I2V-DETACH: Source Grounding Detachment for Unauthorized Image-to-Video Generation

Chanhui Lee[1] **Yeonghwan Song**[2] **Yewon Kang**[1] **Jeany Son**[1]
[1]POSTECH [2]GIST
{chanlee, yewon.kang, jeany}@postech.ac.kr yeonghwan.song@gm.gist.ac.kr

https://chanhuilee1111.github.io/I2V-DETACH/



Figure 1: Goal of I2V-DETACH against diffusion-based I2V generation. Unlike prior baselines that often result in temporal degradation or fidelity reduction, *source grounding detachment* objective directly drives video generation that no longer faithfully preserves visual content of source-image.

Abstract

Recent diffusion models have advanced image-conditioned video generation, but their accessibility raises risks of unauthorized manipulation. Image immunization has emerged as a promising defense for Image-to-Video (I2V) protection, yet existing methods often rely on degrading visual quality or temporal coherence. In this work, we revisit I2V immunization as reducing source-content preservation, shifting the focus from output degradation to weakening source-conditioned content propagation. To this end, we propose I2V-DETACH, an image immunization method that prevents unauthorized I2V generation through source grounding detachment. I2V models preserve source content by propagating source-image cues into video latents during denoising. Accordingly, I2V-DETACH disrupts this process in two complementary ways: it suppresses source-video grounding to reduce source cue propagation, and strengthens intra-video interactions to shift denoising toward video-side dynamics with reduced reliance on the source image. To obtain informative optimization signals, we initialize auxiliary video latents with source-oriented states that remain within the source-conditioned generation regime while deviating from source-preserving trajectories. Extensive experiments across diverse diffusion architectures show that I2V-DETACH substantially reduces source-content preservation, improves over prior immunization methods, and exhibits strong black-box transferability.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

1 Introduction

Diffusion models [21, 13] have enabled high-fidelity visual content manipulation under natural-language guidance, spanning image synthesis [41, 16, 29, 28] and video generation [22, 52, 27, 46, 57, 20]. In particular, image-conditioned video generation allows users to preserve and animate the visually grounded content of a reference image, but also increases the risk of unauthorized image use, including portrait misuse, privacy violations, and malicious visual manipulation [36, 49, 37]. Protecting private images from such misuse has therefore become an urgent problem.

Image immunization provides a proactive defense by adding imperceptible perturbations to source images before they are used by generative models. Early studies mainly focused on image manipulation tasks such as image editing [42, 33, 26, 3, 44], inpainting [40, 10, 25], and instruction-guided image editing [9, 47]. I2VGuard [17] contrastively aligns intermediate representations with a target video, VidFreeze [12] suppresses motion synthesis by manipulating attention dynamics, and Anti-I2V [45] enhances robustness against purification by optimizing perturbations jointly in the CIE Lab color space and the frequency domain.

While effective, existing defenses often overlook a key property of I2V generation: videos are generated under strong reference conditioning, where the source image serves as a persistent and dominant reference throughout denoising [54, 7, 51]. As a result, prior defenses often manifest as perceptual degradation, temporal artifacts, or unstable generation, while the generated video can still retain recognizable source identity, structure, and appearance, as shown in Figure 1. We therefore shift the goal of I2V immunization from degrading generated videos to reducing the preservation of protected source content. Importantly, this does not require eliminating all source-related cues; instead, the goal is to reduce source preservation rather than treating generation degradation itself as the primary objective.

To this end, we propose I2V-DETACH, a novel image immunization method for diffusion-based I2V models based on *source grounding detachment*. I2V-DETACH directly targets the source-video interactions that act as the main pathway for propagating source-specific cues into video latents during denoising. Specifically, it learns an imperceptible perturbation that weakens the grounding between the protected source image and source-conditioned video latents, thereby reducing the generated video's dependence on the protected source.

I2V-DETACH optimizes two complementary objectives for source grounding detachment. A ***source-video decoupling loss*** suppresses source-to-video interactions, while an ***intra-video coupling loss*** strengthens interactions among video latents, increasing the degree to which video latents evolve through video-side interactions rather than being anchored to the source image. To provide informative optimization signals, we further introduce source-divergent video initialization, where auxiliary video latents are drawn from the source-conditioned generation process rather than random initialization. These latents remain source-related while deviating from source-preserving trajectories, enabling effective detachment without merely destabilizing generation. As a result, I2V-DETACH makes generated videos less strongly grounded in the protected source, as shown in Figure 1.

We perform extensive experiments to compare our method with existing I2I immunization methods [9, 25], as well as I2V immunization methods [45, 17] whose objectives are closely related to ours, across diffusion-based I2V models with diverse architectures [52, 20, 57]. We further extend our method to the universal adversarial perturbation [32, 31, 30] setting and demonstrate that our objective is better suited to learning universal perturbations than those of existing baselines. We conduct comprehensive evaluations from multiple perspectives, including quantitative metrics and human evaluation, to validate the effectiveness of the proposed method. The results demonstrate the superiority of our method compared with existing I2V immunization baselines, which we attribute to explicitly tailoring source-image grounding disruption to the distinctive characteristics of I2V generation.

The main contributions can be summarized as follows:

- We reformulate I2V immunization as reducing source-conditioned content preservation, showing that protected source content can remain recognizable even when output quality or temporal dynamics are degraded.
- We propose I2V-DETACH, a source-grounding detachment framework that directly intervenes in the source-to-video pathway of I2V denoising by separating source-video grounding from intra-video interactions.

- We introduce source-video decoupling and intra-video coupling objectives with source-oriented video latent initialization, enabling targeted suppression of source cue retrieval while promoting video-side dynamics.
- We demonstrate consistent source-preservation reduction across diverse I2V architectures, adapted I2I baselines, and I2V immunization methods, with strong robustness and black-box transferability.

2 Related Works

Image-to-Video Diffusion Models. Modern image-to-video (I2V) generation has largely evolved by augmenting text-to-image diffusion models with temporal modeling modules, enabling the synthesis of temporally coherent videos from a single image. Early methods such as Tune-A-Video [50] and AnimateDiff [19] showed that even lightweight temporal extensions can produce convincing image animation. Stable Video Diffusion [5] further advanced latent video diffusion at scale and has since become a widely adopted open-source benchmark. More recent models have improved generation quality and controllability through more powerful spatio-temporal architectures, including Lumiere [4] and diffusion transformer-based models such as CogVideoX [52]. At the same time, LTX-Video [20] emphasizes efficient generation, while Wan [46] lowers the barrier to high-quality I2V synthesis as a large-scale open-source model. Community-led projects such as VideoCrafter2 [8] and Open-Sora [57] have further democratized access to I2V technology. As I2V models become increasingly realistic and accessible, protecting source images from misuse in video synthesis has emerged as an important problem.

