I2VSHIELD: FOCUSED ADVERSARIAL PROTECTION AGAINST HARMFUL IMAGE-TO-VIDEO GENERATION

Anonymous authors
Paper under double-blind review

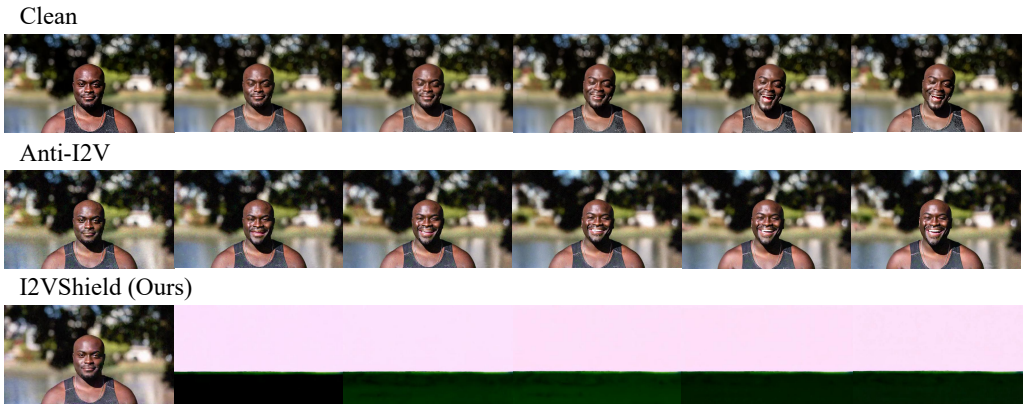

Figure 1: **Visual results of image-level protections on I2V generation.** All videos are generated by Wan2.1-I2V-14B (Wan et al., 2025) from the same prompt; only the perturbation applied to the input image differs, and all perturbations share the same ℓ_∞ budget $\varepsilon = 16/255$. **Top:** the unprotected image yields a temporally coherent, high-fidelity video whose motion faithfully follows the text prompt. **Middle:** the image protected by Anti-I2V (Vu et al., 2026), the strongest prior protection we are aware of, yields a video of comparable quality, preserving subject identity, background, and prompt-consistent motion. **Bottom:** the image protected by I2VShield yields a video in which every frame after the conditioning frame collapses into a saturated, structureless color field.

ABSTRACT

Modern image-to-video (I2V) diffusion models turn any photograph into a high-fidelity video, making non-consensual deepfakes cheap to produce. A natural defense is adversarial protection: adding an imperceptible perturbation (under a small budget) to an image before publication so that generation from it fails. Yet existing methods inherit the text-to-image (T2I) recipe, which barely dents modern I2V backbones. We trace this failure to the recipe's two pillars, neither paying off under I2V. First, the 3D VAE is far harder to move than its 2D counterpart, so the same budget buys far less disruption here. Second, and more decisively, I2V's static conditioning leaves random-timestep averaging both unnecessary and self-diluting: the perturbation never enters the noisy latent, and we measure cross-timestep gradients to be near-orthogonal (cosine I2V 0.07 vs T2I 0.60). Probing where the condition acts, we locate its influence at the largest noise level, decaying monotonically thereafter, so averaging spends most of the budget where the condition barely responds. Together, these findings identify what decides protection strength: not which objective is attacked, but where along the timestep axis the budget is spent. Guided by this diagnosis, we propose **I2VShield**, which forgoes the VAE objective and commits the entire budget to the largest noise level, where the condition's influence peaks. Across 9 I2V models and 2 datasets, it lowers DINO similarity from 0.92 to 0.18 on average, against 0.80 for the strongest, outperforming even at a $16\times$ smaller budget (0.43 vs 0.80 DINO). We further release HQFace-I2V, the first benchmark purpose-built for this setting.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

Diffusion models (Ho et al., 2020; Song et al., 2021; Lipman et al., 2023) have driven rapid progress in video generation, where image-to-video (I2V) generation offers pixel-level control by animating a single reference image. Modern large-scale I2V models (Yang et al., 2025; Wan et al., 2025; HaCohen et al., 2024) are trained on web-scale video with billions of parameters, and animate an arbitrary photograph into multi-second, high-resolution video following text-specified motion. As shown in Fig. 1 (Top), the generated frames evolve in accordance with the given text prompt while faithfully preserving the subject's identity and scene structure.

However, the same fidelity turns I2V models into a direct vector for abuse: any photograph posted online becomes a reusable source for fabricating video of a real person, and the added motion and temporal narrative make such forgeries far more persuasive than static images (Pang et al., 2025). Deployed I2V systems have already been reported to animate a public figure's photo into sexually explicit video without jailbreaking (AI Incident Database, 2025), and identity-preserving video forgery has enabled multi-million-dollar fraud (CNN Business, 2024).

A promising defense is proactive adversarial protection, first developed against text-to-image (T2I) models (Salman et al., 2023; Shan et al., 2023; Liang et al., 2023; Liang & Wu, 2023; Xue et al., 2024). Before publishing an image, its owner adds an imperceptible perturbation (constrained by a small ℓ_∞ budget), obtained either by attacking the VAE encoder to corrupt its latent representation or by optimizing the diffusion denoising objective at random timesteps, so that conditioning on it degrades the result. More recently, this idea has been carried over to the I2V setting (Gui et al., 2025; Chowdhury et al., 2025; Vu et al., 2026), where similar perturbations promise to suppress object motion or corrupt generated frames, yet fall short against video diffusion backbones. As shown in Fig. 1 (Middle), Anti-I2V (Vu et al., 2026), the state-of-the-art protection in this setting, offers limited defense against Wan2.1-I2V-14B (Wan et al., 2025): the protected image still yields a high-quality video that preserves both subject identity and coherent motion.

Beyond this, no benchmark specifically targets the I2V protection setting. Existing work either does not release its evaluation data (Gui et al., 2025; Chowdhury et al., 2025), or repurposes datasets assembled for other tasks, such as low-resolution action-recognition footage and interview-style face videos (Vu et al., 2026). Their inputs are video frames rather than real photographs, and each image is paired with a single prompt. Protection is therefore never tested on high-quality user photos and diverse prompts that real attackers would exploit.

To bridge these gaps, in this paper, we first introduce **HQFace-I2V**, the first benchmark specifically designed for I2V adversarial protection. Unlike prior evaluation suites that rely on video frames or synthesized images, HQFace-I2V is built from high-quality real-world portrait photographs. Each image is paired with multiple prompts, ensuring that protection performance is assessed across diverse animations of the same identity rather than a single fixed generation.

We then systematically investigate why existing I2V protection methods consistently fail in practice, and our analysis uncovers two underlying causes. First, we find that the 3D VAE employed by I2V models is markedly harder to move than its T2I counterpart: under the same perturbation budget, the latent deviation it admits is substantially smaller in magnitude. Attacking the VAE encoder therefore buys far less disruption here than in T2I, making it a poor use of a fixed budget.

Second, and more decisively, we identify a fundamental mismatch between I2V's static-conditioning design and the optimization strategy inherited from T2I. In T2I, the perturbation directly enters the noisy latent through forward diffusion, so every noise level must be covered by the attack, and we also find that averaging over timesteps can accumulate the attack signal. In I2V, however, neither premise survives. First, the perturbation never enters the noisy latent: it remains a static condition throughout denoising, so no noise level needs covering. Second, we measure how attack gradients at different timesteps align: they are near-orthogonal in I2V (cosine 0.07) but well aligned in T2I (0.60), so averaging dilutes the attack signal rather than accumulating it. Probing how the image condition drives denoising, we further find its influence concentrated at high timesteps and decaying monotonically as sampling proceeds, so the budget pays off most at the highest noise levels and least once it has faded. Random-timestep averaging thus both dilutes its own attack signal and spends most of its budget where the image condition barely responds.

Building on these insights, we propose **I2VShield**, a focused adversarial protection strategy against malicious I2V generation. Unlike prior methods that combine VAE latent perturbation with diffusion optimization over randomly sampled timesteps, I2VShield strictly removes both inherited components, adding nothing in their place: it forgoes any auxiliary VAE objective and, rather than averaging over timesteps, concentrates the entire attack budget on the largest noise level, where the image condition's influence peaks. The resulting objective is a single term at a single timestep, and every component it drops is one our diagnosis shows to be ineffective or self-diluting under I2V. To the best of our knowledge, and as illustrated in Fig. 1 (Bottom), I2VShield is the first method capable of inducing catastrophic degradation in videos generated by modern I2V models.

Extensive experiments across 9 I2V models spanning 4 architectures, evaluated on HQFace-I2V and a public benchmark, validate our approach. I2VShield outperforms 6 baselines, including 2 I2V-specific state-of-the-art methods, cutting average DINO similarity from 0.92 to 0.18 versus 0.80 for the strongest. Further experiments on generalization and robustness confirm these gains, with I2VShield still surpassing existing defenses at a $16\times$ smaller budget (0.43 vs 0.80 DINO).

Our contributions are summarized as follows: ① We introduce HQFace-I2V, the first benchmark purpose-built for I2V adversarial protection, featuring high-quality real-world portraits paired with diverse animation prompts. ② We diagnose why existing protections fail: the 3D VAE is far harder to move, so budget spent there buys little, while under I2V's static conditioning cross-timestep gradients are near-orthogonal (cosine 0.07 vs T2I's 0.60), so averaging dilutes the attack signal. ③ Identifying where the budget is spent as what decides protection strength, we propose I2VShield, which forgoes the VAE objective and commits the entire budget to the largest noise level, where the image condition's influence peaks, obtaining protection by removal rather than addition. ④ Across 9 I2V models and 4 architectures, I2VShield outperforms 6 baselines, cutting average DINO to 0.18 against 0.80 for the strongest, and winning even at a $16\times$ smaller budget.

2 Related Work

2.1 Image-To-Video Diffusion Model

Diffusion models (Ho et al., 2020; Song et al., 2021) and flow-based formulations (Lipman et al., 2023) have driven rapid advances in video generation, evolving from UNet-based architectures to Diffusion Transformers (DiTs) and from discrete sampling to continuous flow matching (Yang et al., 2025; Lipman et al., 2023; Wan et al., 2025; HaCohen et al., 2024). Image-to-video generation, which animates a reference image into temporally coherent motion while preserving subject identity, has become a key paradigm, with the central technical challenge being how image information is injected into the generative process. Early foundational models such as SVD (Blattmann et al., 2023) and DynamiCrafter (Xing et al., 2024) established the dual-stream template: a VAE-encoded latent concatenated with noisy input for pixel-level appearance, and a semantic pathway (e.g., CLIP embeddings via cross-attention or Query Transformer) for high-level semantic context. Modern large-scale I2V models (CogVideoX (Yang et al., 2025), LTX-Video (HaCohen et al., 2024), Wan (Wan et al., 2025)) differ in the concrete mechanisms for injecting image information (e.g., channel-wise latent concatenation or first-frame pinning), yet converge on the same underlying design: the reference image, typically the first frame, is treated as a static condition that remains fixed throughout the entire denoising trajectory. Despite such implementation diversity, these models uniformly achieve high-fidelity animation with faithful identity preservation, a capability that enables malicious synthesis and motivates the development of proactive adversarial protection.

2.2 Adversarial Protection

Adversarial protection against diffusion models. Proactive adversarial protection adds adversarial perturbations to images before publication, so that a generative model conditioned or fine-tuned on them produces degraded output. Early work targets text-to-image models along two attack surfaces: the VAE encoder and the diffusion denoising process. Encoder-side methods such as PhotoGuard (Salman et al., 2023) and Glaze (Shan et al., 2023) push the image's latent representation toward a target latent; denoising-side methods such as AdvDM (Liang et al., 2023) optimize against the diffusion objective evaluated at randomly sampled timesteps. Later methods go further: Mist (Liang & Wu, 2023) attacks both surfaces jointly, while SDS (Xue et al., 2024) accelerates

denoising-side optimization through Score Distillation Sampling. The denoising-side objectives further share an implicit assumption: the adversarial perturbation is injected into the noisy latent through forward diffusion, hence multi-timestep optimization is necessary and gradients from different noise levels naturally cooperate. As we will show, this assumption does not carry over to I2V, where the adversarial image serves as a static condition throughout denoising.

Image-to-video protection and evaluation. Adversarial protection has recently been extended to the I2V setting, where the goal becomes disrupting both visual quality and temporal coherence. I2VGuard (Gui et al., 2025) combines a VAE-side objective with temporal-attention disruption, and VidFreeze (Chowdhury et al., 2025) suppresses attention alone to freeze motion, while Anti-I2V (Vu et al., 2026), the current state-of-the-art, collapses internal representations onto shallow features and anchors them to an unrelated image. All three, however, inherit the T2I recipe—a VAE-side objective, or a diffusion loss spread over sampled timesteps—and offer limited defense against modern backbones such as Wan (Wan et al., 2025). Moreover, no purpose-built benchmark exists for this setting: prior works either do not release their data (Gui et al., 2025; Chowdhury et al., 2025) or repurpose datasets of video frames, each paired with a single prompt (Vu et al., 2026), inputs unlike the real photographs and diverse prompts real attackers would exploit.

In addition to adversarial protection, we discuss further related work in App. A.

3 HQFACE-I2V

Existing evaluations of I2V adversarial protection either do not release their data (Gui et al., 2025; Chowdhury et al., 2025) or repurpose datasets of video frames paired with a single prompt (Vu et al., 2026), inputs and prompts unlike those real attackers would exploit. To enable rigorous evaluation in this setting, we introduce **HQFace-I2V**, the first benchmark purpose-built for I2V adversarial protection. It is built from high-quality real-world portrait photographs: we queried the Pexels API (Pexels, 2026) with 20 semantically diverse search phrases and verified every candidate with a VLM (Doubao-1.8 Vision Pro (ByteDance Seed, 2026)) against six criteria (color photograph, real human face, close-up, natural lighting, minimally processed style, everyday background), accepting 80 images spanning varied age, gender, and ethnicity in outdoor, indoor, and urban settings, and center-cropped and resized all of them to 1280×720 (16:9). For each image, the same VLM generated three animation prompts grounded in the subject's appearance, covering three distinct motion types: expression-driven motion (smiling, blinking, speaking), head and gaze movement, and environmental dynamics, yielding 240 image–prompt pairs. This multi-prompt design enables cross-prompt evaluation: each protection is assessed across diverse animations of the same identity rather than a single fixed generation, which prior single-prompt suites cannot expose (e.g., defenses that collapse when the attacker's prompt changes). Curation procedures, dataset statistics, file formats, and a comparison with prior evaluation resources are provided in App. B.

4 I2VSHIELD

4.1 PROBLEM SETUP

We begin by establishing the notation and formalizing the adversarial protection problem for I2V generation. An I2V model comprises a 3D VAE with encoder E and decoder D, and a denoising network ϵ_θ. Let z_t denote the video latent at noise level t. I2V generation is conditioned on a reference image and a text prompt y: the reference image x is encoded once into an image condition $c_{\text{img}} = E(x)$ which, together with y, is injected into the denoising network at every timestep, yielding $\epsilon_\theta(z_t, c_{\text{img}}, y, t)$. The network is trained to reverse the noising process:

$$z_t = \sqrt{\bar{\alpha}_t}\, z_0 + \sqrt{1-\bar{\alpha}_t}\, \epsilon, \quad \epsilon \sim \mathcal{N}(0, I), \tag{1}$$

by minimizing the simplified denoising objective (Ho et al., 2020)

$$\mathcal{L}(\theta) = \mathbb{E}_{t\sim\mathcal{U}(0,T),\, z_0,\, \epsilon}\left[\, \|\epsilon_\theta(z_t, c_{\text{img}}, y, t) - \epsilon\|^2 \,\right]; \tag{2}$$

flow-matching variants (Lipman et al., 2023) replace the noise target with the velocity field $\epsilon - z_0$ under the linear schedule $z_t = (1-t)\, z_0 + t\, \epsilon$ ($t \in [0, 1]$), and our analysis applies to both.

