2025b), and uses the restored frame as a new input image for I2V generation with the same prompt. I2V Subject and I2V Background of the regenerated videos are computed with respect to the original image, and the quantitative results are shown in Table 12, where the results without regeneration are taken from Table 1. The regeneration results of all protections are computed on the first 100 VBench-I2V images. Frame regeneration largely removes the degradation caused by baseline protections, with Subject Consistency rising to between 0.953 and 0.967 after regeneration. For RVDGuard, regeneration also improves Aesthetic Quality and Imaging Quality to 0.524 and 0.615, since SeedVR2 restores the quality of the selected frame. However, the regenerated videos remain temporally inconsistent, with Subject Consistency of 0.762 and Background Consistency of 0.879, the lowest among all protections. They also deviate from the original image, with I2V Subject of 0.720 and I2V Background of 0.748, the lowest among all protections, and only Immune2V with a budget four times larger reaches a comparable I2V Subject of 0.722. These results indicate that even after restoration and regeneration, the videos derived from images protected by RVDGuard do not faithfully reproduce the original image.

G Robustness to the Initial Noise

The objective in Equation (3) is defined on a single initial noise z_0 drawn with seed 42, and the videos in Section 5.2 are generated with the same seed. To verify that the protection does not rely on this particular noise, we generate videos on Wan2.2-TI2V-5B from the same protected images with seed 31337, which is never used during optimization, on the first 100 VBench-I2V images without input transformation. As shown in Table 13, the protection remains effective under the unseen seed: relative to Clean, Subject Consistency drops by 39.1% and Imaging Quality by 28.0%, compared with 31.4% and 23.3% under seed 42, while the relative drops in I2V Subject and Aesthetic Quality, 17.2% and 28.5%, are close to 15.9% and 31.2% under seed 42. These results indicate that the first-step deflection is not tied to the initial noise used during optimization.

H Limitations

Cropping and median filtering, which are not included in the transformation distribution, weaken the protection more than other transformations. Diffusion-based purification, such as DiffPure (Nie et al., 2022) and noisy upscaling (Hönig et al., 2025), is largely unexplored. In addition, we evaluate transferability only on DynamiCrafter, and transfer to other I2V architectures remains unexplored.

Table 12: Quantitative results against frame regeneration on CogVideoX-5B-I2V.

Method	I2V Subject ↓	I2V Background ↓	Subject Consistency ↓	Background Consistency ↓	Aesthetic Quality ↓	Imaging Quality ↓	Motion Smoothness ↓
PhotoGuard	0.877	0.899	0.922	0.940	0.446	0.612	0.973
+ Regen.	0.874	0.904	0.962	0.967	0.476	0.621	0.985
Anti-Prompt	0.865	0.895	0.801	0.876	0.477	0.535	0.965
+ Regen.	0.863	0.887	0.953	0.958	0.528	**0.595**	**0.983**
Immune2V	0.689	0.733	0.906	0.928	0.341	0.679	0.974
+ Regen.	0.722	0.768	0.967	0.967	**0.356**	0.669	0.986
Anti-I2V	0.919	0.946	0.938	0.950	0.596	0.646	0.976
+ Regen.	0.935	0.951	0.963	0.965	0.594	0.669	0.985
I2VGuard	0.889	0.916	0.884	0.921	0.504	0.562	0.975
+ Regen.	0.897	0.920	0.954	0.958	0.539	0.631	**0.983**
RVDGuard	0.739	0.772	0.583	0.791	0.329	0.293	0.991
+ Regen.	**0.720**	**0.748**	**0.762**	**0.879**	0.524	0.615	0.984

Table 13: Protection performance of RVDGuard on Wan2.2-TI2V-5B under different generation seeds.

Seed	Method	Subject Consistency ↓	I2V Subject ↓	Aesthetic Quality ↓	Imaging Quality ↓
42	Clean	0.963	0.985	0.657	0.716
42	RVDGuard	0.661 (−31.4%)	0.828 (−15.9%)	0.452 (−31.2%)	0.549 (−23.3%)
31337	Clean	0.958	0.977	0.627	0.694
31337	RVDGuard	0.583 (−39.1%)	0.809 (−17.2%)	0.448 (−28.5%)	0.500 (−28.0%)