Image Immunization against Diffusion Models. Adversarial attack-based image immunization has emerged as an active research direction for protecting images against unauthorized use and malicious generative manipulation. In image-to-image pipelines, prior works have demonstrated strong protection performance by perturbing VAE encoder latents [42, 26, 3, 9], disrupting cross-attention mechanisms [33, 44], or exploiting denoising outputs [42, 10, 40] to break faithful image editing. With the rapid progress of image-to-video generation, recent attention has naturally shifted toward defending images from misuse in video synthesis. Among the first works to explicitly identify this problem, I2VGuard [17] induced temporal degradation by forcing encoder-level latents, attention outputs, and denoised latents to resemble those of a fixed target video. VidFreeze [12] primarily promotes temporal freezing by imposing temporally static attention patterns. Anti-I2V [45] extended this line of work by encouraging internal attention features to collapse into uninformative representations. While these approaches provide effective protection, generated videos can still preserve the identity and visual content of the source image due to the source-anchored prior of I2V models [54, 7, 51]. In contrast, our method explicitly separates *two attention pathways* to suppress source-video grounding and reinforce video-side interactions, directly breaking this prior.

3 Motivation: Source Grounding in I2V Diffusion Models

Diffusion-based I2V models generate videos by denoising video latents conditioned on a source image and a text prompt. Let Z_τ denote the noisy video latents at timestep τ, c the source-image conditioning states, and t the text embedding. At each denoising step, the model predicts

$$\hat{\epsilon}_\tau = \epsilon_\theta(Z_\tau, \tau, t, c). \tag{1}$$

Since source conditioning is repeatedly injected throughout the denoising process, the source image serves as a persistent visual reference in I2V generation. Consequently, generated videos can remain strongly grounded in the source image even when perturbations mainly induce temporal instability or perceptual degradation.

Although different I2V architectures implement source conditioning in different ways, source information is commonly propagated through attention interactions between video latent states and source-conditioning states. We denote the source-conditioning tokens by $\mathcal{C}$ and the video tokens by $\mathcal{V}$. For query and key features Q and K, we define the pre-softmax attention logit function S as

$$S(Q,K) = \frac{QK^\top}{\sqrt{d_k}}, \quad [S(Q,K)]_{I,J} = \left[S(Q,K)_{ij}\right]_{i\in I,\, j\in J}. \tag{2}$$

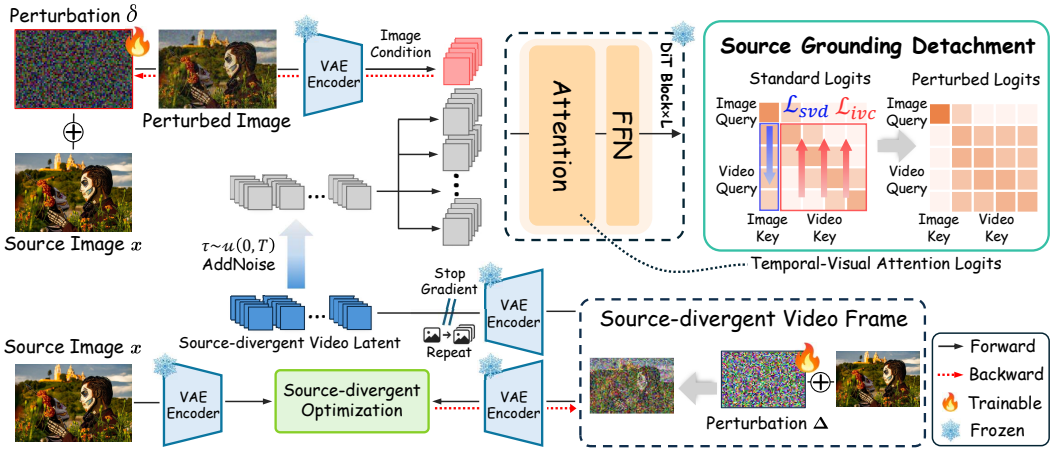

Figure 2: Illustration of the framework of I2V-DETACH. Source grounding detachment is achieved by two complementary losses; *source-video decoupling* and *intra-video coupling*. Jointly optimized with progressively source-divergent video frames, I2V-DETACH drives denoising video latents away from source-bound constraints, disrupting source-content preservation in generated videos.

Here, rows correspond to query tokens and columns correspond to key tokens. For index sets $\mathcal{I}$ and $\mathcal{J}$, $[S(Q,K)]_{\mathcal{I},\mathcal{J}}$ denotes the submatrix obtained by selecting query rows in $\mathcal{I}$ and key columns in $\mathcal{J}$. Under this notation, $[S(Q,K)]_{\mathcal{V},\mathcal{C}}$ measures how strongly video queries retrieve source-conditioning keys. We refer to this interaction as the source-retrieval grounding pathway, which serves as the main source-to-video grounding mechanism. In contrast, $[S(Q,K)]_{\mathcal{V},\mathcal{V}}$ captures interactions among video tokens themselves, corresponding to the video-side interaction pathway.

Figure 3(a) visualizes these token-level attention structures. The blue rectangle indicates the source-conditioning region, where video queries assign salient attention to source-conditioning keys, suggesting that the source image remains an important reference during denoising. The red rectangle indicates the video-side region, where the strong diagonal structure reflects organized interactions among video tokens. These observations motivate our source grounding detachment objective, which suppresses direct source retrieval and promotes video-side interactions, making the denoising process less dominated by source conditioning.

4 I2V-DETACH

In this section, we present I2V-DETACH, an image immunization framework that reduces faithful source-content preservation by detaching the generated video from the protected source image. Instead of inducing generic temporal artifacts or reducing visual fidelity, our goal is to weaken the attention-mediated pathway through which source-image cues are repeatedly propagated into video latents during denoising.

4.1 Overview

Given a source image x and a frozen I2V diffusion denoiser ϵ_θ, I2V-DETACH learns an imperceptible perturbation δ under an ℓ_∞ budget η. The resulting immunized image $x_{\text{adv}} = x + \delta^*$ is optimized by minimizing a source grounding detachment loss over denoising timesteps:

$$\delta^* = \arg \min_{\|\delta\|_\infty \le \eta} \mathbb{E}_{\tau \sim \mathcal{T}} \left[\mathcal{L}_{\text{sgd}}(x + \delta, \tau; \epsilon_\theta) \right]. \tag{3}$$

Our method directly intervenes in the attention-mediated source-to-video grounding pathway of the I2V denoiser at timestep τ. The key idea is to suppress source-conditioned retrieval from video queries to source-conditioning keys, while reinforcing video-side interactions among video tokens.