Having fixed the notation, we turn to the protection problem. Unlike T2I, the image condition is independent of t: an adversarial perturbation δ affects generation only through $c^{\delta}_{\text{img}} = E(x+\delta)$, never through the noisy latent z_t. As we show in Sec. 4.2, this difference breaks the T2I optimization recipe. Adversarial protection seeks δ with $\|\delta\|_\infty \le \varepsilon$ that degrades I2V generation from $x+\delta$ regardless of the prompt. Let $\mathcal{A}(\epsilon_\theta, z_t, c_{\text{img}}, y, t)$ denote a per-timestep degradation measure, whose concrete form varies across methods (e.g., denoising prediction error, internal representation collapse, or attention disruption), and $\mathcal{V}(c_{\text{img}})$ a VAE-side objective (e.g., pushing the encoding toward a target representation). Despite their diverse forms, existing I2V protections all follow the T2I recipe of combining the two terms, evaluating $\mathcal{A}$ at uniformly sampled timesteps:

$$\delta^* = \arg\max_{\|\delta\|_\infty \le \varepsilon} \mathbb{E}_{t\sim\mathcal{U}(0,T)} \left[\mathcal{A}\big(\epsilon_\theta, z_t, c^{\delta}_{\text{img}}, y, t\big) \right] + \mathcal{V}\big(c^{\delta}_{\text{img}}\big), \tag{3}$$

with method-specific weights, and some methods use only one of the two terms. In Sec. 4.2, we analyze why this formulation fails in I2V and identify two causes in the shared structure of Eq. (3): the VAE-side term $\mathcal{V}(c^{\delta}_{\text{img}})$ is absorbed by the robustness of the 3D VAE, while the expectation over t makes gradients from different timesteps cancel under the static condition.

4.2 Diagnosing Existing Methods

Existing methods carry two design choices from the T2I recipe: auxiliary objectives that corrupt the VAE encoding, and an expectation over uniformly sampled timesteps. Under I2V's design, both choices fail, and each failure points to a design decision for our method.

Diagnosis 1: the VAE encoder. We first study the VAE-encoder attack surface. We optimize a perturbation δ within the budget ε to maximize the latent deviation $\|E(x+\delta) - E(x)\|$, and measure the reconstruction gap between $D(E(x+\delta))$ and x with DINO (Oquab et al., 2024), a widely used perceptual metric, on four models: two with 2D VAEs from the Stable Diffusion (SD) series (Rombach et al., 2022) (SD1.5, SD2.1) and two with 3D VAEs (CogVideoX (Yang et al., 2025), Wan (Wan et al., 2025)). As shown in Fig. 2, the same attack yields a markedly smaller reconstruction gap on the 3D VAEs than on the 2D VAEs, and the 3D VAEs stay hard to move even when the budget is raised to $64/255$. A plausible explanation is the 3D VAE's larger capacity and temporal compression, which may make its latent less sensitive to pixel-level changes; regardless of the cause, the perturbation is absorbed at the representation level, leaving encoder-side objectives little room to act. Under a budget fixed by imperceptibility, attacking the encoder buys far less disruption here than in T2I.

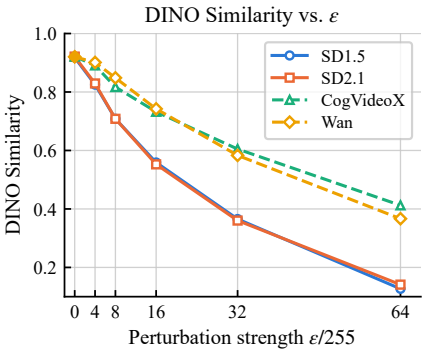

Figure 2: DINO similarity (Oquab et al., 2024) between the original and the perturbed reconstruction versus the perturbation budget ε across different VAEs.

Diagnosis 2: random timestep sampling. As a second step toward understanding why existing protections fail on I2V models, we investigate the timestep axis of the optimization. Our analysis reveals two facts, which we detail in turn: per-timestep gradients do not cooperate in I2V, in contrast to T2I; and the condition's influence peaks at the largest noise level and decays from there.

Gradient alignment across timesteps. The expectation over t averages per-timestep gradients, so whether averaging accumulates a coherent attack signal or dilutes it hinges on their mutual alignment. We therefore compute the gradients $g_t = \nabla_\delta \mathcal{A}(\epsilon_\theta, z_t, c^{\delta}_{\text{img}}, y, t)$ at the original image over timesteps and measure the pairwise cosine similarity between them, in both T2I and I2V. As Fig. 3 (a) shows, the two regimes differ qualitatively: gradients from different timesteps are positively aligned in T2I, whereas in I2V they are near-orthogonal and in places negatively aligned, a mean cross-timestep cosine of 0.07 against 0.60. The I2V gradient field is dispersed across the timestep axis, so the expectation over t dilutes the attack signal rather than accumulating it.

Condition influence over timesteps. To locate the condition's influence, we analyze the attention maps of Wan (Wan et al., 2025). There, the conditioning image is encoded and placed at the first frame, so the attention a frame pays to the first frame's tokens measures how strongly the condition drives that frame's generation. As Fig. 3 (b) shows, this attention is consistently largest at the highest

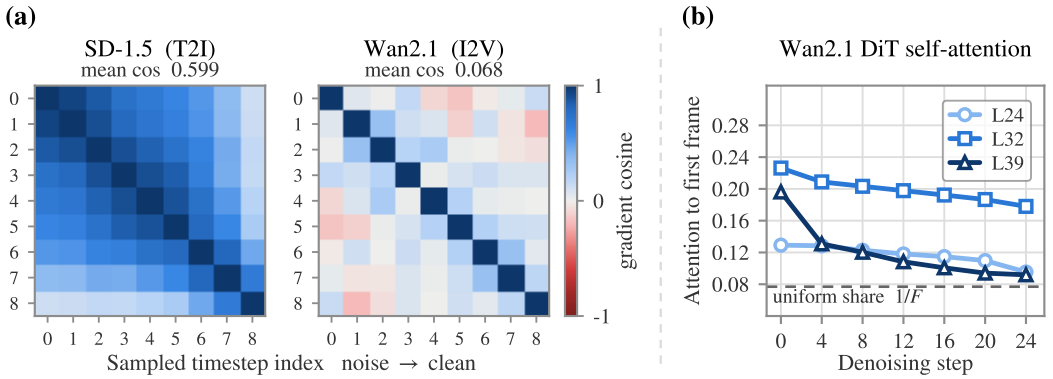

Figure 3: **The timestep axis.** **(a)** Pairwise cosine between per-timestep gradients, taken at the original image. T2I gradients are positively aligned; I2V gradients are near-orthogonal or negatively aligned, so the expectation over t cancels the attack signal. **(b)** Share of self-attention that each latent frame directs at the first (conditioning) frame in Wan (Wan et al., 2025), against the uniform share $1/F$. It is largest at high timesteps and decays monotonically as denoising proceeds.

noise level and decays as denoising proceeds: at the deepest layer it starts at $2.6\times$ the uniform share $1/F$ and falls by $2.1\times$ toward that share by the end of sampling, with nearly two-thirds of the decline spent in the first sampling interval alone. A perturbation acting through the condition therefore has the most leverage where this attention is highest, and progressively less as it decays. Budget placed at the largest noise level thus acts where the condition responds most, while budget spread over the middle and low noise levels acts where it responds least.

These two findings together explain why the T2I recipe fails under I2V's static conditioning. In T2I, random sampling is both necessary and sufficient: the perturbation enters the noisy latent z_t through forward diffusion and must survive every noise level the sampler visits, and, as measured above, gradients cooperate, so the expectation over t accumulates a coherent attack signal. In I2V, neither holds. The perturbation never enters z_t; it acts solely on the static condition $c_{\text{img}} = E(x+\delta)$, so no noise level needs covering. And its gradients are near-orthogonal rather than mutually reinforcing, plausibly because each timestep consults the condition differently, so averaging over t dilutes the attack signal and wastes budget on directions that never accumulate. Since the condition responds most at the largest noise level, we commit the entire budget there, the core design of I2VShield.

Full experimental details and additional results are provided in App. C.

4.3 From Diagnosis to Design

The two diagnoses translate directly into two design decisions, both of which remove a component of Eq. (3) rather than add one. Based on Diagnosis 1, we discard the VAE-side objective $\mathcal{V}(c^{\delta}_{\text{img}})$: under the same budget, a perturbation moves the 3D VAE's encoding far less than that of its 2D counterparts, leaving VAE-side objectives little room to act. Based on Diagnosis 2, we replace the expectation $\mathbb{E}_{t\sim\mathcal{U}(0,T)}$ with a point evaluation at the largest noise level $\tau^\star = 1$: averaging dilutes the attack signal because per-timestep gradients are near-orthogonal, and since the condition's influence peaks there and decays monotonically, this point captures all the leverage the condition offers. What remains is a single term evaluated at a single timestep,

$$\delta^* = \arg\max_{\|\delta\|_\infty \le \varepsilon} \mathcal{A}\big(\epsilon_\theta,\, z_{\tau^\star},\, c^{\delta}_{\text{img}},\, y_{\text{fix}},\, \tau^\star\big), \tag{4}$$

where y_{fix} is a fixed prompt used throughout optimization; since protection must hold for prompts chosen by the attacker, we verify in Sec. 5.3 that the resulting δ transfers across unseen prompts. Formally, we optimize Eq. (4) by sign-gradient ascent with projection onto the ℓ_∞ ball (Alg. 1).

Concretely, we instantiate $\mathcal{A}$ as the divergence between the network's predictions under the perturbed and clean conditions, $\|\epsilon_\theta(z_{\tau^\star}, c^{\delta}_{\text{img}}, y_{\text{fix}}, \tau^\star) - \epsilon_\theta(z_{\tau^\star}, c_{\text{img}}, y_{\text{fix}}, \tau^\star)\|_2^2$ (App. E.5). It measures how far the perturbation moves the first step of generation, where the condition exerts the most leverage, and it needs neither a target image nor a clean video latent. It is also agnostic to the parameterization: the same expression applies whether ϵ_θ predicts noise, v or a velocity, so one objective

serves every backbone we evaluate. Finally, $z_{\tau^\star}$ is resampled at every iteration (Alg. 1), so the perturbation is optimized over the distribution of sampler initializations rather than for a single seed, and does not depend on the seed used at generation. Its computational cost is reported in App. J.

5 EXPERIMENT

5.1 EXPERIMENT SETUP

Models and Datasets. We evaluate on 9 I2V models spanning four architectures. The main paper reports three: the flow-matching transformer Wan (Wan et al., 2025), the v-prediction transformer CogVideoX (Yang et al., 2025), and LTX-Video (HaCohen et al., 2024), which applies dynamic time shifting and encodes the reference image directly into its token sequence without a separate image encoder. In all I2V models, the image condition is concentrated in the first frame. The datasets are our high-quality benchmark HQFace-I2V and the public CelebV-Test dataset (Yu et al., 2023). More details on the models and datasets are in App. E.1 and App. E.2.

Baselines. We compare I2VShield with six baselines: the popular T2I protection methods PhotoGuard (Salman et al., 2023), AdvDM (Liang et al., 2023), Mist (Liang & Wu, 2023), and SDS+ (Xue et al., 2024), and two methods designed specifically for I2V, I2VGuard (Gui et al., 2025) and Anti-I2V (Vu et al., 2026), the state-of-the-art protection in this setting, both inheriting the T2I optimization recipe. All attacks are white-box against the target model and share the same ℓ_∞ budget $\varepsilon = 16/255$; each method keeps its original objective and iteration count, detailed in App. E.3.

Metrics. We evaluate protection effectiveness with two complementary groups. The first consists of three conventional metrics: DINO (Oquab et al., 2024), the feature similarity between the generated video and the input image, IQ (Huang et al., 2024), the VBench imaging quality, a no-reference score of per-frame distortion, and SC (Huang et al., 2024), the VBench subject consistency. The second consists of two large multimodal models (LMMs) acting as judges: VQA^2 (Jia et al., 2025), which predicts a continuous quality score from five-level quality logits without access to the prompt, and FineVQ (Duan et al., 2025), a fine-grained video quality VLM; both judge the generated video without reference to the input image, aligning the evaluation with human perception. Lower values indicate stronger protection throughout. These five metrics span the key facets of protection degradation, from semantic deviation and frame-level quality loss to consistency collapse and the perceptual judgement of two independent LMMs. More details are in App. E.4.

Implementation Details. We optimize in pixel space with sign-PGD using a step size of $\varepsilon/16$, $N = 100$ iterations, and a random start inside the ℓ_∞ ball, maximizing the velocity divergence between the perturbed and clean trajectories at the largest noise level $\tau^\star = 1$, under the fixed prompt $y_{\text{fix}} = P_1$ and a fresh sampler initialization at every iteration. Full hyperparameters are in App. E.5.

5.2 COMPARISON WITH EXISTING METHODS

In this section, we compare I2VShield with existing protections. As shown in Tab. 1, we conduct experiments on three I2V backbones (i.e., Wan2.1-I2V-14B, LTX-Video-2B-v0.9, and CogVideoX1.5-5B-I2V) and two datasets (i.e., HQFace-I2V and CelebV-Test (Yu et al., 2023)), where every method is implemented as a white-box attack against the target model and is projected onto the same ℓ_∞ ball of $\varepsilon = 16/255$, and each image is animated with a single prompt under a fixed seed. I2VShield wins all 30 backbone–dataset–metric cells, by wide margins: on HQFace-I2V it drives DINO to 0.11–0.27, IQ to 0.27–0.40, and SC to 0.59–0.61, while the strongest baseline stalls at 0.75, 0.57, and 0.92, up to $7\times$ higher. Averaged over the six settings, the best baseline erases 0.16 of the unprotected DINO 0.92 and I2VShield 0.73, a $4.6\times$ larger drop. Both LMM judges concur: on Wan, FineVQ falls $57.2 \rightarrow 26.0$ and VQA^2 $0.996 \rightarrow 0.601$, past the best baseline (38.3 and 0.710). CelebV-Test behaves identically, and Fig. 1 visualizes the Wan results on HQFace-I2V.

Notably, the six baselines differ in both attack surface and objective, yet all land in the same narrow band (DINO 0.75–0.90 on Wan and CogVideoX1.5), where I2VShield reaches 0.17 and 0.27. What decides protection strength is therefore where along the timestep axis the budget is spent, not which objective or attack surface spends it. More visualizations are in App. K.1.

Table 1: Comparison of I2VShield and baseline methods on three I2V backbones and two datasets under an identical perturbation budget $\varepsilon = 16/255$, averaged over 80 image–prompt pairs. Lower values indicate stronger protection for all five metrics. I2VShield outperforms every baseline on every backbone, dataset and metric. Best in **bold**, strongest baseline underlined.