Our optimization has two complementary components. The *source-video decoupling* objective suppresses attention logits from video queries to condition-image keys, reducing direct retrieval of protected source cues. To obtain informative optimization signals, we further construct auxiliary source-divergent video latents that originate from the source image but progressively deviate from

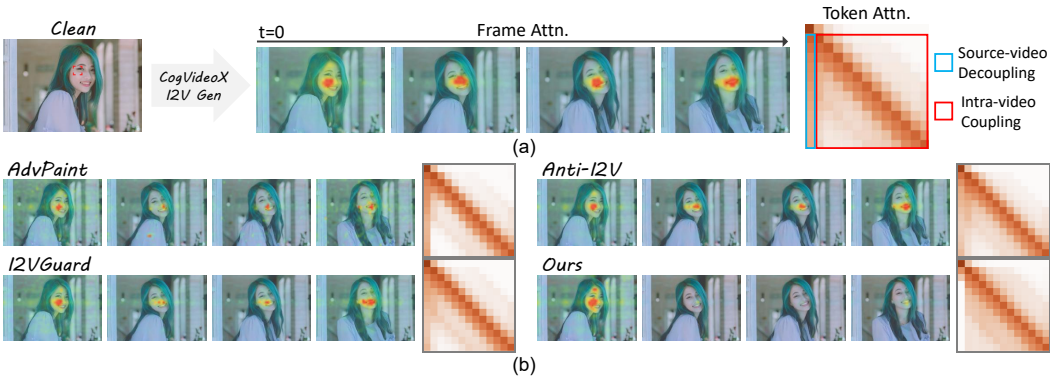

Figure 3: Frame- and token-level attention visualizations from a single denoising pass using video latents generated from the clean source image. In the frame-level attention maps, we highlight the most attended video tokens, with their corresponding spatial regions marked on the clean frame. (a) shows the result for the clean source image, while (b) shows the results for immunized images obtained by different methods.

simple source repetition. The *intra-video coupling* objective reinforces attention logits among video tokens, encouraging denoising to proceed through video-side interactions rather than repeatedly returning to the source image as an anchor. The overall framework is shown in Figure 2. The following sections define these components and present the final optimization objective.

4.2 Source Grounding Detachment Optimization

Source-divergent Video Latent Construction. Effective source-video decoupling requires video latents that provide meaningful source-conditioned signals. Random or source-unrelated latents are weakly coupled with the I2V denoising process in the early timestep, whereas simple source repetition can reinforce shortcut anchoring to the input image. We therefore construct source-divergent video latents that start from the source image but gradually move away from the clean source latent.

At each outer perturbation iteration n, we update Δ^{n-1} once using a small PGD-ascent step with step size β. We then encode the source-divergent frame $x + \Delta^n$ into a single latent and replicate it across the video-side positions to obtain source-divergent latents $\bar{Z}^n$:

$$\bar{Z}^n = \mathcal{R}_{F-1}(\mathrm{sg}[\bar{z}^n]),\ \ \bar{z}^n = \mathcal{E}(x+\Delta^n),\ \ \Delta^n = \Delta^{n-1} + \beta \cdot \mathrm{sign}\left(\nabla_{\Delta^{n-1}} \left\| \mathcal{E}(x+\Delta^{n-1}) - \mathcal{E}(x) \right\|_2^2\right), \tag{4}$$

where $\mathcal{E}$ denotes the 3D-VAE encoder applied to the input after clamping it to $[0, 1]$, F is the total number of visual latents, $\mathcal{R}_m(z)$ denotes the operation that repeats the latent z m times along the temporal dimension, and $\mathrm{sg}[\cdot]$ denotes the stop-gradient operation. After sampling a diffusion timestep $\tau \sim \mathcal{T}$, we add forward-process noise to obtain the source-divergent video latents $\bar{Z}^n_\tau$. During the update of δ, these latents are detached, ensuring that the optimization modifies only the source-image perturbation.

The step-wise update of Δ prevents the auxiliary video latents from becoming either source-unrelated or trivially source-repeated. By keeping the latents source-conditioned while gradually deviating from exact source copies, it provides informative grounding signals for $\mathcal{L}_{\mathrm{svd}}$ across a broader range of source-related states. Given the auxiliary video latents $\bar{Z}^n_\tau$, we run the frozen denoiser and compute two complementary losses: source-video decoupling and intra-video coupling.

Source-Video Decoupling Loss. To disrupt source-guided attention in I2V models, we directly suppress the pre-softmax attention logits from video queries to condition-image keys across selected layers L. Operating at the logit level allows us to isolate this query-key alignment before attention aggregation mixes in responses from other frame keys, as follows:

$$\mathcal{L}_{\mathrm{svd}} = \frac{1}{|L|} \sum_{\ell \in L} \frac{1}{|\mathcal{V}|\,|\mathcal{C}|} \sum_{i \in \mathcal{V}} \sum_{c \in \mathcal{C}} \left[S(Q^\ell, K^\ell) \right]_{ic}, \tag{5}$$

where $S(\cdot, \cdot)$ denotes the pre-softmax attention logit function defined in Eq. (2). Both Q^ℓ and K^ℓ are obtained from the joint latent sequence $[z_{\mathrm{adv}}; \bar{Z}^n_\tau]$ and at layer ℓ and timestep τ and iteration n. The

subscripts $\mathcal{C}$ and $\mathcal{V}$ indicate indexing over the source-token and video-token subsets, respectively. For architectures without a unified full-attention sequence, Q^ℓ and K^ℓ are taken from the corresponding source-to-video or video-side attention module. By suppressing these responses, we prevent video queries from retrieving source-image information from condition-image keys, weakening source-guided visual grounding during denoising.

Intra-video Coupling Loss. We further introduce intra-video coupling loss, which encourages video-side self-interactions to weaken temporal anchoring to the conditioning image. Specifically, we maximize the attention logits among the video-side latents:

$$\mathcal{L}_{\text{ivc}} = -\frac{1}{|L|}\sum_{\ell\in L}\frac{1}{|\mathcal{V}|^2}\sum_{i\in\mathcal{V}}\sum_{j\in\mathcal{V}}\left[S\big(Q^\ell, K^\ell\big)\right]_{ij}. \tag{6}$$

This reinforces query–key matching among video tokens, increasing the response of video queries to video keys and countering the condition-dominant attention pattern. Denoising is thus steered toward video-side interactions, hindering a return to the source-guided trajectory.

Overall Objective. We optimize the adversarial perturbation δ by jointly minimizing the source-video decoupling loss and the video-side interaction loss:

$$\mathcal{L}_{\text{sgd}} = \mathcal{L}_{\text{svd}} + \lambda\mathcal{L}_{\text{ivc}}, \tag{7}$$

where λ is a hyperparameter set to 1 in all experiments. The effect of the proposed loss is illustrated in Figure 3(b). Compared with prior methods [9, 25, 17, 45], I2V-DETACH substantially weakens salient attention between the protected source image and clean-source generated video latents, indicating effective detachment from the source-video anchor. The token-level attention maps also show strengthened video-side interactions, consistent with our intra-video coupling objective. This trend is further quantified by the attention mass in Tables 9, 10, and 11 of Appendix B.1, where I2V-DETACH markedly reduces source-video attention mass while increasing video-video attention mass. Together, these results show that denoising is shifted away from source-dominated retrieval and toward source-suppressed video-side interactions, leading to visually source-detached videos.

4.3 Overall Algorithm

The overall I2V-DETACH framework and optimization procedure are shown in Figure 2 and Algorithm 1, respectively. We first perform a one-step update of the source-divergent video latents using Eq. (4) with a small step size. We then sample a diffusion timestep and add the corresponding forward-process noise to the video latents. The perturbation δ is then updated using our objective in Eq. (7), which disrupts source grounding and prevents the generated video from faithfully preserving the visual content of the source image.

Algorithm 1 I2V-DETACH

Input: Source image x, I2V model ϵ_θ, perturbation budget η, step size α, iterations N, diffusion timestep set $\mathcal{T}$, visual latent count F
Output: Immunized image x_{adv}.

Initialize $\delta \leftarrow 0$, $\quad \Delta^0 \sim \mathcal{N}(0, 0.01^2)$
for $n = 1, \ldots, N$ **do**
  $z_{adv} \leftarrow \mathcal{E}(\text{Clip}(x+\delta, 0, 1))$ // Encode immunized image
  Update Δ^n & source-divergent video latent by Eq. (4)
  $\bar{Z}^n \leftarrow \mathcal{R}_{F-1}(\text{sg}[\mathcal{E}(\text{Clip}(x+\Delta^n, 0, 1))])$
  Sample diffusion timestep $\tau \sim \mathcal{T}$
  Add the forward-process noise to video latents
  $\bar{Z}^{\text{n}}_\tau \leftarrow \text{AddNoise}(\bar{Z}^{\text{n}}, \tau; \epsilon_\theta)$
  $L_{\text{sgd}} \leftarrow L_{\text{svd}} + \lambda L_{\text{ivc}}$ // Compute loss using Eq. (5) and (6)
  $\delta \leftarrow \delta - \alpha \cdot \text{sign}(\nabla_\delta L_{\text{sgd}})$ // Update perturbation
  $\delta \leftarrow \text{Clip}(\delta, -\eta, \eta)$ // Project δ into ℓ_∞ ball
end for
return $x_{\text{adv}} = \text{Clip}(x+\delta, 0, 1)$

5 Experiments

5.1 Experimental Setup

Implementation details. To examine the generality of our method across diverse DiT-based I2V architectures, we evaluate three representative models feasible for video generation and perturbation optimization within an 80GB memory budget: CogVideoX-5B [52] with MMDiT full-attention, LTX-Video-2B [20] with cross-attention, and DiT-based OpenSora 1.2 [57]. For our method, we sample diffusion timestep from the first 10% of the denoising step, set the step sizes $\alpha = 1/255$ and $\beta = 2/255$, and use the 25–75% layer range for all models. For a fair comparison, we set perturbation budget $\eta = 8/255$ following I2VGuard [17] and details for baselines are provided in Appendix A.1.

Baselines. We first compare our method with state-of-the-art I2V immunization methods, including I2VGuard [17] and Anti-I2V [45], which are specifically designed to defend against malicious image-to-video generation. We further include EditShield [9], which operates on latent-level representations,



Figure 4: Qualitative results for our method across diverse diffusion-based I2V architectures [52, 20, 57]. Regardless of the underlying model, our method consistently produces visually detached videos.

Table 1: White-box performance on CogVideoX [52], LTX-Video [20], and OpenSora 1.2 [57]. *Space* denotes the perturbation domain used for optimization. *V/P* denotes that methods require a *pre-generated video* and *prompt* during perturbation optimization, respectively. * refers to the variant that incorporates purification-aware optimization. The best score is shown in **bold** and the second-best is underlined.

Model			CogVideoX					LTX-Video					OpenSora 1.2				
Method	Space	V/P	IS ↓	SC ↓	DS ↓	AQ	IQ	IS ↓	SC ↓	DS ↓	AQ	IQ	IS ↓	SC ↓	DS ↓	AQ	IQ
Clean	-	-	97.28	94.83	89.17	60.46	70.03	97.12	94.22	81.04	57.85	67.60	95.73	95.17	87.30	56.95	66.10
EditShield [9]	RGB	✗/✗	95.97	93.77	86.09	60.72	65.41	92.63	86.85	69.62	51.04	57.04	86.39	88.98	70.08	53.52	62.64
AdvPaint [25]	RGB	✗/✓	93.27	89.90	78.42	57.11	63.64	93.87	90.33	73.24	54.43	59.18	84.47	83.39	58.80	48.92	56.10
I2VGuard [17]	RGB	✓/✓	94.39	92.52	82.99	58.65	63.39	94.10	89.67	73.37	54.47	61.08	92.20	92.68	77.68	56.10	62.19
Ours	RGB	✗/✗	86.39	80.63	65.54	57.87	66.76	89.85	81.98	60.88	50.35	58.51	79.62	73.92	52.10	48.95	52.38
Anti-I2V* [45]	Lab+DCT	✓/✓	93.13	92.42	82.34	60.97	65.38	93.80	89.50	75.12	53.40	63.77	91.64	91.97	77.33	55.30	62.03
Ours*	RGB+DCT	✗/✗	**83.52**	**78.30**	**63.33**	59.52	67.55	**88.70**	**81.08**	**58.94**	49.73	58.52	**78.95**	**73.70**	**51.40**	48.98	52.55

and AdvPaint [25], which perturbs attention embedding representations. Although originally developed for I2I settings, these methods are relevant to our study because they intervene on representation spaces closely associated with source-content preservation. This selection allows us to evaluate our method against both task-specific I2V baselines and objective-relevant I2I baselines.

Evaluation dataset and metrics. Since there is no benchmark designed for I2V immunization, we sample 100 image-prompt pairs from VBench [23], widely used for I2V generation, including 80 human-related images and 20 non-human images. For evaluation, we use *I2V Subject* (IS) and *Subject Consistency* (SC) from VBench [23, 24], along with *DINO Similarity* (DS) [39] between the source image and generated frames. These metrics assess how strongly the generated video preserves the source subject and visual content, while lower values indicate stronger immunization. We also include quality metrics to monitor the perceptual quality of generated videos, *aesthetic quality* (AQ) and *imaging quality* (IQ) from VBench.

5.2 Main Results

Qualitative comparison. We first qualitatively compare our method against the baselines on CogVideoX [52]. As shown in Figure 1, I2V-DETACH produces visually source-detached videos, whereas the baselines mainly induce perceptual degradation while still retaining source-related content. We also present qualitative results on diverse diffusion-based I2V models in Figure 4. While

Table 2: Human evaluation results. We evaluate each method in terms of *Source Reflection* (Src. Ref.), *Protection Assessment* (Pro. Ass.), and *Overall Effectiveness* (Ovr. Eff.). ES, AP, IG, and AI denote EditShield [9], AdvPaint [25], I2VGuard [17], and Anti-I2V [45], respectively.

Model	CogVideoX					LTX-Video					OpenSora 1.2				
Method	ES	AP	IG	AI	Ours	ES	AP	IG	AI	Ours	ES	AP	IG	AI	Ours
Src. Ref. ↓	4.34	3.33	4.13	4.33	**1.46**	2.46	2.99	4.09	3.06	**1.49**	3.76	2.94	4.04	3.86	**1.64**
Pro. Ass. ↑	0.20	0.20	0.20	0.27	**4.12**	0.85	0.20	0.85	0.45	**2.65**	0.32	0.90	0.22	0.27	**3.27**
Ovr. Eff. ↓	4.10	2.50	3.05	4.00	**1.35**	2.25	3.30	4.45	3.60	**1.40**	4.00	2.10	3.95	3.55	**1.40**

Table 3: Robustness evaluation against transformation and diffusion-based purification techniques. * refers to the variant that incorporates purification-aware optimization. The best score is shown in **bold** and the second-best is underlined.