	HQFace-I2V (ours)					CelebV-Test				
Method	DINO	IQ	SC	VQA^2	FineVQ	DINO	IQ	SC	VQA^2	FineVQ
Wan2.1-I2V-14B (Wan et al., 2025)										
Clean (no protection)	0.953	0.696	0.973	0.996	57.2	0.876	0.622	0.921	0.795	37.6
PhotoGuard (Salman et al., 2023)	0.854	0.644	0.943	0.769	44.5	0.793	0.558	0.896	0.641	33.2
AdvDM (Liang et al., 2023)	0.853	0.643	0.943	0.786	43.6	0.791	0.563	0.874	0.652	33.3
SDS+ (Xue et al., 2024)	0.846	0.634	0.928	0.777	43.2	0.776	0.551	0.871	0.657	32.5
Mist (Liang & Wu, 2023)	0.858	0.641	0.937	0.778	43.3	0.791	0.566	0.873	0.666	32.9
I2VGuard (Gui et al., 2025)	0.833	0.650	0.940	0.786	38.3	0.806	0.544	0.894	0.672	30.8
Anti-I2V (Vu et al., 2026)	0.804	0.611	0.941	0.710	41.3	0.757	0.551	0.875	0.636	32.8
I2VShield (ours)	**0.173**	**0.396**	**0.614**	**0.601**	**26.0**	**0.200**	**0.385**	**0.607**	**0.549**	**26.3**
LTX-Video-2B-v0.9 (HaCohen et al., 2024)										
Clean (no protection)	0.977	0.702	0.998	0.999	56.6	0.874	0.503	0.926	0.724	32.2
PhotoGuard (Salman et al., 2023)	0.840	0.630	0.975	0.700	43.0	0.630	0.425	0.845	0.585	25.0
AdvDM (Liang et al., 2023)	0.822	0.631	0.977	0.688	42.8	0.649	0.413	0.853	0.580	25.3
SDS+ (Xue et al., 2024)	0.833	0.634	0.981	0.708	43.8	0.624	0.424	0.837	0.578	25.2
Mist (Liang & Wu, 2023)	0.821	0.619	0.977	0.682	42.8	0.651	0.420	0.854	0.582	24.9
I2VGuard (Gui et al., 2025)	0.832	0.591	0.930	0.750	38.2	0.598	0.396	0.792	0.577	22.8
Anti-I2V (Vu et al., 2026)	0.753	0.570	0.961	0.640	32.7	0.517	0.420	0.774	0.577	23.5
I2VShield (ours)	**0.113**	**0.282**	**0.592**	**0.496**	**18.7**	**0.137**	**0.276**	**0.586**	**0.503**	**17.6**
CogVideoX1.5-5B-I2V (Yang et al., 2025)										
Clean (no protection)	0.945	0.688	0.967	0.997	64.1	0.886	0.600	0.930	0.776	40.0
PhotoGuard (Salman et al., 2023)	0.898	0.672	0.955	0.942	52.6	0.855	0.584	0.926	0.687	36.2
AdvDM (Liang et al., 2023)	0.898	0.678	0.929	0.869	53.5	0.846	0.579	0.909	0.683	39.7
SDS+ (Xue et al., 2024)	0.900	0.670	0.933	0.865	52.9	0.855	0.587	0.914	0.686	39.9
Mist (Liang & Wu, 2023)	0.900	0.677	0.931	0.864	53.1	0.846	0.576	0.910	0.685	39.6
I2VGuard (Gui et al., 2025)	0.903	0.622	0.956	0.879	46.4	0.861	0.529	0.934	0.671	34.4
Anti-I2V (Vu et al., 2026)	0.877	0.669	0.917	0.777	48.5	0.843	0.583	0.903	0.667	40.0
I2VShield (ours)	**0.271**	**0.269**	**0.599**	**0.656**	**29.0**	**0.212**	**0.259**	**0.577**	**0.599**	**26.0**

5.3 Generalization and Robustness

In this section, we sweep the perturbation budget and test unseen prompts, showing that the protection holds under both; further generalization and robustness results are in App. F.

Perturbation Budgets. A data owner may prefer a smaller perturbation for better imperceptibility, since images are published to be seen. We therefore experiment on Wan2.1-I2V-14B with HQFace-I2V (random 20 images): the I2VShield objective and optimizer are kept unchanged, and only the budget is swept over $\varepsilon \in \{2, 4, 8, 16, 32\}/255$, with Anti-I2V, the strongest baseline in Tab. 1, swept alongside. In Fig. 4 the y-axis reports DINO and the x-axis the budget (PSNR in parentheses). Our method outperforms Anti-I2V under every budget, and the two curves never approach each other: the gap between them widens from 0.50 at $\varepsilon = 2/255$ to 0.75 at $\varepsilon = 32/255$. At $\varepsilon = 2/255$, where the protected image stays at a PSNR of 45.2 dB and the perturbation is visually imperceptible, it already achieves a DINO of 0.426, lower than that of Anti-I2V at $\varepsilon = 32/255$ (0.808), a budget $16\times$ larger. Our method thus dominates the protection–imperceptibility trade-off, with visualizations at each budget provided in App. K.2.

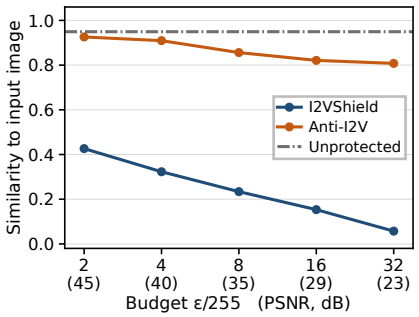

Figure 4: Generalization to lower ε.

Unseen Prompts. In a real-world scenario, the data owner does not know which prompt a malicious user will animate the protected image with, so the protection must hold under unseen prompts. To simulate this, we conduct this experiment on Wan2.1-I2V-14B with HQFace-I2V (80 images), which pairs every image with three prompts: the perturbation optimized under P_1 is reused unchanged,

Table 2: Generalization to unseen prompts on Wan2.1-I2V-14B with HQFace-I2V.

	DINO			IQ			SC			VQA^2			FineVQ		
Method	P_1	P_2	P_3	P_1	P_2	P_3	P_1	P_2	P_3	P_1	P_2	P_3	P_1	P_2	P_3
Unprotected	0.953	0.954	0.951	0.696	0.700	0.695	0.973	0.973	0.972	0.996	0.998	0.997	57.2	56.6	56.6
I2VShield (ours)	0.173	0.189	0.198	0.396	0.425	0.413	0.614	0.602	0.617	0.601	0.607	0.612	26.0	27.3	27.4

with no further optimization, so that the prompt driving the generation, moved to the held-out P_2 or P_3, is the only variable. The Unprotected row of Tab. 2 is the clean generation under the same prompt, and lower values indicate stronger protection throughout. Our method remains effective under all five metrics: the DINO similarity increases only from 0.173 to 0.189 (P_2) and 0.198 (P_3) and the IQ from 0.396 to 0.425 and 0.413, against an unprotected level that is itself flat across the three prompts, while the SC and the two LMM judges are nearly unchanged and stay far below their unprotected level. I2VShield is therefore shown to generalize to whatever prompt the attacker chooses, not just the one it was optimized under.

5.4 Ablation Study

Attacked Timestep. We attribute our method's effectiveness to concentrating the entire budget on the largest noise level. To verify this, we conduct this experiment on Wan2.1-I2V-14B with HQFace-I2V (random 10 images): the I2VShield objective and optimizer are kept unchanged, and only the attacked timestep $\tau^\star$ of Alg. 1 is moved to an interior step, reached by rerolling the trajectory at every optimization iteration. In Fig. 5 the curve reports DINO averaged over the ten images and the dash-dotted line the unprotected level. Attacking $\tau^\star$ achieves a DINO of 0.089, whereas protection weakens as the attacked step moves toward lower noise: the interior steps rise to 0.48–0.74 and the smallest noise level degrades to 0.85, close to the clean value of 0.95, while the imperceptibility remains nearly identical, at a PSNR of 28.8–29.9 dB throughout. Protection thus decays along the sampler step in the same direction as the conditioning attention does in Fig. 3(b), corroborating Diagnosis 2. The gain therefore stems from where along the trajectory the budget is spent, not from the perturbation's magnitude.

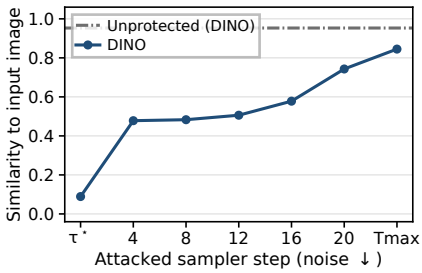

Figure 5: Ablation of the sampler step.

Auxiliary Objectives. Since I2VShield removes the components inherited from the T2I recipe, a natural question is whether any should be added back. We therefore conduct this experiment on Wan2.1-I2V-14B with HQFace-I2V (random 20 images), attaching the characteristic signal of each baseline to our objective at $\tau^\star$. The added terms are a semantic denoising term, Anti-I2V's internal representation anchor (IRA) and collapse (IRC), and Mist's texture term. The exact form of each is given in App. E.6. The budget and generation configuration are kept unchanged, so the objective is the only variable. As shown in Tab. 3, every added term either stays within noise of the single-term objective or degrades it, the IRC and texture terms clearly so, demonstrating that the objective in Eq. (4) is already sufficient.

Table 3: Ablation of auxiliary objectives.

Objective	DINO	SC
$\mathcal{A}$ (ours)	**0.153**	**0.582**
+ semantic	**0.153**	0.602
+ IRA	**0.153**	0.584
+ IRC	0.179	0.588
+ texture	0.189	0.605

6 Conclusion

We revisit adversarial protection against malicious I2V generation and show that what decides protection strength is not which objective is attacked, but where along the timestep axis the budget is spent. I2VShield acts on this, committing the entire budget to the largest noise level. It wins all 30 backbone–dataset–metric cells of our main comparison, cutting DINO from 0.92 to 0.18 against 0.80 for the strongest baseline, a $4.6\times$ larger drop, and still beats it at a $16\times$ smaller budget. We further release HQFace-I2V, and hope this work invites further study of protection in the I2V setting.

REFERENCES

AI Incident Database. Incident 1165: Grok Imagine reportedly produces non-consensual Taylor Swift deepfake nudes without explicit prompting, 2025. URL https://incidentdatabase.ai/cite/1165/.

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, 2023. URL https://arxiv.org/abs/2311.15127.

ByteDance Seed. Seed1.8 model card: Towards generalized real-world agency, 2026. URL https://arxiv.org/abs/2603.20633.

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

Guibin Chen, Dixuan Lin, Jiangping Yang, Chunze Lin, Junchen Zhu, et al. SkyReels-V2: Infinite-length film generative model. *arXiv preprint arXiv:2504.13074*, 2025.

Rohit Chowdhury, Aniruddha Bala, Rohan Jaiswal, and Siddharth Roheda. Vid-freeze: Protecting images from malicious image-to-video generation via temporal freezing. *arXiv preprint arXiv:2509.23279*, 2025.

CNN Business. Arup revealed as victim of $25 million deepfake scam involving Hong Kong employee, May 2024. URL https://www.cnn.com/2024/05/16/tech/arup-deepfake-scam-loss-hong-kong-intl-hnk.

Yingqian Cui, Jie Ren, Yuping Lin, Han Xu, Pengfei He, Yue Xing, Lingjuan Lyu, Wenqi Fan, Hui Liu, and Jiliang Tang. Ft-shield: A watermark against unauthorized fine-tuning in text-to-image diffusion models. *SIGKDD Explor.*, 26:76–88, 2024.

Huiyu Duan, Qiang Hu, Jiarui Wang, Liu Yang, Zitong Xu, Lu Liu, Xiongkuo Min, Chunlei Cai, Tianxiao Ye, Xiaoyun Zhang, and Guangtao Zhai. FineVQ: Fine-grained user generated content video quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3206–3217, 2025.

Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2426–2436, 2023.

Dongnan Gui, Xun Guo, Wengang Zhou, and Yan Lu. I2vguard: Safeguarding images against misuse in diffusion-based image-to-video models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12595–12604, 2025. doi: 10.1109/CVPR52734.2025.01175.

Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, Poriya Panet, Sapir Weissbuch, Victor Kulikov, Yaki Bitterman, Zeev Melumian, and Ofir Bibi. Ltx-video: Realtime video latent diffusion, 2024. URL https://arxiv.org/abs/2501.00103.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21807–21818, 2024.

Ziheng Jia, Zicheng Zhang, Jiaying Qian, Haoning Wu, Wei Sun, Chunyi Li, Xiaohong Liu, Weisi Lin, Guangtao Zhai, and Xiongkuo Min. VQA^2: Visual question answering for video quality assessment. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025. doi: 10.1145/3746027.3754696.

Boheng Li, Yanhao Wei, Yankai Fu, Zhenting Wang, Yiming Li, Jie Zhang, Run Wang, and Tianwei Zhang. Towards Reliable Verification of Unauthorized Data Usage in Personalized Text-to-Image Diffusion Models . In *2025 IEEE Symposium on Security and Privacy (SP)*, pp. 2564–2582, Los Alamitos, CA, USA, May 2025. IEEE Computer Society. doi: 10.1109/SP61157.2025.00073.

Chumeng Liang and Xiaoyu Wu. Mist: Towards improved adversarial examples for diffusion models, 2023. URL https://arxiv.org/abs/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 *Proceedings of the 40th International Conference on Machine Learning*, pp. 20763–20786, 2023.

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*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

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

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=a68SUt6zFt. Featured Certification.

Yan Pang, Baicheng Chen, Yang Zhang, and Tianhao Wang. Vgmshield: Mitigating misuse of video generative models, 2025. URL https://arxiv.org/abs/2402.13126.

Pexels. Pexels — free stock photos & videos, 2026. URL https://www.pexels.com.

Ed Pizzi, Sreya Dutta Roy, Sugosh Nagavara Ravindra, Priya Goyal, and Matthijs Douze. A self-supervised descriptor for image copy detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14512–14522, 2022. doi: 10.1109/CVPR52688.2022.01413.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=di52zR8xgf.

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

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, June 2022.

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.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 29894–29918. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/salman23a.html.

Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22522–22531. IEEE, 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 *Proceedings of the 32nd USENIX Conference on Security Symposium*, SEC '23, USA, 2023. USENIX Association. ISBN 978-1-939133-37-3.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=St1giarCHLP.

Duc Vu, Anh Nguyen, Chi Tran, and Anh Tran. Anti-i2v: Safeguarding your photos from malicious image-to-video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 37621–37631, June 2026.

Team Wan, 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, Mengyang Feng, 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. *arXiv preprint arXiv:2503.20314*, 2025.

Zhenting Wang, Chen Chen, Lingjuan Lyu, Dimitris N. Metaxas, and Shiqing Ma. DIAGNOSIS: Detecting unauthorized data usages in text-to-image diffusion models. In *The Twelfth International Conference on Learning Representations*, 2024.

Yanhao Wei, Xiaokang Zhao, Boheng Li, Yang Zhang, and Run Wang. RECOVER: Reliable detection of unauthorized data usage in text-to-image diffusion models via inversion robustness. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=HYZ3ooGuew.

Xiaoyu Wu, Yang Hua, Chumeng Liang, Jiaru Zhang, Hao Wang, Tao Song, and Haibing Guan. CGI-DM: Digital Copyright Authentication for Diffusion Models via Contrasting Gradient Inversion . In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10812–10821, Los Alamitos, CA, USA, June 2024. IEEE Computer Society. doi: 10.1109/CVPR52733.2024.01028.

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*, pp. 399–417. Springer, 2024.

Naen Xu, Jinghuai Zhang, Changjiang Li, Zhi Chen, Chunyi Zhou, Qingming Li, Tianyu Du, and Shouling Ji. Videoeraser: Concept erasure in text-to-video diffusion models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 5965–5994, 2025.

Haotian Xue, Chumeng Liang, Xiaoyu Wu, and Yongxin Chen. Toward effective protection against diffusion-based mimicry through score distillation. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=NzxCMe88HX.

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*, 2025. URL https://openreview.net/forum?id=LQzN6TRFg9.

Xiaoyu Ye, Songjie Cheng, Yongtao Wang, Yajiao Xiong, and Yishen Li. T2vunlearning: A concept erasing method for text-to-video diffusion models. *arXiv preprint arXiv:2505.17550*, 2025.

Jaehong Yoon, Shoubin Yu, Vaidehi Patil, Huaxiu Yao, and Mohit Bansal. SAFREE: Training-free and adaptive guard for safe text-to-image and video generation. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=hgTFotBRKl.

Jianhui Yu, Hao Zhu, Liming Jiang, Chen Change Loy, Weidong Cai, and Wayne Wu. CelebV-Text: A large-scale facial text-video dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19305–19314, 2023.

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/CVF conference on computer vision and pattern recognition*, pp. 586–595. IEEE, 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 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

A Further Related Work

A.1 Safety Protections against NSFW Generation

A parallel line of work prevents the malicious use of generative models at the *model level*, with NSFW content as the primary target. In T2I models, Safe Latent Diffusion (SLD) (Schramowski et al., 2023) steers the denoising trajectory away from inappropriate concepts through training-free safety guidance, and Erased Stable Diffusion (ESD) (Gandikota et al., 2023) erases concepts by editing model weights. This paradigm has been extended to video generation, where SAFREE (Yoon et al., 2025) offers a training-free guard for both T2I and T2V models, while VideoEraser (Xu et al., 2025) and T2VUnlearning (Ye et al., 2025) erase concepts from T2V models in a training-free and a fine-tuning-based manner, respectively. All these approaches, however, act on the model side—modifying weights, filtering prompts, or steering sampling—whereas the defender in our setting is the *data owner*, who controls only the image before it is published. These model-level safeguards are therefore out of our scope.

A.2 Ownership Verification and Provenance Tracing

A second line of defenses detects, after the fact, whether a diffusion model has been trained or fine-tuned on protected data. **Watermark-based** methods embed an invisible coating into images before release, so that models trained on them reproduce the signal, which is then detected by hypothesis testing: DIAGNOSIS (Wang et al., 2024) injects memorization signals, FT-Shield (Cui et al., 2024) ties the watermark to the fine-tuning objective so that it is learned early, and SIREN (Li et al., 2025) optimizes the coating to survive personalization, enabling black-box verification. **Watermark-free** methods instead invert the suspicious model: CGI-DM (Wu et al., 2024) authenticates training data by contrasting gradient inversion against a clean reference model, and RECOVER (Wei et al., 2026) exploits the inversion robustness that models acquire for images they were fine-tuned on, requiring neither image modification nor a reference model.

Ownership verification, however, remains complementary to—rather than a substitute for—adversarial protection. Detection is reactive: it confirms misuse only after training and, typically, after harmful content has been generated. Verification further yields only evidence of infringement, with recourse still depending on legal enforcement. Adversarial protection, in contrast, is proactive and data-level: the perturbation travels with the image and degrades generation in any model conditioned on it, preventing the misuse itself rather than detecting it afterwards.

B HQFACE-I2V: BENCHMARK DETAILS

Overview. HQFace-I2V consists of 80 high-quality real-world portrait photographs, 240 English animation prompts (3 per image), and 10 reference videos produced by a commercial I2V model. The 19 candidates rejected during curation are retained as negative examples. All images are offered under the permissive Pexels license (Pexels, 2026), and the benchmark is intended solely for developing and evaluating defenses against malicious I2V generation.

Image collection and filtering. Candidate photographs were retrieved from the Pexels API (Pexels, 2026) using 20 semantically diverse search phrases (e.g., "natural face portrait daylight" and "casual headshot outdoor natural light"), with landscape orientation, a large-size filter, and a minimum source resolution of 1280×720. Every candidate was then verified by a vision-language model, Doubao-1.8 Vision Pro (ByteDance Seed, 2026), which accepted an image only if all six criteria held: (i) a color photograph; (ii) a real human face, excluding animals, statues, and cartoons; (iii) a close-up portrait rather than a full-body or distant view; (iv) sufficient natural lighting with clearly visible facial details; (v) a natural, minimally processed style; and (vi) an everyday background rather than a pure studio backdrop. 80 images passed verification and 19 were rejected. Source resolutions range from 5523×4000 to 10000×6667; the accepted images come from 61 distinct photographers and depict subjects of both genders and varied apparent ages and ethnicities in outdoor, indoor, and urban settings. All images were center-cropped to a 16:9 aspect ratio and resized to 1280×720.

Prompt design. For each accepted image, the same VLM generated three English animation prompts grounded in the subject's actual appearance (gender, age, hairstyle, clothing) and covering three qualitatively distinct motion types: expression-driven motion (smiling, blinking, speaking), head and gaze movement (turning, glancing, tilting), and environmental dynamics (hair in breeze, shifting light). The 240 resulting prompts are 30–51 words long. This multi-prompt design supports cross-prompt evaluation, enabling a more comprehensive assessment under a realistic threat model in which the data owner has no knowledge of the prompts used by infringers for I2V generation.

Reference videos. Ten of the 240 image–prompt pairs were rendered with a commercial I2V service, Doubao-Seedance-1.0-Pro-Fast, with a 5-second duration, an adaptive aspect ratio (yielding 1920×1088 output), an unfixed camera, and no watermark. The videos cover 4 images (three images with all three prompts, and one image with its first prompt) and are included to illustrate the quality of the benchmark inputs under a state-of-the-art generator.

Comparison with prior evaluation resources. As summarized in the main paper, prior I2V protection works either do not release their evaluation data (Gui et al., 2025; Chowdhury et al., 2025) or repurpose datasets assembled for unrelated tasks, whose inputs are video frames paired with a single prompt (Vu et al., 2026). HQFace-I2V instead provides real photographs, three prompts per identity, and full public release, enabling the cross-prompt evaluation these suites cannot support.

Ethics and intended use. The photographs are offered by Pexels under a license that permits free use and modification without attribution. HQFace-I2V is released exclusively for research on defending against malicious image-to-video synthesis, and must not be used to train generative models whose purpose is to synthesize video of real, identifiable individuals.

C DETAILS OF THE DIAGNOSTIC EXPERIMENTS

This section gives the implementation details of the diagnostic experiments of Sec. 4.2 and reports the full measurements behind them. Unless stated otherwise, the I2V model is Wan2.1-I2V-14B and the images are drawn from HQFace-I2V with the prompt P_1.

C.1 VAE ROBUSTNESS

Implementation Details. We attack the encoder of four autoencoders directly: the 2D SD1.5 and SD2.1 VAEs, and the 3D Wan2.1 and CogVideoX1.5 VAEs. All models use the same sign-PGD attack, with 100 iterations, random initialization, step size $\varepsilon/16$, and budgets $\varepsilon \in \{4, 8, 16, 32, 64\}/255$. The objective maximizes $\|E(x+\delta) - E(x)\|_2^2$ in latent space. We use the $F = 1$ capacity-control setting for the 3D VAEs, feed the same 20 images at 512×512, and

report means over the 20 examples. This control substantially reduces the latent-capacity advantage that the 3D VAEs receive when a single image is repeated over $F = 17$ frames. The primary metric is DINO similarity (Oquab et al., 2024) between the original image and the reconstruction of the perturbed image; lower similarity means a more successful attack. We additionally report CLIP-I similarity (Radford et al., 2021), SSCD copy-detection similarity (Pizzi et al., 2022), and pixel-level LPIPS (Zhang et al., 2018). For CLIP, SSCD, and DINO, larger similarity indicates greater retained content; for LPIPS, a smaller distance indicates greater retained content.

Additional Results. DINO is a strong perceptual measure, but a conclusion about encoder robustness should not depend on a single feature space. CLIP captures semantic and visual similarity, SSCD is designed to detect whether two images remain copies of the same content, and LPIPS measures perceptual distance directly in pixel-derived feature space. We therefore use these complementary metrics to test whether the DINO ordering reflects a broader phenomenon. Table 4 reports the complete $F = 1$ control grid for the four VAEs, the five perturbation budgets, and all four metrics. Each entry is the mean over 20 images. The similarity rows compare the original image with the adversarial reconstruction, whereas the LPIPS rows use the same pair and therefore increase as the attack moves the reconstruction farther from the source. The three additional metrics agree with DINO at every budget: the 3D VAEs retain substantially higher similarity than the 2D VAEs. At $\varepsilon = 16/255$, the mean DINO similarity is 0.737 for the 3D VAEs versus 0.556 for the 2D VAEs; the corresponding CLIP values are 0.789 versus 0.682, and SSCD values are 0.807 versus 0.660. LPIPS shows the same separation in the opposite direction: 0.678 for 3D versus 0.868 for 2D. At $\varepsilon = 64/255$, the DINO gap widens to 0.389 versus 0.134, while LPIPS reaches 0.996 versus 1.136. Thus, the conclusion that 3D VAE encoders absorb more of the input perturbation is not specific to DINO or to one perceptual representation.

Table 4: Encoder robustness in the $F = 1$ capacity-control experiment. Entries are means over 20 images for the reconstruction of the perturbed image against the original image. DINO, CLIP-I, and SSCD are cosine similarities (higher is more robust); LPIPS is a perceptual distance (lower is more robust). The attack uses the same latent-space sign-PGD objective at all budgets.

Autoencoder	Metric	Clean	$\varepsilon = 4$	8	16	32	64
SD1.5 VAE (Rombach et al., 2022)	DINO	0.918	0.825	0.708	0.559	0.365	0.127
	CLIP-I	0.898	0.807	0.736	0.672	0.643	0.582
	SSCD	0.980	0.887	0.791	0.663	0.495	0.297
	LPIPS	0.035	0.415	0.652	0.886	1.072	1.200
SD2.1 VAE (Rombach et al., 2022)	DINO	0.921	0.829	0.708	0.552	0.360	0.141
	CLIP-I	0.903	0.812	0.764	0.693	0.642	0.577
	SSCD	0.987	0.893	0.793	0.656	0.499	0.308
	LPIPS	0.029	0.417	0.655	0.850	0.989	1.073
Wan VAE (Wan et al., 2025)	DINO	0.920	0.901	0.848	0.742	0.583	0.367
	CLIP-I	0.905	0.863	0.820	0.765	0.707	0.659
	SSCD	0.997	0.957	0.907	0.809	0.665	0.468
	LPIPS	0.013	0.224	0.470	0.691	0.831	0.949
CogVideoX VAE (Yang et al., 2025)	DINO	0.921	0.891	0.817	0.732	0.605	0.412
	CLIP-I	0.905	0.882	0.858	0.814	0.747	0.689
	SSCD	0.997	0.961	0.907	0.806	0.652	0.442
	LPIPS	0.013	0.225	0.425	0.665	0.889	1.042

C.2 Gradient Alignment Across Timesteps

Implementation Details. We compare a T2I regime, Stable Diffusion v1.5 (Rombach et al., 2022) at 512×512 in fp32, with an I2V regime, Wan2.1-I2V-14B. In each regime we sample nine timesteps uniformly over the model's own schedule, since the two noise schedules are not commensurable. At each of them we take the gradient $g_t = \nabla_\delta \mathcal{A}(\epsilon_\theta, z_t, c^\delta_{\text{img}}, y, t)$ of the same per-timestep objective, average it over $K = 16$ noise draws so that the comparison is not dominated by the sampling noise of a single draw, flatten it, and compute the cosine similarity of all 36 unordered pairs. The gradients are taken at the original image, the common starting point of every method under comparison, so that the two regimes are read at the same point rather than wherever their optimizers have since moved. Two settings are specific to this diagnostic. We disable classifier-free guidance, so that a

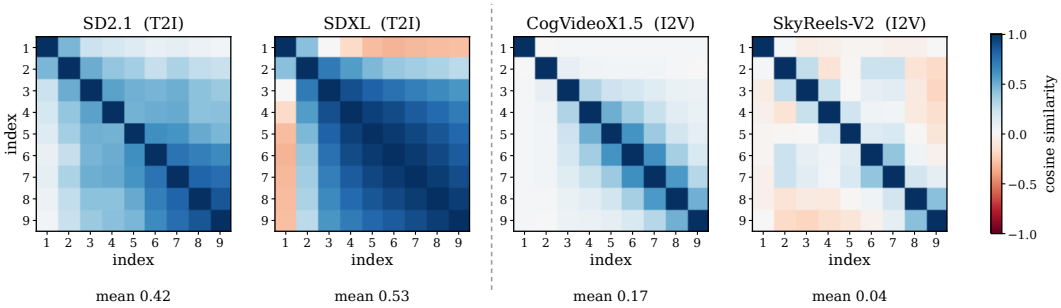

Figure 6: **Gradient alignment on four further models.** Pairwise cosine between the gradients measured at the nine indices of each model's own schedule, taken at the original image under the protocol of App. C.2 and averaged over 20 images. Entry (i, j) is the cosine between the gradients at index i and index j, so the matrix is symmetric and the diagonal is 1 by construction. Indices are numbered in the order they are sampled, and the two regimes are not compared index by index, only by how dispersed each matrix is. All panels share one colour scale with zero at its neutral midpoint, so blue is positive alignment and red negative. The mean over the 36 unordered off-diagonal pairs appears below each panel.

single conditional branch is measured, and we let δ reach the loss only through the image-condition pathway, the conditioning latent and its CLIP embedding, holding z_0 frozen, which is the pathway an I2V protection actually controls. Wan is run at 9 frames rather than the 17 of the main experiments; its first latent frame, which is pinned to the conditioning image, is excluded from the loss, so that the gradient reflects how the condition drives the generated frames rather than how well it reproduces itself. Everything else is held fixed across timesteps, including the image, the prompt, and the sampler initialization. We use 20 images at a fixed seed.

Additional Results. The contrast in Fig. 3(a) is read off a single model per regime. On its own it cannot separate a property of the regime from a property of the two particular backbones that carry it, and the design decision it supports—evaluating at one point rather than averaging—rests on that separation. We therefore repeat the measurement on four further models, two of each regime: the T2I models SD2.1 (Rombach et al., 2022) and SDXL (Podell et al., 2024), and the I2V models CogVideoX1.5-5B-I2V (Yang et al., 2025) and SkyReels-V2-I2V-1.3B-540P (Chen et al., 2025). The protocol is the one above, unchanged: nine indices along each model's own schedule, the same per-index objective averaged over $K = 16$ noise draws, gradients taken at the original image, and the cosine computed over all 36 unordered pairs.

Figure 6 reports the four matrices. The two regimes separate as clearly as in Fig. 3(a). The T2I matrices are positively aligned almost everywhere and dark away from the diagonal, with mean cosines of 0.42 on SD2.1 and 0.53 on SDXL, whereas the I2V matrices are pale over most of their area, at 0.17 on CogVideoX1.5 and 0.04 on SkyReels-V2; half of SkyReels-V2's pairs are negative. Taken together with SD1.5 at 0.60 and Wan2.1 at 0.07, the six models give 0.42 to 0.60 for T2I against 0.04 to 0.17 for I2V, two ranges that do not overlap and are separated by a gap wider than the spread within either.

The dispersion is therefore a property of the I2V regime rather than of any one backbone: it holds across two architectures and an order of magnitude of scale, and its T2I counterpart holds across three model generations. Averaging over the axis dilutes the attack signal in every I2V model we measure, which is what the point evaluation of Sec. 4.3 is built on.