Target Model	JPEG [18]			Gaussian noise			Adverse Clean [14]			GridPure [55]		
Method	IS ↓	SC ↓	DS ↓	IS ↓	SC ↓	DS ↓	IS ↓	SC ↓	DS ↓	IS ↓	SC ↓	DS ↓
EditShield [9]	94.95	92.74	84.37	96.03	94.19	86.55	95.30	93.81	84.84	96.30	94.36	87.00
AdvPaint [25]	94.42	92.30	83.17	94.14	96.03	86.55	95.49	94.28	85.66	95.53	93.50	85.48
I2VGuard [17]	93.69	92.29	82.38	94.19	92.72	82.85	94.66	93.10	83.26	94.55	93.12	83.27
Ours	90.21	86.11	74.26	87.57	81.85	67.54	90.31	86.09	73.62	91.18	86.12	74.81
Anti-I2V* [45]	93.33	92.32	83.08	92.92	92.56	82.22	94.23	92.88	84.49	93.39	92.47	83.25
Ours*	**84.98**	**80.14**	**66.57**	**83.80**	**78.89**	**64.32**	**88.10**	**83.92**	**70.51**	**85.06**	**79.60**	**64.83**

the resulting generation patterns differ across architectures, a common trend is clearly observed: all three models fail to faithfully preserve the visual content of the source image. This indicates that the proposed perturbation consistently breaks source grounding, rather than merely inducing temporal degradation. More qualitative comparisons and discussion are provided in Appendix D.1.

Quantitative comparison. We evaluate immunization performance across diverse diffusion-based I2V models. As shown in Table 1, our method consistently outperforms the baselines by a substantial margin on consistency metrics (*i.e.*, IS, SC, and DS), while requiring neither additional reference videos nor prompts, in contrast to the baselines [25, 17, 45]. In addition, I2V-DETACH demonstrates strong generalizability across diverse attention architectures [52, 20, 57]. We further evaluate an RGB+DCT variant, denoted as *Ours**, to validate compatibility with robustness-oriented perturbation designs [3, 45]. Its improved performance suggests that our objective can be effectively extended beyond RGB-space perturbations to the DCT domain. AQ/IQ trends differ by model, likely because each I2V model fails under immune perturbation in different ways. CogVideoX [52] tends to preserve clean alternative videos, whereas LTX-Video [20] severely collapses and OpenSora 1.2 [57] partially collapses while still producing source-deviated videos, but the key point is the broken source grounding rather than the quality score.

Human Evaluation. We conduct human evaluation with 20 crowdsourced participants along three dimensions: (i) *Source Reflection* (SR), how faithfully the generated video preserves the visual content of the source image on a 1–5 scale; (ii) *Protection Assessment* (PA), select one method that best protects the source image from a protection perspective; and (iii) *Overall Effectiveness* (OE), rank the methods according to their overall protection effectiveness. For the human evaluation, we randomly sample 10 videos from each model and method. All scores are converted to a 5-point scale. Lower is better for SR and OE while higher scores indicate stronger protection for PA. As shown in Table 2, the lowest SR indicates that human evaluators consistently find that I2V-DETACH least preserves the visual content of the source image across all models. Furthermore, our method is most frequently selected in PA and receives the best ranking in OE, suggesting that source grounding detachment provides a suitable protection criterion for unauthorized I2V generation. Details are provided in Appendix A.6.

5.3 Robustness Evaluation against Purification

We evaluate the robustness of I2V-DETACH against perturbation purification. Specifically, we consider preprocessing-based purification, including JPEG compression [18], Gaussian noise, and Adverse Cleaner [14], as well as diffusion-based GridPure [55]. As shown in Table 3, our original RGB-space method already achieves the strongest protection performance under perturbation purification compared with baselines. Furthermore, the additional robustness of *Ours**, obtained by incorporating DCT perturbations, shows that our objective can be integrated with existing purification-aware techniques for improved robustness.

Table 4: Transferability evaluation on LTX-Video [20] and OpenSora 1.2 [57]. CogVideoX [52] is used for victim model. The best score is shown in **bold** and the second-best is underlined.

Target Model	LTX-Video					OpenSora 1.2				
Method	IS ↓	SC ↓	DS ↓	AQ	IQ	IS ↓	SC ↓	DS ↓	AQ	IQ
EditShield [9]	94.42	91.23	77.72	53.09	62.52	94.86	94.51	84.51	54.85	63.07
AdvPaint [25]	94.67	91.41	76.02	53.93	62.30	94.69	94.28	83.46	54.93	62.99
I2VGuard [17]	93.82	91.29	76.36	53.35	61.95	93.26	94.08	81.02	54.80	61.84
Anti-I2V [45]	92.57	90.21	74.25	54.17	62.97	93.74	94.55	82.33	58.63	63.44
Ours	**92.06**	**88.26**	**72.30**	51.23	60.28	**92.62**	**93.58**	**79.07**	54.56	59.35

Table 5: Imperceptibility comparison with baselines. The best score for each metric within each model is shown in **bold**.

Model	CogVideoX [52]				LTX-Video [20]				OpenSora 1.2 [57]			
Method	PSNR ↑	SSIM ↑	LPIPS ↓	DISTS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	DISTS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	DISTS ↓
EditShield	31.54	0.779	0.174	0.210	31.55	0.786	0.388	0.181	31.43	0.778	0.400	0.232
AdvPaint	33.06	0.831	0.166	0.162	33.15	0.840	0.357	0.171	32.14	0.829	0.403	0.234
I2VGuard	30.72	0.743	0.229	0.172	29.99	0.723	0.340	0.146	30.36	0.744	0.388	0.193
Anti-I2V	28.66	0.727	0.237	0.207	32.55	0.843	0.387	0.171	25.14	0.628	**0.387**	**0.191**
Ours	**33.49**	**0.848**	**0.155**	**0.139**	**33.48**	**0.853**	**0.320**	**0.145**	**32.61**	**0.837**	0.397	0.214

Table 6: Protection performance on OpenSora 1.2 under matched imperceptibility (DISTS ≈ 0.19). AdvPaint [25] and I2V-DETACH are re-optimized using a perturbation budget of $6/255$.

Category	Metric	ES $(8/255)$	AP $(6/255)$	IG $(8/255)$	AI $(8/255)$	Ours $(6/255)$
Imperceptibility	LPIPS ↓	0.400	0.377	0.388	0.387	**0.354**
	DISTS ↓	0.232	0.193	0.193	0.191	**0.187**
Performance	IS ↓	86.39	88.08	92.20	91.64	**82.92**
	SC ↓	88.98	86.03	92.68	91.97	**77.62**
	DS ↓	70.08	66.81	77.68	77.33	**56.51**

Table 7: Protection performance of image-specific methods and their universal variants. Lower values indicate stronger protection. ES, AP, IG, and AI denote EditShield [9], AdvPaint [25], I2VGuard [17], and Anti-I2V [45], respectively.