C.3 Condition Influence Across Timesteps

Implementation Details. Wan2.1-I2V-14B encodes the conditioning image into the first latent frame, so the attention that later frames pay to the first frame's tokens is a direct read-out of how strongly the condition drives generation. We record the frame-to-frame attention matrix of transformer blocks 24 and 39 at seven points of the 25-step sampling schedule, reduced over heads and over the spatial tokens within a frame. Rows are normalized, so the uniform reference is $1/F$ with

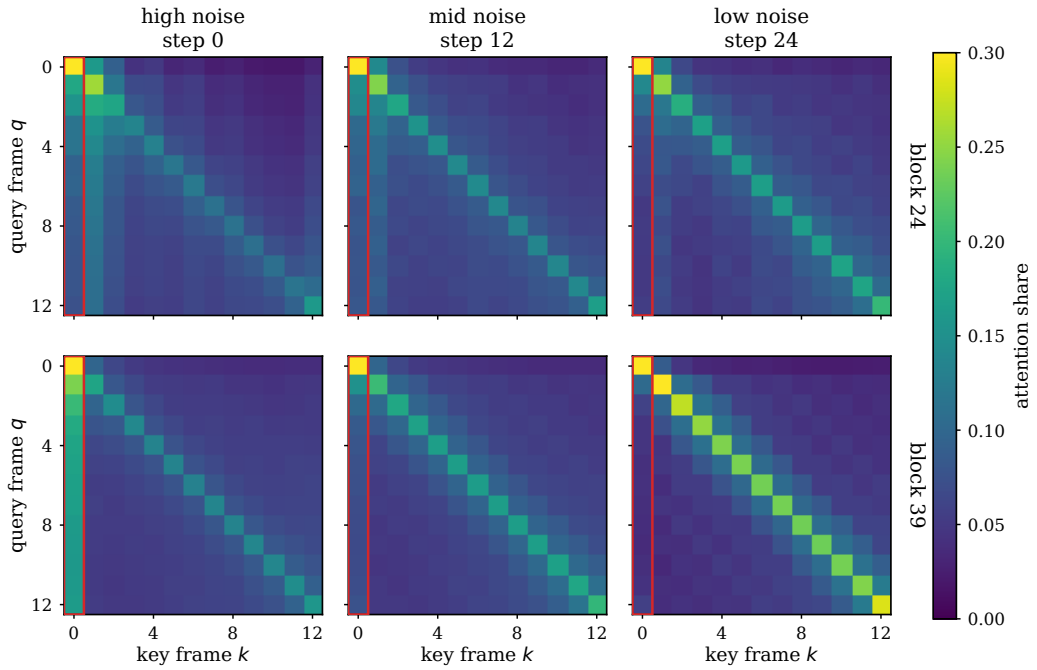

Figure 7: Frame-to-frame self-attention in blocks 24 (top) and 39 (bottom) of Wan2.1-I2V-14B, averaged over the 20 HQFace-I2V images at $F = 13$ latent frames (conditional branch). Columns are three points of the 25-step sampling schedule, from the largest to the smallest noise level. Entry (q, k) is the share of attention that query frame q directs at key frame k; rows sum to one, so the uniform reference is $1/F = 0.077$. The conditioning frame is the first column, outlined in red. All six panels share one colour scale, from which the diagonal cell $(0, 0)$ is excluded and clipped, since the conditioning frame attending to itself runs far above every other cell and says nothing about how the remaining frames use the condition.

F the number of latent frames. The diagnostic runs at $F = 13$ latent frames, and we average the matrices over 20 images on the conditional branch.

Additional Results. The main paper summarizes this measurement with one number per sampling step, the share of attention that goes to the conditioning frame. That number says how much attention leaves the conditioning frame but not where it goes instead, and it compresses a whole matrix into a scalar. We therefore report the matrices themselves, which show both at once and make the movement directly visible. Figure 7 gives the averaged attention matrices of blocks 24 and 39 at the largest noise level, an intermediate step, and the smallest noise level; block 39 is the block whose conditioning column decays most, and block 24 is shown so that the effect is not read off a single depth. Entry (q, k) is the share of attention that frame q directs at frame k, and the conditioning column is outlined in red; the statistics below exclude the cell $(0, 0)$, as the colour scale of Fig. 7 does. At the largest noise level that column is the dominant structure of the matrix: in block 39 it carries $2.3\times$ the uniform share and is $2.9\times$ brighter than the off-diagonal background, against $2.3\times$ for the main diagonal. As denoising proceeds the two exchange roles. By the last sampling step the column of block 39 has fallen to $0.79\times$ the background, below it rather than above, while the diagonal has risen to $4.0\times$; block 24 moves the same way, its column falling from $1.50\times$ to $1.00\times$ the background. Since the image condition enters the network only through the first frame, the attention that frame receives is the channel through which the condition acts, and that channel is widest at the largest noise level and closes as sampling proceeds.

D Overall Algorithm

Algorithm 1 I2VShield

Require: image x, budget ε, steps N, step size α, prompt y_{fix}, timestep $\tau^\star = 1$

$\delta \leftarrow \mathcal{U}(-\varepsilon, \varepsilon)$
for $n = 1$ to N **do**
  $c^{\delta}_{\text{img}} \leftarrow E(x + \delta)$ ▷ image condition, independent of τ
  $\epsilon \sim \mathcal{N}(0, I)$; $z_1 \leftarrow \epsilon$ ▷ sampler initialization
  $z_{\tau^\star} \leftarrow \text{Rollout}\big(\epsilon_\theta,\, z_1,\, c^{\delta}_{\text{img}},\, y_{\text{fix}},\, 1 \to \tau^\star\big)$ ▷ identity when $\tau^\star = 1$
  $g \leftarrow \nabla_\delta \mathcal{A}\big(\epsilon_\theta, z_{\tau^\star}, c^{\delta}_{\text{img}}, y_{\text{fix}}, \tau^\star\big)$
  $\delta \leftarrow \Pi_{\|\cdot\|_\infty \le \varepsilon}\big(\delta + \alpha \cdot \text{sign}(g)\big)$
end for
return protected image $x + \delta$

E Experiment Setup Details

E.1 Models

All experiments use publicly available I2V releases, without any fine-tuning of our own. Table 5 lists the nine checkpoints, which span 1.3B to 27B parameters and four model families. Three of them carry the main comparison of Tab. 1, one per family. **Wan2.1-I2V-14B** (Wan et al., 2025) is a 14B-parameter flow-matching diffusion transformer whose conditioning image enters through channel-wise concatenation of VAE latents together with CLIP image embeddings, and is sampled with an Euler flow-matching scheduler. **LTX-Video** (HaCohen et al., 2024) is a 2B-parameter flow-matching transformer whose reference image is encoded by its VAE and prepended to the token sequence; it applies its own dynamic time shifting during sampling and uses the negative prompt "worst quality, inconsistent motion, blurry, jittery, distorted" recommended by its authors. **CogVideoX1.5-5B-I2V** (Yang et al., 2025) is a 5B-parameter diffusion transformer trained with v-prediction, which pins the encoded condition along the frame axis of the video latent and is sampled with a DDIM scheduler.

The remaining six checkpoints are the additional models evaluated in App. F.1, chosen so that the three main backbones are not the only evidence for any claim. Three of them vary the checkpoint within a family we already attack: **CogVideoX-5B-I2V** (Yang et al., 2025) is the earlier v-prediction model of the CogVideoX family, and **LTX-Video 0.9.5** and **LTX-Video 0.9.7** are later LTX releases,

the second of which also scales the transformer from 2B to 13B parameters. Two vary the generation within the Wan family: **Wan2.2-TI2V-5B** is the newer Wan release, which replaces the autoencoder with a more strongly compressing one and handles text-to-video and image-to-video in a single model, and we additionally attack a publicly available NSFW fine-tune of **Wan2.2-I2V-A14B**, a two-expert mixture-of-experts model with 27B total and 14B active parameters per step, in a release distilled for few-step sampling. It is the closest available proxy for the deployed systems that motivate this work. The last is **SkyReels-V2-I2V-1.3B-540P** (Chen et al., 2025), a separately trained I2V family an order of magnitude smaller than Wan2.1-I2V-14B, which lets us check that the protection does not depend on model scale.

Table 5: The nine public I2V checkpoints used in this work, none of them fine-tuned by us.

Model	Hugging Face checkpoint	Params
Main comparison (Tab. 1)		
Wan2.1-I2V-14B (Wan et al., 2025)	`Wan-AI/Wan2.1-I2V-14B-480P`	14B
LTX-Video 0.9 (HaCohen et al., 2024)	`Lightricks/LTX-Video`, `ltx-video-2b-v0.9`	2B
CogVideoX1.5-5B-I2V (Yang et al., 2025)	`zai-org/CogVideoX1.5-5B-I2V`	5B
Additional models (App. F.1)		
CogVideoX-5B-I2V (Yang et al., 2025)	`zai-org/CogVideoX-5b-I2V`	5B
LTX-Video 0.9.5 (HaCohen et al., 2024)	`Lightricks/LTX-Video-0.9.5`	2B
LTX-Video 0.9.7 (HaCohen et al., 2024)	`Lightricks/LTX-Video-0.9.7-dev`	13B
Wan2.2-TI2V-5B (Wan et al., 2025)	`Wan-AI/Wan2.2-TI2V-5B-Diffusers`	5B
Wan2.2-I2V-A14B, NSFW	`fdk6566/wan2.2_14b_i2v_480p_lightning_nsfw_diffusers`	27B
SkyReels-V2-I2V-1.3B-540P (Chen et al., 2025)	`Skywork/SkyReels-V2-I2V-1.3B-540P`	1.3B

Table 6 lists the generation configuration of the three backbones of the main comparison, which is identical for clean and protected inputs, and every video is generated with a fixed seed of 42. One configuration deviates from the official defaults for comparability: CogVideoX1.5-5B-I2V generates 17 frames with 25 inference steps (official defaults: 81 frames and 50 steps), matching the frame and step counts used for the other backbones.

Table 6: Generation configuration of the three backbones of the main comparison, kept identical for clean and protected inputs.

Model	Resolution	Frames	Steps	CFG (shift)	fps
Wan2.1-I2V-14B (Wan et al., 2025)	480×832	17	25	5.0 (5.0)	16
LTX-Video (HaCohen et al., 2024)	480×704	41	25	3.0 (dynamic)	24
CogVideoX1.5-5B-I2V (Yang et al., 2025)	768×1360	17	25	6.0 $(-)^{\dagger}$	8

† DDIM scheduler (v-prediction), which has no shift parameter.

E.2 Datasets

HQFace-I2V comprises 80 high-resolution real-world portraits offered under the Pexels license (Pexels, 2026), center-cropped to a 16:9 aspect ratio and resized to 1280×720, each paired with three VLM-generated animation prompts (30–51 words, mean 41.6) covering expression-driven motion, head and gaze movement, and environmental dynamics; full curation details are in App. B.

CelebV-Test is drawn from CelebV-Text (Yu et al., 2023). We follow Anti-I2V (Vu et al., 2026), the strongest prior protection, which evaluates on this source, so that the comparison is not confined to the benchmark we release ourselves. CelebV-Text collects 70,000 in-the-wild face video clips, each annotated with 20 sentences covering six attribute groups, general and detailed appearance, lighting condition and direction, action, and emotion, for 1.4M descriptions in total. Its clips are in-the-wild footage rather than curated photography, so they are lower in resolution and less controlled than the portraits of HQFace-I2V, which is the contrast we want from a second dataset. We take the first frame of 80 clips as the conditioning images and pair each with one of that clip's captions as the prompt, giving 80 image–prompt pairs to match HQFace-I2V.

Usage. The main comparison uses one prompt per image—P_1, the first of the three prompts—on both datasets (80 image–prompt pairs each); the remaining two prompts of HQFace-I2V are used in the cross-prompt evaluation. Before generation, every image is resized to the native input

resolution of the target model (Table 6). The same 80 images are used across all experiments, so clean and protected videos are directly comparable. The supplementary experiments (perturbation-budget sweep, purification robustness, inference-parameter ablation, prompt transfer, and cross-model transfer) run on HQFace-I2V with P_1 unless stated otherwise.

E.3 BASELINES

Every baseline is implemented as a white-box attack on the target model and is projected onto the same ℓ_∞ ball of radius $\varepsilon = 16/255$ in pixel space, so that all methods are compared under an identical perturbation budget. All of them are optimized by sign-gradient descent with step size $\varepsilon/16$ and a random start inside the ball, under the same fixed seed, and each retains the objective and iteration count of its original implementation. We write $c^\delta = E(x+\delta)$ for the conditioning latent of the perturbed image, x_{tgt} for the fixed target image used by the targeted encoder attacks, and $\epsilon_\theta(z_t, c^\delta, y, t)$ for the network output, which is a velocity prediction on the flow-matching backbones and a noise prediction otherwise.

Provenance. PhotoGuard, AdvDM, SDS+ and Mist were designed for T2I, where the protected image is itself the latent being denoised. We port them to I2V with the smallest change that makes them applicable: the perturbation now enters through the image condition c^δ instead of through the noisy latent, and the latent their objectives are evaluated on is the video latent of the target model. Their loss functions and iteration counts are otherwise those of their public releases. I2VGuard and Anti-I2V have no public code, so we reproduce them from their papers, following the loss terms, weightings and iteration counts reported there.

Sampling trajectories. The objectives below are written against a noisy video latent z_t and the noise target that produced it. In I2V that pair does not exist before generation: only the conditioning image is given, and the clean video latent it should be noised from has to be produced by running the sampler first. Every baseline that touches the denoising network therefore rests on a rollout, and PhotoGuard is the sole exception, since its objective never leaves the encoder. In our implementation the rollout is not repeated at every PGD iteration: a bank of clean sampling trajectories is generated once and shared by every baseline, so that all methods are evaluated on identical latents and differ only in their objective, with the reference step taken at the midpoint of the schedule. I2VShield evaluates its objective at $\tau^\star = 1$, where the latent is the sampler's own Gaussian initialization.

PhotoGuard (Salman et al., 2023) is an encoder-side attack that does not touch the denoising network. It pulls the conditioning latent toward that of a fixed target image,

$$\min_{\|\delta\|_\infty \le \varepsilon} \big\|E(x+\delta) - E(x_{\mathrm{tgt}})\big\|_2^2, \tag{1}$$

so each of its 100 iterations costs one encoder forward and backward pass and no denoising call.

AdvDM (Liang et al., 2023) is the denoising-side counterpart. It maximizes the denoising prediction error at uniformly sampled timesteps,

$$\max_{\|\delta\|_\infty \le \varepsilon} \mathbb{E}_{t,\,\epsilon}\big\|\epsilon_\theta(z_t, c^\delta, y, t) - \epsilon\big\|_2^2, \tag{2}$$

for 100 iterations, so the budget is spread over the whole timestep axis. This is the instance of Eq. (3) with $\mathcal{V} \equiv 0$ that Diagnosis 2 addresses directly.

SDS+ (Xue et al., 2024) attacks the same signal as AdvDM. The two maximize the same quantity and differ only in how the ascent direction is obtained: Score Distillation Sampling replaces backpropagation through the denoising network with a surrogate gradient that drops its Jacobian, which leaves the attacked signal unchanged while making each iteration substantially cheaper (100 iterations).

Mist (Liang & Wu, 2023) attacks both surfaces at once, combining the encoder term of PhotoGuard with the denoising term of AdvDM,

$$\min_{\|\delta\|_\infty \le \varepsilon} \big\|E(x+\delta) - E(x_{\mathrm{tgt}})\big\|_2^2 - \lambda\, \mathbb{E}_{t,\,\epsilon}\big\|\epsilon_\theta(z_t, c^\delta, y, t) - \epsilon\big\|_2^2, \tag{3}$$

with the weight $\lambda = 10^4$ of its reference implementation and 100 iterations.