Method	*Image-specific*						*UAP-variant*				
	Clean	ES [9]	AP [25]	IG [17]	AI [45]	Ours	ES [9]	AP [25]	IG [17]	AI [45]	Ours
IS ↓	97.28	95.97	93.27	94.39	93.13	**86.39**	96.32	95.58	95.37	95.08	**93.31**
SC ↓	94.83	93.77	89.90	92.52	92.42	**80.63**	94.45	94.01	92.74	93.87	**91.74**
DS ↓	89.17	86.09	78.42	82.99	82.34	**65.54**	87.57	85.73	86.43	85.77	**83.54**

5.4 Black-box Transferability

We evaluate the black-box transferability between different diffusion architectures. In particular, we use CogVideoX (full-attention) [52] as the victim model, and LTX-Video (cross-attention) [20] and OpenSora 1.2 (DiT-based) [57] as the black-box models. As shown in Table 4, our method consistently maintains stronger image protection than competing methods, demonstrating strong black-box transferability to unseen diffusion models with different attention architectures. Additional black-box evaluations, including experiments using LTX-Video and OpenSora 1.2 as surrogate models and transfer to large-scale and commercial models, are provided in Appendix C.6.

5.5 Imperceptibility Analysis

We compare imperceptibility across methods using PSNR, DISTS [15], LPIPS [53], and SSIM [48], as a shared perturbation budget of $\eta = 8/255$ does not guarantee comparable perceptual visibility. As shown in Table 5, I2V-DETACH achieves the best imperceptibility across all metrics on CogVideoX [52] and LTX-Video [20].

On OpenSora 1.2 [57], however, I2V-DETACH yields slightly lower LPIPS and DISTS scores. We therefore re-optimize I2V-DETACH with $\eta = 6/255$, obtaining a DISTS score of approximately 0.19, comparable to I2VGuard [17] and Anti-I2V [45]. We also evaluate AdvPaint [25], the second-best baseline in Table 6, at a comparable perceptual distortion level. As shown in Table 6, I2V-DETACH still demonstrates superior protection performance even when compared under similar perceptual imperceptibility.

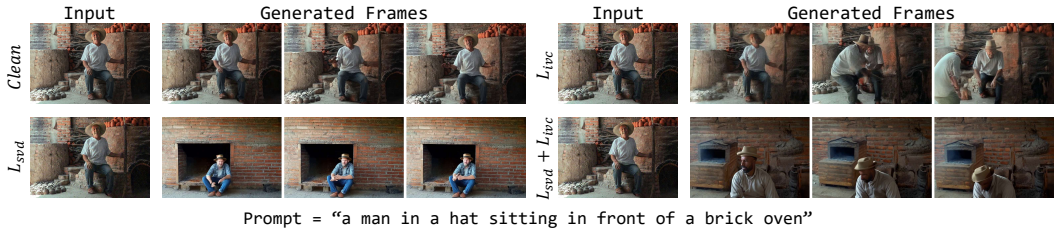

Figure 5: Qualitative ablation for the proposed two loss functions on CogVideoX [52].

Table 8: Ablation studies for the losses and video frames. *Black*, *Source*, and *Rand* denote repeated black frames, repeated source frames, and random noisy latents at the corresponding timestep, respectively, while *SDV* denotes the proposed source-divergent video.

Method	$\mathcal{L}_{\text{svd}}$				$\mathcal{L}_{\text{ivc}}$				$\mathcal{L}_{\text{svd}}+\mathcal{L}_{\text{ivc}}$			
Video	*Unrelated*		*Related*		*Unrelated*		*Related*		*Unrelated*		*Related*	
	Rand	Black	Source	SDV	Rand	Black	Source	SDV	Rand	Black	Source	SDV
IS ↓	95.29	95.06	92.83	**91.83**	90.09	89.39	90.38	**89.31**	89.75	89.14	88.45	**86.39**
SC ↓	91.59	90.98	90.49	**87.49**	85.80	**85.07**	86.50	85.37	84.41	84.23	83.54	**80.63**
DS ↓	83.16	82.55	80.93	**78.24**	74.77	73.03	75.32	**72.37**	72.62	72.60	70.78	**65.54**

5.6 Extension to Universal Adversarial Perturbation

I2V-DETACH introduces a novel source grounding detachment objective for I2V immunization, but image-specific optimization remains costly due to backpropagation through the I2V model. To explore an extension to universal adversarial perturbations (UAPs) [30] that protect different images without per-image optimization, we train UAP variants of I2V-DETACH and all baselines. We train all UAP variants on 250 VBench [23] images disjoint from the evaluation set, using CogVideoX [52] as the surrogate with a shared ℓ_∞ budget of $8/255$. Table 7 shows that I2V-DETACH leads all UAP variants across all three metrics, outperforming even image-specific EditShield [9] and I2VGuard [17]. Despite a gap from its image-specific counterpart, these results support our objective's applicability to universal image protection.

5.7 Ablation Studies

We conduct ablation experiments to examine the effectiveness of the proposed components. We use CogVideoX [52] as the victim model.

Impact of Loss Functions. We ablate the two losses, $\mathcal{L}_{\text{svd}}$ and $\mathcal{L}_{\text{ivc}}$. As shown in Table 8, under our proposed source-divergent video (SDV) setting, each loss improves over the all baselines as in Table 1, while their combination achieves the best performance, confirming their complementary roles. Qualitatively, Figure 5 shows that $\mathcal{L}_{\text{svd}}$ alone detaches the generated video from the source to some extent, but the result may still preserve source-like identity cues. In contrast, $\mathcal{L}_{\text{ivc}}$ alone increases video dynamics but does not detach the generation from the source. The two losses, jointly designed for source grounding detachment, produce stronger source detachment when combined.

Video Latent Construction. We investigate the effect of video initialization on I2V-DETACH. As shown in Table 8, $\mathcal{L}_{\text{svd}}$ is effective mainly with source-related videos, whereas $\mathcal{L}_{\text{ivc}}$ is relatively insensitive to the video type. Moreover, $\mathcal{L}_{\text{svd}} + \mathcal{L}_{\text{ivc}}$ substantially improves over $\mathcal{L}_{\text{ivc}}$ with source-related videos, but yields only marginal gains with source-unrelated videos. This is expected, since our objective is not to drive generation toward source-unrelated videos, but to detach it from source grounding. This indicates that source-related videos provide the necessary source-grounding signals for $\mathcal{L}_{\text{svd}}$ to meaningfully suppress source reliance, while our source-divergent videos offer more diverse video-side signals than simple repeated source images.

6 Conclusion

We introduced a novel I2V immunization objective *source grounding detachment*, shifting the goal from temporal degradation to disrupting source-guided visual grounding. We presented I2V-DETACH, which prevents unauthorized video generation by suppressing source-video responses and amplifying intra-video interactions, thereby weakening the source-dominant attention pathway. We further introduced source-divergent video construction, which provides diverse source-oriented signals for our objective. Extensive experiments across diverse diffusion-based I2V models demonstrate that our method consistently achieves stronger immunization performance than existing baselines.

Acknowledgments

This work was supported by the AI Graduate School Program at POSTECH (RS-2019-II191906 (5%)), the NRF grants (RS-2026-25491789 (35%)), and the IITP grants (RS-2026-25518317 (30%), RS-2026-25616370 (30%)) funded by MSIT, Korea.

References

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

[2] Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2018.

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

[4] Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Guanghui Liu, Amit Raj, et al. Lumiere: A space-time diffusion model for video generation. In *SIGGRAPH Asia*, 2024.