I2VGuard (Gui et al., 2025) sums a spatial, a contrastive and a temporal term,

$$\min_{\|\delta\|_\infty \le \varepsilon} \alpha \big\| E(x+\delta) \big\|_2^2 + \beta\, \mathcal{L}_{\text{con}} + \gamma\, \mathcal{L}_{\text{att}}. \tag{4}$$

The spatial term drives the conditioning latent toward a degraded target, and the temporal term $\mathcal{L}_{\text{att}}$ pushes the temporal attention maps of the perturbed input away from those of the clean one, read at the tapped block 20. The contrastive term acts on the clean-latent prediction $\hat{z}_0^\delta$ obtained under the perturbed condition, pulling it toward a target latent z_{tg} while pushing it away from the clean latent z_0 up to a margin τ_1,

$$\mathcal{L}_{\text{con}} = \big\| \hat{z}_0^\delta - z_{\text{tg}} \big\|_2^2 + \max\big(0,\ \tau_1 - \big\| \hat{z}_0^\delta - z_0 \big\|_2^2\big). \tag{5}$$

The three terms are equally weighted, $\alpha = \beta = \gamma = 1$, the two margins are $\tau_1 = \tau_2 = 0.5$, and the attack runs for 50 iterations.

Anti-I2V (Vu et al., 2026) adds two internal-representation terms to the denoising error, with equal weights (200 iterations):

$$\max_{\|\delta\|_\infty \le \varepsilon} \mathbb{E}_{t,\epsilon} \big\| \epsilon_\theta(z_t, c^\delta, y, t) - \epsilon \big\|_2^2 - \big\| h_{\text{deep}}(c^\delta) - h_{\text{shallow}}(c^\delta) \big\|_2^2 - \big\| E(x+\delta) \big\|_2^2, \tag{6}$$

where h_{shallow} and h_{deep} are the hidden features read at the tapped blocks 2 and 20. The second term is the internal representation collapse, which makes the deep blocks reproduce the shallow ones, and the third is the internal representation anchor, which pulls the conditioning latent toward the origin, the encoding of a black target image. We keep the three attack terms at their original equal weighting and omit the CLIP and LPIPS auxiliary terms of the reference implementation, since the ℓ_∞ projection applied to every method already bounds the distortion those terms regularize.

$\varepsilon = 16/255$ serves only as the default budget for the main comparison and does not limit our method. We additionally sweep $\varepsilon \in \{2, 4, 8, 16, 32\}/255$ for our method and Anti-I2V to study the protection–imperceptibility trade-off, with results reported in the budget-sweep analysis.

E.4 Metrics

Protection effectiveness is measured by how much the generated video degrades relative to the clean generation. All metrics are computed per video and averaged over samples.

Conventional metrics. DINO (Oquab et al., 2024) is the mean cosine similarity between DINO features of the generated frames and those of the input image. IQ follows the imaging quality metric of VBench (Huang et al., 2024), a no-reference per-frame assessment of low-level distortion such as blur, noise and over-exposure, averaged over frames and reported on a 0–1 scale; unlike DINO it does not consult the input image, so it registers that the frames have degraded rather than that they have moved away from the source. SC follows the subject consistency metric of VBench (Huang et al., 2024), computed as the mean DINO similarity of each frame to the first frame and to neighboring frames. Lower values of all three indicate stronger protection. For perturbation imperceptibility, we report PSNR, SSIM, and LPIPS of the protected image against the original.

LMM-as-a-judge metrics. VQA^2 (Jia et al., 2025) is a user-generated-content video quality assessor that emits logits over five quality levels, from bad to excellent, and converts them into a continuous score by probability-weighted averaging; it is prompt-free and rates perceptual quality from the video alone. FineVQ (Duan et al., 2025) is an InternVL-based video quality VLM that rates overall, blur, color, noise, artifact, and temporal quality on a 0–100 scale; we report its overall score. Both models judge the generated videos directly, without reference to the input image, complementing the reference-based conventional metrics.

E.5 Implementation Details

All methods optimize in the pixel domain and project the perturbation onto the ℓ_∞ ball of radius ε ($16/255$ in the main comparison) after every update. We use sign-PGD with step size $\varepsilon/16$ and $N = 100$ iterations, initialized uniformly at random inside the ball, and maximize the prediction divergence

$$\mathcal{A}(\delta) = \big\| \epsilon_\theta(z_{\tau^\star}, c^\delta, y_{\text{fix}}, \tau^\star) - \epsilon_\theta(z_{\tau^\star}, c^0, y_{\text{fix}}, \tau^\star) \big\|_2^2, \tag{7}$$

where $c^\delta = E(x+\delta)$ and $c^0 = E(x)$ are the conditioning latents of the perturbed and the clean image. On the flow-matching backbones ϵ_θ outputs a velocity, so we also refer to $\mathcal{A}$ as the velocity divergence. Since $z_{\tau^\star}$ is the sampler's initialization, each iteration consists of one forward and one backward pass through ϵ_θ. The baselines retain their original objectives and iteration counts (App. E.3) and are likewise projected onto the same ε-ball. The generation configuration of each model (Table 6) is identical for clean and protected inputs.

E.6 Auxiliary Objectives

The ablation of Sec. 5.4 holds everything about I2VShield fixed—the attacked timestep $\tau^\star$, sign-PGD with step size $\varepsilon/16$ and $N = 100$ iterations, the budget $\varepsilon = 16/255$, the random start, and the generation configuration—and changes only the objective. Each variant maximizes

$$\mathcal{A}(\delta) + \lambda\, \mathcal{A}_{\text{aux}}(\delta), \tag{8}$$

where $\mathcal{A}$ is the prediction divergence of Eq. (7) and $\mathcal{A}_{\text{aux}}$ is one of the four terms below. We use $\lambda = 1$ for all of them, the weight each term carries in the method it is taken from: the Anti-I2V terms are combined without explicit coefficients, and Mist's texture term has unit weight, since its `mist_rate` $= 10^4$ multiplies the semantic term that this variant does not contain.

Semantic. The denoising term of Anti-I2V, which is also the quantity AdvDM maximizes, pinned to $\tau^\star$ rather than sampled over t:

$$\mathcal{A}_{\text{sem}}(\delta) = \left\|\epsilon_\theta\left(z_{\tau^\star}, c^\delta, y_{\text{fix}}, \tau^\star\right) - \epsilon\right\|_2^2. \tag{9}$$

Internal representation collapse. The IRC term of Anti-I2V, which drives the deep blocks to reproduce the shallow ones. On the 40 blocks of Wan2.1-I2V-14B we tap block 2 as shallow and block 38 as deep:

$$\mathcal{A}_{\text{irc}}(\delta) = -\left\|h_{38}(c^\delta) - h_2(c^\delta)\right\|_2^2. \tag{10}$$

Internal representation anchor. The VAE-side IRA term of Anti-I2V, which pulls the conditioning latent toward the origin, the encoding of a black target image:

$$\mathcal{A}_{\text{ira}}(\delta) = -\left\|E(x+\delta)\right\|_2^2. \tag{11}$$

Texture. The textural term of Mist, which pulls the conditioning latent toward that of a fixed target image:

$$\mathcal{A}_{\text{tex}}(\delta) = -\left\|E(x+\delta) - E(x_{\text{tgt}})\right\|_2^2. \tag{12}$$

The minus signs place all four terms in the same maximization as $\mathcal{A}$: the last three are distances that their own methods minimize, so maximizing their negation reproduces the original signal.

F Additional Experiments on Generalization and Robustness

This section continues Sec. 5.3, which sweeps the perturbation budget and tests unseen prompts, with the generalization and robustness results that do not fit within the page budget of the main paper. Generalization is widened along two axes the main paper cannot cover: across model families and checkpoints (App. F.1), and across backbones, by transferring each perturbation to models it was not optimized against (App. F.4). Robustness is tested against two operations that may be applied to a protected image: purifying it before generation (App. F.2) and changing the generation settings (App. F.3). Unless stated otherwise, every experiment uses Wan2.1-I2V-14B on HQFace-I2V with P_1 and the budget $\varepsilon = 16/255$.

F.1 More Attacked Models

The main comparison of Sec. 5.2 establishes the protection on three backbones, one per architecture family, against six baselines and on two datasets. We extend the same evaluation to the six further checkpoints of Tab. 5, so that the claim rests on the breadth of publicly available I2V releases rather than on one representative per family. Together they span the four families, successive releases

within them, a range of scale from 1.3B to 27B parameters, and a publicly available NSFW fine-tune that is the closest proxy we have to a deployed system. Nothing about the method is adapted to them: the objective of Eq. (7), the budget $\varepsilon = 16/255$, the optimizer, the step size and the 100 iterations are those of the main comparison, and each model is attacked on 20 HQFace-I2V images with P_1 under its own default generation configuration. Table 7 reports the result. The protection carries to every checkpoint: DINO falls by 0.41 to 0.86, from unprotected values that sit near the ceiling. The largest drop is on Wan2.2-TI2V-5B, the newest Wan release, which goes from 0.977 to 0.116 and is therefore protected as strongly as the backbones of the main comparison. The NSFW fine-tune of Wan2.2-I2V-A14B falls from 0.909 to 0.461 even though it is distilled for few-step sampling, a schedule much shorter than the 25 steps the method is optimized against. Since no hyperparameter is tuned per model, the single-timestep objective transfers across families, across successive releases of the same family, and across an order of magnitude of scale.

Table 7: Protection on six further I2V checkpoints (20 HQFace-I2V images, prompt P_1, $\varepsilon = 16/255$). Lower is stronger. The objective, budget, optimizer and iteration count are those of the main comparison, and no hyperparameter is tuned per model.

	DINO	
Model	Clean	Ours
Wan2.2-TI2V-5B (Wan et al., 2025)	0.977	0.116
Wan2.2-I2V-A14B, NSFW	0.909	0.461
CogVideoX-5B-I2V (Yang et al., 2025)	0.967	0.405
LTX-Video 0.9.5 (HaCohen et al., 2024)	0.976	0.564
LTX-Video 0.9.7 (HaCohen et al., 2024)	0.979	0.434
SkyReels-V2-1.3B (Chen et al., 2025)	0.935	0.392

F.2 Robustness to Purification

We sweep five purifications at five strengths each: the generic operations JPEG compression, Gaussian blur and additive Gaussian noise, and the published purifiers GrIDPure (Zhao et al., 2024) and Impress (Cao et al., 2023), which run an independent Stable Diffusion v1.5. GrIDPure is the grid-wise, small-step refinement of DiffPure (Nie et al., 2022) introduced to preserve image structure while purifying, and we evaluate it in place of DiffPure itself. Each is applied to the already-optimized images, which are then regenerated, so no new attack is run and the perturbation is exactly the one evaluated in Tab. 1. Anti-I2V, the strongest baseline there, is swept alongside on the same images. All runs use 20 HQFace-I2V images on Wan2.1-I2V-14B.

Figure 8 reports the sweep, across which the protection falls into three regimes. Over eleven of the twenty-five settings it is left fully intact, staying within 0.11 of its un-purified value of 0.184: $q = 90$ JPEG (0.296), blur up to $\sigma = 0.5$ (0.193), noise up to scale 0.02 (0.213), the weakest GrIDPure setting (0.182), and the entire range of Impress, which never removes the perturbation at any strength (0.200 to 0.297). Three settings reduce it without closing the gap: noise at scale 0.03 (0.358), GrIDPure at 0.5 (0.556) and JPEG at $q = 75$ (0.703). The remaining eleven—JPEG from $q = 60$, blur from $\sigma = 1.0$, noise from scale 0.05, and GrIDPure from 0.3 onward—restore the similarity to between 0.81 and 0.90, where the two curves meet; these are also the settings that visibly degrade the image being purified. Throughout the sweep the protected generation stays below the unprotected level of 0.947.

Against Anti-I2V the margin is 0.54 to 0.65 over the first regime and 0.15 to 0.49 over the second, in the latter case still no smaller than the gap separating Anti-I2V itself from the unprotected generation, before the two converge in the third, where Anti-I2V occupies 0.84 to 0.91. Its curve is flat across all twenty-five settings, but the two are not comparable in shape: each method's response is bounded by the protection it starts from, and Anti-I2V moves within a narrower range between its initial 0.82 and the clean 0.947, whereas I2VShield starts at 0.184. Stability across the sweep is therefore not, on its own, a measure of purification robustness.

Sensitivity to strong purification is shared across protections of this kind. All of them, ours included, are small-norm, high-frequency signals in pixel space, which is what a purifier is built to remove, and GrIDPure and Impress were designed against protective perturbations in general rather than against any particular objective. The constraint applies to I2VShield no differently than to the methods it is

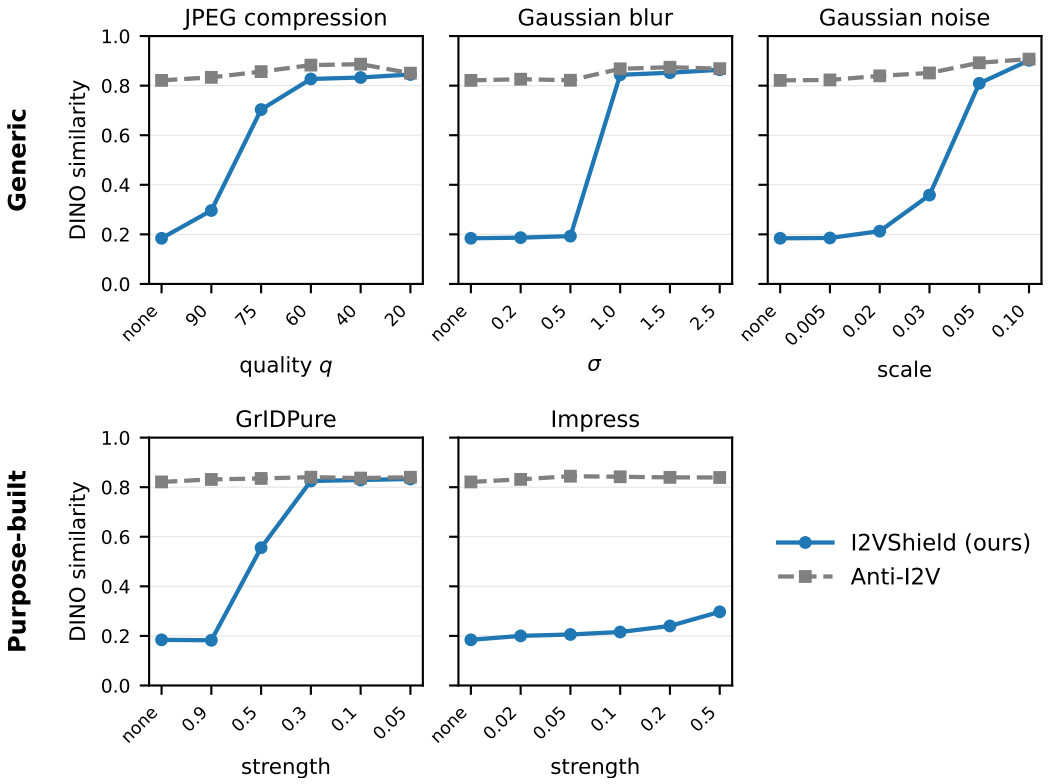

Figure 8: Protection after purification, applied to the already-protected images (20 images, Wan2.1-I2V-14B, HQFace-I2V). The top row holds the three generic image operations, the bottom row the two purifiers built to strip protective perturbations. Within each panel the settings run from the weakest to the strongest purification, with the leftmost point the un-purified image. Lower is stronger protection.

compared with: we add no new objective and no new attack surface, and change only where along the timestep axis the budget is spent. Wherever the perturbation survives, that single change is what separates the two curves—over fourteen of the twenty-five settings I2VShield remains well ahead of the strongest prior protection, and over the remaining eleven the two are comparable. Retaining the full strength of a pixel-space protection under the most aggressive of these operations remains open, and we do not address it here.

F.3 Generalization to Inference and Optimization Settings

The perturbation is optimized against a single generation configuration, so we vary one parameter at a time around the default of Tab. 6 and report the DINO similarity of the protected generation, the same quantity as in Tab. 1. The four generation parameters are the number of inference steps, the guidance scale, the flow-matching shift and the number of generated frames; we additionally vary the attack iteration count N of Alg. 1, a property of the optimization rather than of generation, to check that our default is not over-tuned. All runs use 20 HQFace-I2V images on Wan2.1-I2V-14B.

Table 8 collects the result: the protection holds under every generation setting we vary, staying between 0.045 and 0.330 against an unprotected level of 0.87 to 0.96. What these parameters change is only where in that band it lands, and two of them move it in the direction Diagnosis 2 predicts.

Raising the guidance scale drives the similarity monotonically down, $0.237 \rightarrow 0.045$ from 3 to 10, as expected when a larger weight on the conditional branch amplifies a perturbation that acts through the condition. Lowering the shift does the same, $0.228 \rightarrow 0.097$ from 9 to 3, and this one is a direct

Table 8: Protection under generation settings that differ from the one the perturbation was optimized against (20 images, Wan2.1-I2V-14B, HQFace-I2V). Entries are the DINO similarity of the protected generation; lower is stronger. The unprotected level ranges from 0.87 to 0.96 over these settings, reaching its minimum at guidance scale 10. Defaults are underlined; every row varies one parameter around the same default configuration, whose value is measured once and repeated. The last row varies the attack budget rather than a generation setting.

Parameter	Setting / DINO			
Inference steps	15 / 0.109	25 / 0.184	50 / 0.239	100 / 0.330
Guidance scale	3 / 0.237	5 / 0.184	7 / 0.084	10 / 0.045
Shift	3 / 0.097	5 / 0.184	7 / 0.206	9 / 0.228
Frames	9 / 0.133	17 / 0.184	25 / 0.152	
Attack iterations N	25 / 0.486	50 / 0.281	75 / 0.216	100 / 0.184

test of Diagnosis 2: a smaller shift places more denoising steps at high noise, where the condition's leverage is concentrated. The frame count barely matters, at 0.133 to 0.184 across 9 to 25 frames, so the protection carries across video lengths it was not optimized for.

The sampling grid is the second such test, and it moves the protection the other way for the same reason: a finer grid shortens the first sampling interval, so the highest-noise step—the one the budget is committed to—governs a smaller share of the trajectory. At 100 steps, four times finer than the 25-step trajectory the perturbation was optimized on and twice the largest official default among our backbones, the similarity is 0.330, a drop of 0.617 below the unprotected 0.947 against 0.763 at the default grid, and still far below what any baseline of Tab. 1 reaches at its own default configuration. At this level the generated video is destroyed rather than merely degraded: 0.330 lies far outside the 0.75–0.90 band within which every baseline leaves the subject recognizable and the motion coherent (Fig. 1), and below the 0.426 that already constitutes effective protection at $\varepsilon = 2/255$ in Sec. 5.3. Protection therefore tracks how much of the sampling trajectory sits at high noise, which is Diagnosis 2 read through a generation parameter rather than through the attack.

Finally, the attack iteration count behaves as an optimization budget should, falling from 0.486 at $N = 25$ to 0.184 at $N = 100$. At $N = 25$ it already outperforms every baseline, so $N = 100$ is a lower bound on what the objective can reach rather than over-optimization.

F.4 Cross-Model Transferability

To evaluate how far the perturbation transfers between backbones, we feed each perturbation to the two models it was not optimized against, which generate with their own official configuration, images being resized rather than cropped when resolutions differ. Anti-I2V is transferred across the same pairs on the same images, so the two are measured under identical conditions, and every entry is compared against the clean generation of the column model. The six transfers require no new attack, and all runs use 20 HQFace-I2V images.

Table 9 reports them. Averaged over the six, I2VShield lowers DINO by 0.103 below the clean level, against 0.077 for Anti-I2V, and it is the stronger of the two in four pairs: transferred out of LTX-Video it reaches 0.823 on Wan2.1 and 0.801 on CogVideoX1.5, against 0.869 and 0.862, and transferred out of CogVideoX1.5 it reaches 0.896 and 0.919, against 0.929 and 0.940. In the two transfers out of Wan2.1 the methods are level, at 0.930 against 0.928 on LTX-Video and 0.887 against 0.884 on CogVideoX1.5. Neither retains much of its protection once the backbone is substituted: the drop below the clean level ranges from 0.068 to 0.162 for I2VShield and from 0.047 to 0.107 for Anti-I2V, against clean values of 0.963 to 0.998.

Transfer across architectures is a long-standing difficulty for adversarial perturbations in general: a perturbation optimized against one network's parameters has no reason to align with another's, and the difficulty grows with the architectural distance between them. The three backbones here differ in prediction target, scheduler, and the mechanism by which the image enters the network, so the distance between any two of them is considerable. It constrains every method in the comparison alike: Anti-I2V spreads its budget over the whole timestep axis rather than concentrating it, and transfers no better. What the present work settles is the question that is tractable once a backbone

Table 9: Cross-model transfer of the perturbation (20 images, HQFace-I2V, $\varepsilon = 16/255$). A perturbation optimized against the row model is fed to the column model, which generates with its own official configuration. Lower is stronger.

Method	Optimized against	Generated by		
		Wan2.1	LTX-Video	CogVideoX1.5
Clean (no protection)	—	0.976	0.998	0.963
I2VShield (ours)	Wan2.1	—	0.930	0.887
	LTX-Video	0.823	—	0.801
	CogVideoX1.5	0.896	0.919	—
Anti-I2V (Vu et al., 2026)	Wan2.1	—	0.928	0.884
	LTX-Video	0.869	—	0.862
	CogVideoX1.5	0.929	0.940	—

is fixed, and there a single relocation of the budget along the timestep axis, with nothing added to the objective and no new attack surface, is enough to take DINO from 0.92 to 0.18. Narrowing the transfer gap is a broader question, and we leave it to future work.

G Ethics Statement

This work develops a defense against the non-consensual animation of real people's photographs, and its central object—an adversarial perturbation—is dual-use by construction: the same optimization that protects an image degrades generation from it, and could in principle be applied to images whose animation is legitimate. We consider the balance favourable, since the perturbation acts only on images its own owner chooses to protect and leaves every other input untouched, but we state the property rather than leave it implicit.

The benchmark we release requires more care, as it consists of photographs of real, identifiable people. All 80 images are offered by their photographers under the Pexels license (Pexels, 2026), which permits free use and modification, and no image was obtained by scraping or used outside the terms of that license. No personal information beyond the photograph itself is collected or distributed, and no human subjects were recruited. HQFace-I2V is released solely for research on defending against malicious image-to-video synthesis; as stated in App. B, it must not be used to train generative models whose purpose is to synthesize video of real, identifiable individuals. The videos generated in the course of our experiments are degraded outputs produced to measure protection strength, and we release only those needed to document the benchmark.

One of the nine models requires separate mention. Among the checkpoints of App. E.1 is a publicly available NSFW fine-tune of Wan2.2-I2V-A14B, released by a third party on Hugging Face and listed in Tab. 5. We include it because the abuse that motivates this work is carried out with models of exactly this kind, and a protection evaluated only on general-purpose checkpoints would leave open whether it holds where it matters most. Our use of it is confined to that evaluation. We do not fine-tune, modify or redistribute the model; it is driven only by the benign animation prompts of HQFace-I2V, never by prompts requesting explicit content; its outputs serve solely to compute the similarity scores of Tab. 7; and no video generated by it appears in the paper or in any material we release. The checkpoint is identified in Tab. 5 for reproducibility, not as an endorsement.

H Reproducibility Statement

Every component needed to reproduce our results is specified in the paper. The nine I2V checkpoints are public releases used without any fine-tuning of our own, listed with their exact Hugging Face identifiers in Tab. 5; the generation configuration of each backbone, identical for clean and protected inputs and run under a fixed seed, is given in Tab. 6. The method is stated as Alg. 1, its objective as Eq. (7), and its hyperparameters—pixel-space sign-PGD, step size $\varepsilon/16$, $N = 100$ iterations, random start, $\tau^\star = 1$—in App. E.5. The six baselines are reproduced from their original objectives and iteration counts, written out in App. E.3 together with the two deviations we make and our reasons for them. The five evaluation metrics are defined in App. E.4, and the diagnostic experiments

of Sec. 4.2 are specified in App. C, including the settings particular to each. Of the two datasets, CelebV-Test is drawn from the public CelebV-Text (Yu et al., 2023), and HQFace-I2V is released with this paper, with its curation procedure, prompts and file formats documented in App. B. We release the code for the method, for the baselines as we reproduce them, and for the evaluation pipeline, so that every table and figure in the paper can be regenerated from the public checkpoints and the released data.

I USE OF AI TOOLS

We used large language models at several stages of this work. For the research itself, they assisted in searching for and discussing related work, in refining the framing and interpretation of experimental results, and in writing and debugging parts of the code used to run the experiments. For the manuscript, they were used to translate text into English, to draft several appendix sections, to produce the plots from the measurements behind them, and to revise wording throughout. A vision-language model was additionally used to construct HQFace-I2V, filtering the candidate photographs and generating the animation prompts, as described in App. B; the reference videos released with the benchmark were produced by a commercial I2V service.

No experimental result reported in this paper was produced by a language model. The method, the diagnostic experiments, and every number in the tables and figures come from runs the authors executed on the models and data described in App. E. The authors verified the AI-assisted contributions: code was checked by inspection and by execution, each reported value was traced back to the output of the run that produced it, cited work was read in the original, and AI-assisted text was checked for accuracy and for originality. The authors take full responsibility for the content of the paper.

J COST ANALYSIS

We report the computational cost of each protection. Every method is run on the same GPU, one NVIDIA RTX PRO 6000 (96 GB) of a four-GPU node with 22 vCPUs of an Intel Xeon Platinum 8470Q, on the same images and against Wan2.1-I2V-14B under the configuration of Tab. 6, so the numbers differ only in the optimization. We report the time to protect one image and the peak GPU memory over protection and generation, averaged over the 80 HQFace-I2V images, at each method's own prescribed iteration count (App. E.3). Generation is excluded from the time, and takes 62.7 s per video at a peak of 44.7 GiB under the same configuration.

Table 10: Cost of protecting a single image on Wan2.1-I2V-14B, measured on one RTX PRO 6000 (96 GB). Iteration counts are those each method prescribes; the baselines whose objectives need a sampling trajectory read it from the shared precomputed bank of App. E.3.

Method	Iterations	Time (s)	s / iteration	Peak memory (GiB)
PhotoGuard (Salman et al., 2023)	100	63.6	0.64	44.7
AdvDM (Liang et al., 2023)	100	944.9	9.45	84.8
SDS+ (Xue et al., 2024)	100	941.5	9.42	84.8
Mist (Liang & Wu, 2023)	100	942.5	9.43	84.8
I2VGuard (Gui et al., 2025)	50	491.7	9.83	84.8
Anti-I2V (Vu et al., 2026)	200	1852.4	9.26	84.9
I2VShield (ours)	100	969.4	9.69	84.8

Table 10 reports the result. PhotoGuard runs at 0.64 s per iteration, since its objective never leaves the encoder, and its peak memory of 44.7 GiB is set by generation rather than by the attack. Every method that reaches the denoising network costs 9.3 to 9.8 s per iteration and peaks at 84.8 to 84.9 GiB, I2VShield among them at 9.69 s and 84.8 GiB: each iteration of these methods is one forward and one backward pass, the baselines reading their noisy latents from the shared trajectory bank (App. E.3). The per-image totals therefore follow the iteration count each method prescribes, and I2VShield, at 100 iterations, has a cost comparable to that of AdvDM, SDS+ and Mist.

K More Visualization Results

K.1 Comparison with Existing Methods

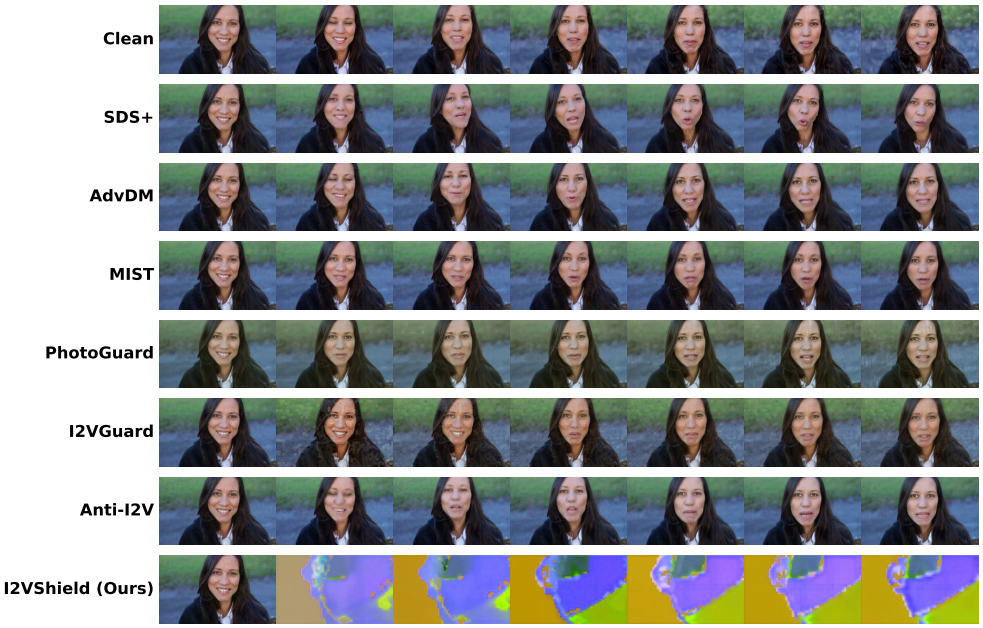

Figure 9: Visualization of the main comparison.

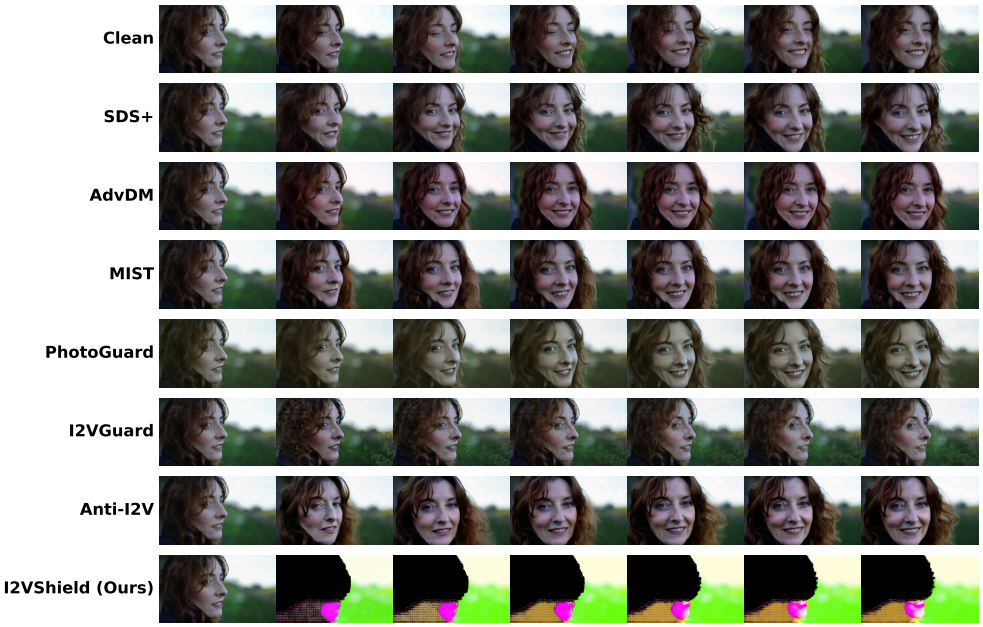

Figure 10: Visualization of the main comparison.