[5] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

[6] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2023.

[7] Minghong Cai, Xiaodong Cun, Xiaoyu Li, Wenze Liu, Zhaoyang Zhang, Yong Zhang, Ying Shan, and Xiangyu Yue. Ditctrl: Exploring attention control in multi-modal diffusion transformer for tuning-free multi-prompt longer video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 2025.

[8] Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 2024.

[9] Ruoxi Chen, Haibo Jin, Yixin Liu, Jinyin Chen, Haohan Wang, and Lichao Sun. Editshield: Protecting unauthorized image editing by instruction-guided diffusion models. In *European Conference on Computer Vision (ECCV)*, 2024.

[10] June Suk Choi, Kyungmin Lee, Jongheon Jeong, Saining Xie, Jinwoo Shin, and Kimin Lee. Diffusionguard: A robust defense against malicious diffusion-based image editing. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025.

[11] June Suk Choi, Kyungmin Lee, Sihyun Yu, Yisol Choi, Jinwoo Shin, and Kimin Lee. Enhancing motion dynamics of image-to-video models via adaptive low-pass guidance. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 2026.

[12] 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.

[13] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 45(9):10850–10869, 2023.

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

[15] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 44(5):2567–2581, 2020.

[16] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning (ICML)*, 2024.

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

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

[19] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.

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

[21] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[22] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.

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

[24] Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, et al. Vbench++: Comprehensive and versatile benchmark suite for video generative models. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2025.

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

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

[27] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

[28] Black Forest Labs. FLUX.2: Frontier Visual Intelligence. https://bfl.ai/blog/flux-2, 2025.

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

[30] Chanhui Lee, Donggyu Choi, Seunghyun Shin, Hae-Gon Jeon, and Jeany Son. Universal image immunization against diffusion-based image editing via semantic injection. In *European Conference on Computer Vision (ECCV)*, 2026.

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

[32] Yiran Liu, Xin Feng, Yunlong Wang, Wu Yang, and Di Ming. Trm-uap: Enhancing the transferability of data-free universal adversarial perturbation via truncated ratio maximization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

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

[34] Luma AI. Luma introduces ray3.2 model & api: Complete creative control for video generation. https://lumalabs.ai/news/introducing-ray-3-2, June 2026. Accessed: 2026-09-08.

[35] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018.

[36] Momina Masood, Mariam Nawaz, Khalid Mahmood Malik, Ali Javed, Aun Irtaza, and Hafiz Malik. Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward. *Applied intelligence*, 53(4):3974–4026, 2023.

[37] Yisroel Mirsky and Wenke Lee. The creation and detection of deepfakes: A survey. *ACM computing surveys (CSUR)*, 54(1):1–41, 2021.

[38] Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Anima Anandkumar. Diffusion models for adversarial purification. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2022.

[39] 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 (TMLR)*, 2024.

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

[41] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2022.

[42] Hadi Salman, Alaa Khaddaj, Guillaume Leclerc, Andrew Ilyas, and Aleksander Madry. Raising the cost of malicious ai-powered image editing. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2023.

[43] Seunghyun Shin, Jifei Song, Wooseok Jeon, Hae-Gon Jeon, and Jiankang Deng. Trimotion: Modality-agnostic camera control for video generation. In *European Conference on Computer Vision (ECCV)*, 2026.

[44] Matteo Trippodo, Federico Becattini, and Lorenzo Seidenari. Immunizing images from text to image editing via adversarial cross-attention. In *Proceedings of the 33rd ACM International Conference on Multimedia (ACM MM)*, 2025.

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

[46] Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 3(4):6, 2025.

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

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

[49] Mika Westerlund. The emergence of deepfake technology: A review. *Technology innovation management review*, 9(11), 2019.

[50] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[51] Haocheng Xi, Shuo Yang, Yilong Zhao, Chenfeng Xu, Muyang Li, Xiuyu Li, Yujun Lin, Han Cai, Jintao Zhang, Dacheng Li, et al. Sparse video-gen: Accelerating video diffusion transformers with spatial-temporal sparsity. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2025.

[52] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025.

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

[54] Min Zhao, Hongzhou Zhu, Chendong Xiang, Kaiwen Zheng, Chongxuan Li, and Jun Zhu. Identifying and solving conditional image leakage in image-to-video diffusion model. *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[55] Zhengyue Zhao, Jinhao Duan, Kaidi Xu, Chenan Wang, Rui Zhang, Zidong Du, Qi Guo, and Xing Hu. Can protective perturbation safeguard personal data from being exploited by stable diffusion? In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 2024.

[56] Zangwei Zheng, Xiangyu Peng, Yuxuan Lou, Chenhui Shen, Tom Young, Xinying Guo, Binluo Wang, Hang Xu, Hongxin Liu, Mingyan Jiang, et al. Open-sora 2.0: Training a commercial-level video generation model in $200 k. *arXiv preprint arXiv:2503.09642*, 2025.

[57] Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all. *arXiv preprint arXiv:2412.20404*, 2024.

I2V-DETACH:Preventing Unauthorized Image-to-Video Generations via source grounding Detachment

-Supplementary Materials-

A Experimental Details **16**

A.1 Implementation Setup . . . 16

A.2 Threat Models . . . 16

A.3 Evaluation Dataset and Metrics . . . 17

A.4 Baseline Selection and Implementation Details . . . 18

A.5 Details of Purification . . . 18

A.6 Details of Human Evaluation . . . 19

B Additional Analysis **20**

B.1 Attention Mass Shift . . . 20

B.2 Analysis of Source-divergent Video . . . 20

C Additional Experiments **21**

C.1 Effect of Layer Selection . . . 21

C.2 Perturbation Budgets . . . 21

C.3 Source-divergent Video Optimization . . . 21

C.4 Range of Timestep Sampling . . . 22

C.5 Optimization with Diverse Video . . . 22

C.6 Additional Black-box Transferability . . . 23

D Qualitative Results **23**

D.1 Comparison with Baselines . . . 23

D.2 Qualitative Results Beyond Human-Related Images . . . 24

D.3 Black-box Results . . . 24

E Failure Cases **24**

F Limitations and Discussion **24**

G Broader Impact **25**

H Future Work **25**

I Reproducibility **25**

A Experimental Details

A.1 Implementation Setup

For our method, we set the step sizes for δ and Δ to $1/255$ and $2/255$, respectively. We use the 25–75% layer range across all I2V models. Since our objective targets video-token interactions commonly present in I2V models, we apply it to the corresponding attention module of each architecture without additional model-specific design: full attention in CogVideoX [52], self-attention in LTX-Video [20], and temporal attention in OpenSora 1.2 [57]. For a fair comparison, we set the perturbation budget to $\eta = 8/255$ and use 50 attack iterations. We conduct all experiments on an NVIDIA A100 GPU, and the optimization time is approximately 10 minutes with 62 GB memory.

We sample diffusion timesteps from the first 10% of the denoising trajectory, where coarse video structure is established under strong source-image guidance [11]. By intervening before the denoising trajectory becomes stabilized, I2V-DETACH more effectively disrupts source grounding and prevents the generated video from faithfully preserving the source-image content.