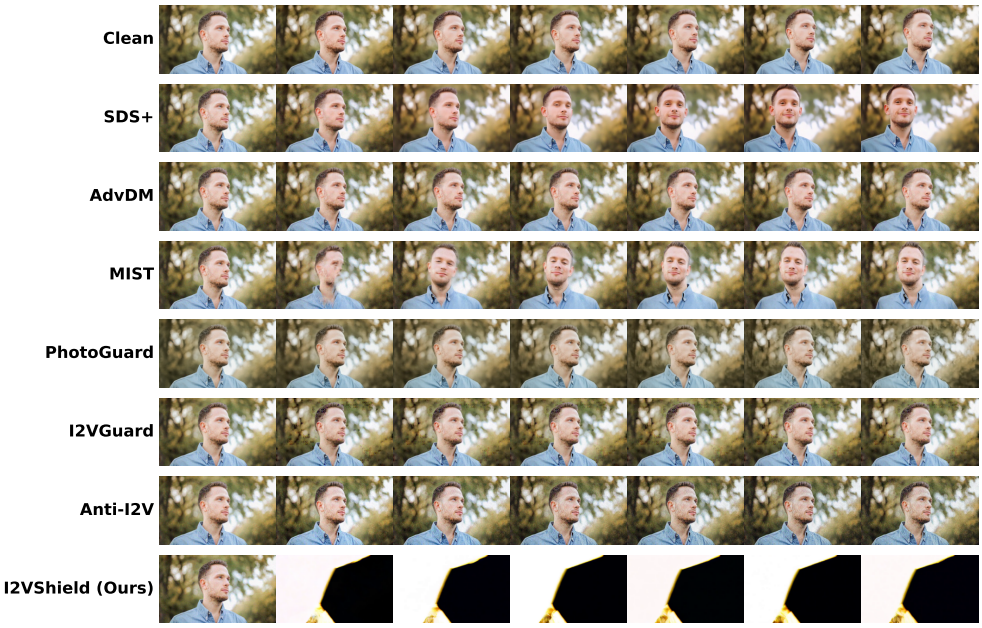

Figure 11: Visualization of the main comparison.

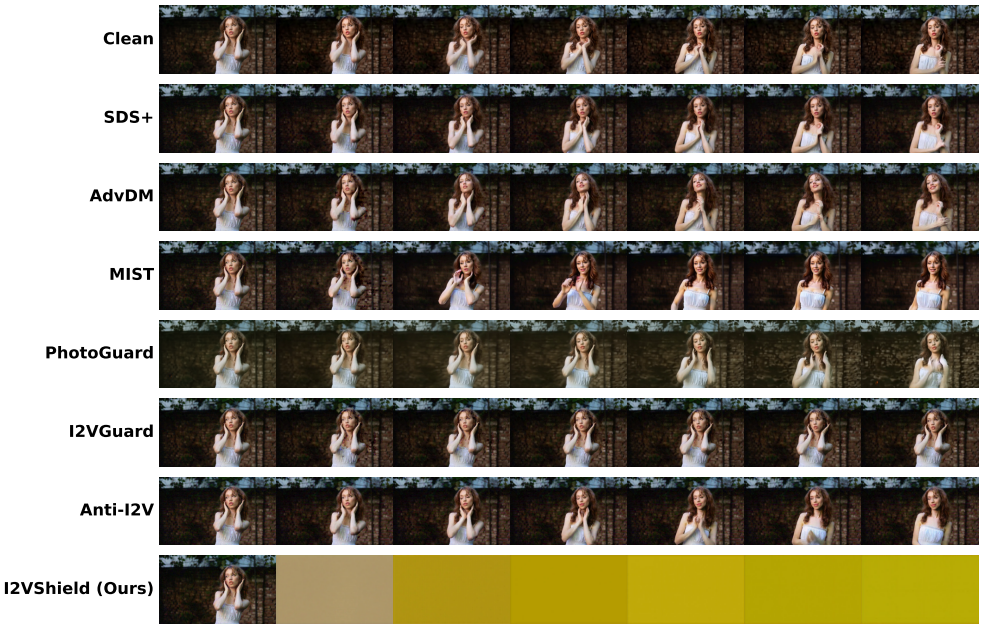

Figure 12: Visualization of the main comparison.

K.2 Perturbation Budgets

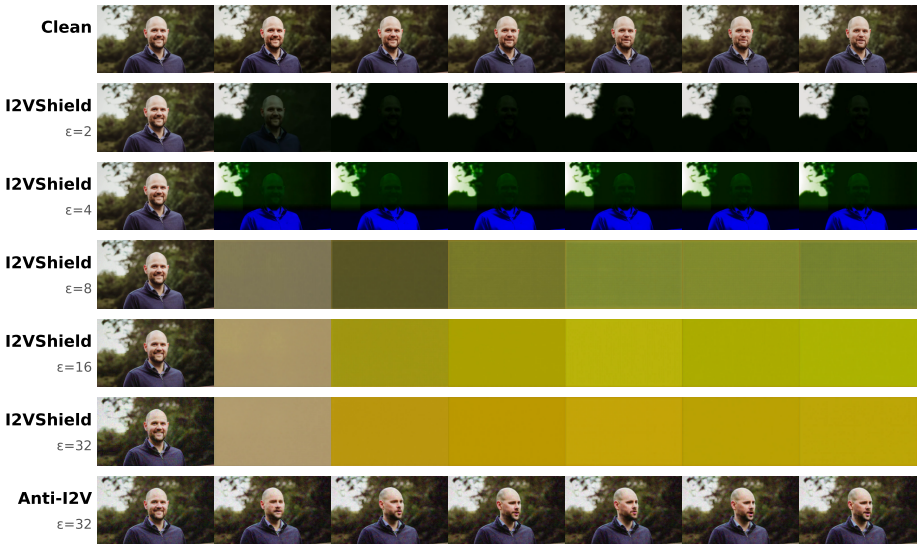

Figure 13: Visualization of the perturbation budget experiment.

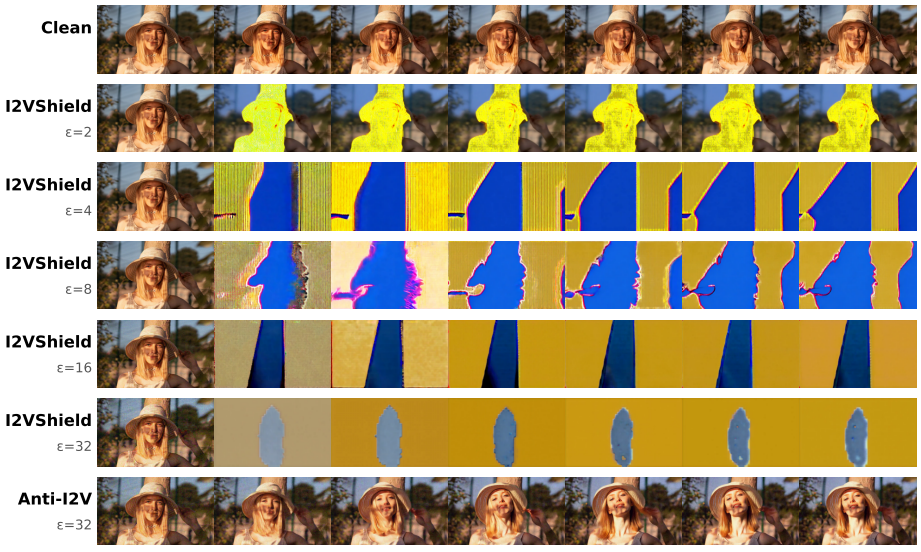

Figure 14: Visualization of the perturbation budget experiment.

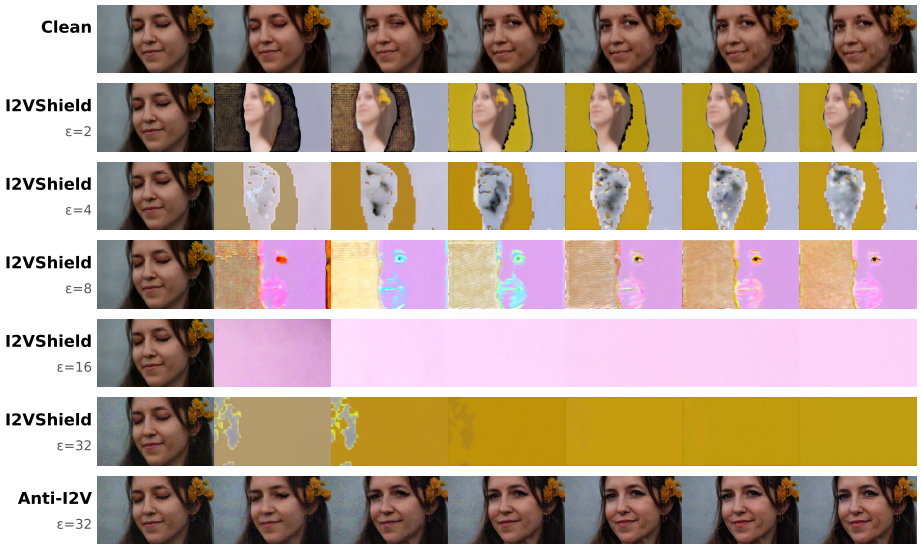

Figure 15: Visualization of the perturbation budget experiment.

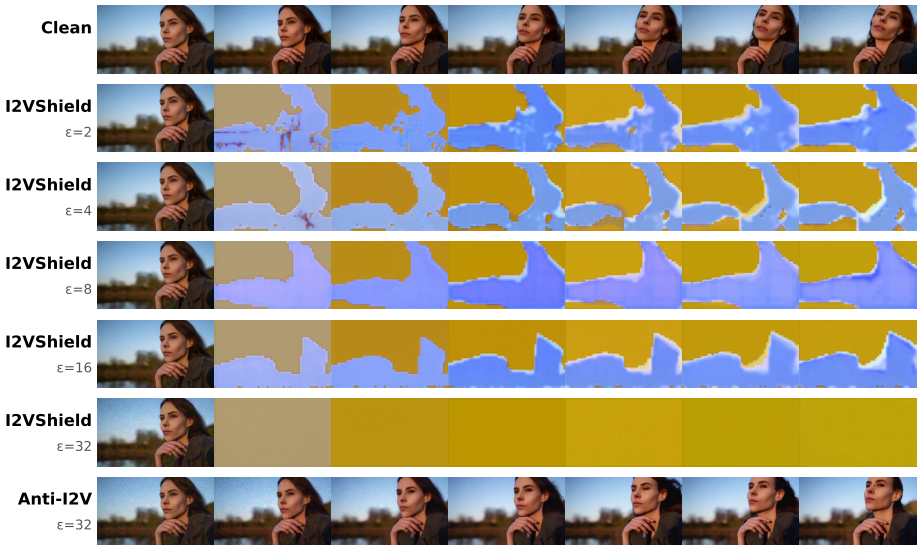

Figure 16: Visualization of the perturbation budget experiment.

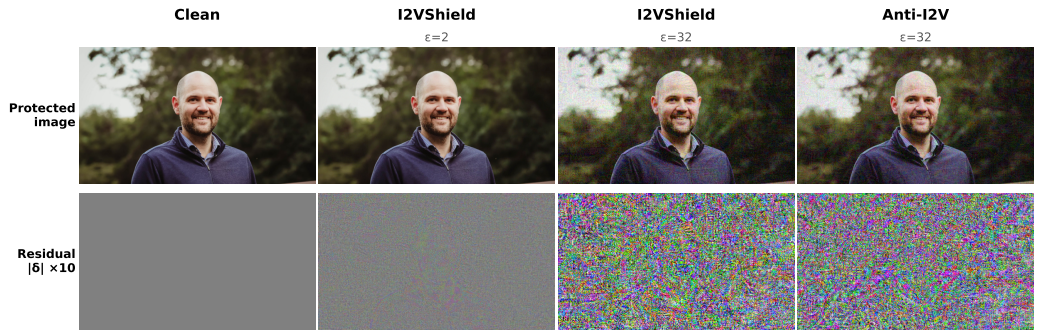

Figure 17: Visualization of the perturbation.

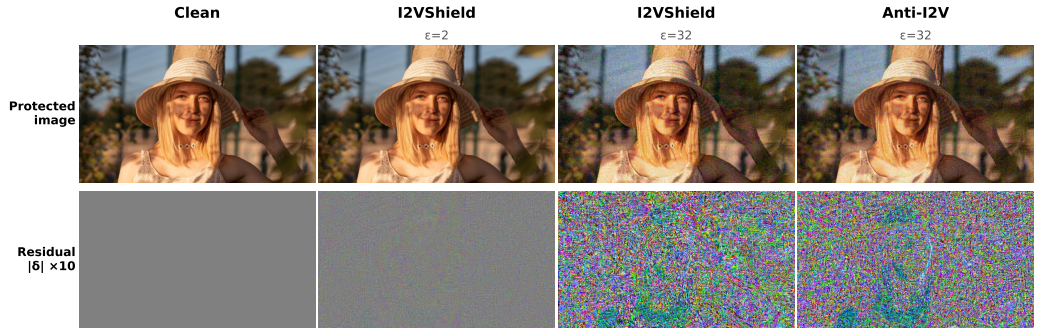

Figure 18: Visualization of the perturbation.

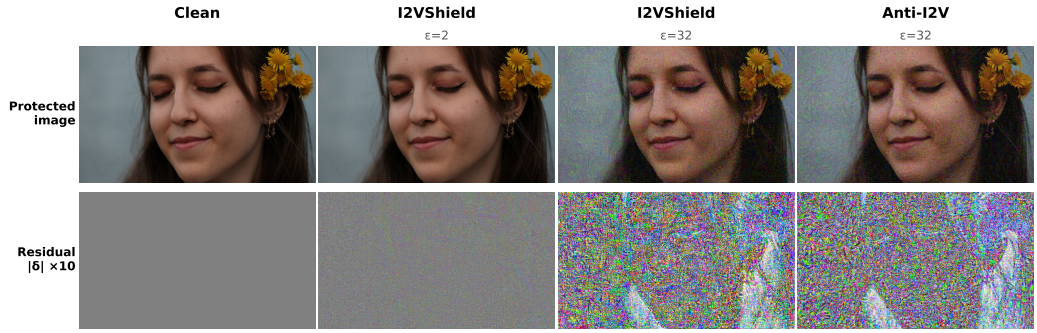

Figure 19: Visualization of the perturbation.

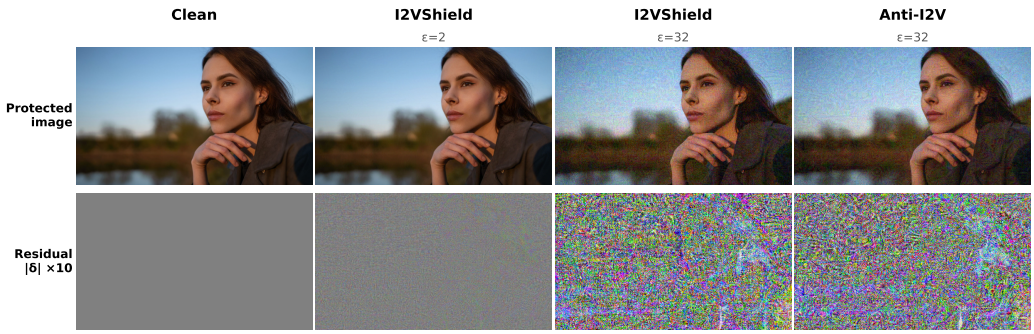

Figure 20: Visualization of the perturbation.