A.2 Threat Models

We consider I2V models with diverse attention architectures–MMDit full-attention, MMDit cross-attention, and Dit-based–to evaluate the generality of I2V-DETACH. We select representative models from each architecture that can be reasonably used for immunization on a single NVIDIA A100 80GB GPU, considering factors such as model size and the number of frames required for high-quality video generation.

Full-attention. In full-attention I2V models, condition, text, and video tokens are jointly processed within a unified attention module, allowing direct interaction across all modalities. Representative models include Hunyuan-Video [27] and OpenSora 2.0 [56]. In this category, we use CogVideoX-5B [52] as the primary model.

For CogVideoX-5B [52], text tokens precede visual tokens. Let N_{text} denote the number of text tokens, s_p the number of tokens per latent frame, and F the number of latent frames. We instantiate the losses as follows:

- **Source-video decoupling.** We use video-side queries from frames $1, \dots, F-1$ and source/condition keys from frame 0:

$$Q_{\mathcal{V}} = q_{\ell,h}[N_{\text{text}} + s_p : N_{\text{text}} + Fs_p], \quad K_{\mathcal{C}} = k_{\ell,h}[N_{\text{text}} : N_{\text{text}} + s_p].$$

- **Intra-video coupling.** We use both queries and keys from the video-side frames $1, \dots, F-1$:

$$Q_{\mathcal{V}} = q_{\ell,h}[N_{\text{text}} + s_p : N_{\text{text}} + Fs_p], \quad K_{\mathcal{V}} = k_{\ell,h}[N_{\text{text}} + s_p : N_{\text{text}} + Fs_p].$$

Cross-attention. Cross-attention-based I2V models separate video latent interactions from conditioning signals, where self-attention organizes video latents while cross-attention injects source and text information. Representative models include Wan [46] family. We use LTX-Video-2B [20] in our experiments.

For LTX-Video-2B [20], we apply our losses to self-attention over frame-major visual tokens. Let s_p denote the number of tokens per latent frame and F the number of latent frames. We instantiate the losses as follows:

- **Source-video decoupling.** We use video-side queries from frames $1, \dots, F-1$ and source/condition keys from frame 0:

$$Q_{\mathcal{V}} = q_{\ell,h}[s_p : Fs_p], \quad K_{\mathcal{C}} = k_{\ell,h}[0 : s_p].$$

- **Intra-video coupling.** We use both queries and keys from the video-side frames $1, \dots, F-1$:

$$Q_{\mathcal{V}} = q_{\ell,h}[s_p : Fs_p], \quad K_{\mathcal{V}} = k_{\ell,h}[s_p : Fs_p].$$

DiT-based spatio-temporal attention. DiT-based I2V models often decompose spatial and temporal processing into separate attention modules, alternating between spatial and temporal blocks. Representative models include OpenSora variants. In this paper, we use OpenSora 1.2 [57], where spatial attention captures within-frame structure, temporal attention propagates information across frames, and cross-attention incorporates conditioning signals.

For OpenSora 1.2 [57], we compute the losses on temporal attention. Let T denote the number of latent frames. We instantiate the losses as follows:

- **Source-video decoupling.** We use video-side queries $1, \ldots, F-1$ and source/condition keys from frame 0:
$$Q_{\mathcal{V}} = q_{\ell,h}[1:F], \quad K_{\mathcal{C}} = k_{\ell,h}[0:1].$$
- **Intra-video coupling.** We use both queries and keys from the video-side frames $1, \ldots, T-1$:
$$Q_{\mathcal{V}} = q_{\ell,h}[1:F], \quad K_{\mathcal{V}} = k_{\ell,h}[1:F].$$

A.3 Evaluation Dataset and Metrics

Dataset. To ensure a reliable evaluation setting and improve the credibility of the measured performance, we build our evaluation set using 100 image-prompt pairs sampled from VBench [23], consisting of 80 human-related images and 20 non-human images such as animals or natural scenes. Since our evaluation metrics are also adopted from VBench series [23, 24], using samples from the same benchmark provides a more consistent evaluation protocol.

Metrics. We evaluate immunization performance using five metrics: *I2V Subject* (IS), *Subject Consistency* (SC), DINO similarity (DS), *Aesthetic Quality* (AQ), and *Imaging Quality* (IQ), where IS, SC, AQ, and IQ are oriented from VBench. Lower IS, SC, and DS indicate stronger immunization, while AQ and IQ are reported to monitor the perceptual quality of generated videos.

I2V Subject (IS). IS measures how strongly the generated video preserves the subject of the source image while maintaining temporal coherence. For a source image x and generated frames $\{v_t\}_{t=1}^{T}$, we compute the DINO-based cosine similarities between the source image and each frame,
$$c_t = \max\big(0, \cos(\phi(x), \phi(v_t))\big), \tag{8}$$
and between consecutive frames,
$$r_t = \max\big(0, \cos(\phi(v_{t-1}), \phi(v_t))\big), \qquad t = 2, \ldots, T. \tag{9}$$
The final IS score is defined as
$$\mathrm{IS} = 0.4 \cdot \max_t c_t + 0.3 \cdot \frac{1}{T-1} \sum_{t=2}^{T} r_t + 0.3 \cdot \min_t r_t. \tag{10}$$

Subject Consistency (SC). SC measures temporal subject coherence within the generated video. We compute the average, over all frames, of the mean cosine similarity to the first frame and to the immediately preceding frame:
$$\mathrm{SC} = \frac{1}{T-1} \sum_{t=2}^{T} \frac{\max\big(0, \cos(\phi(v_1), \phi(v_t))\big) + \max\big(0, \cos(\phi(v_{t-1}), \phi(v_t))\big)}{2}. \tag{11}$$

DINO Similarity (DS). DS directly measures how strongly the generated frames preserve the visual content of the source image. It is computed as the average cosine similarity between the source image and each generated frame in the DINO feature space:
$$\mathrm{DS} = \frac{1}{T} \sum_{t=1}^{T} \max\big(0, \cos(\phi(x), \phi(v_t))\big). \tag{12}$$

Aesthetic Quality (AQ). AQ measures the perceptual aesthetic quality of the generated video. Following VBench, we compute frame-level aesthetic scores using the LAION aesthetic predictor and average them over all generated frames:
$$\mathrm{AQ} = \frac{1}{T} \sum_{t=1}^{T} a(v_t), \tag{13}$$

where $a(\cdot)$ denotes the aesthetic scoring model.

Imaging Quality (IQ). IQ measures the low-level visual quality of the generated video. Following VBench, we compute MUSIQ scores for each generated frame and average them over the video:

$$\text{IQ} = \frac{1}{T}\sum_{t=1}^{T} q(v_t), \tag{14}$$

where $q(\cdot)$ denotes the MUSIQ-based image quality predictor.

A.4 Baseline Selection and Implementation Details

Image immunization has been actively studied in I2I generation [42, 44, 25, 26, 33, 47, 10]. Among prior methods, we adapt EditShield [9] and AdvPaint [25] as representative baselines, since they effectively disrupt how diffusion models interpret the source image at the encoder and attention levels, respectively. In the I2V immunizations, we select two state-of-the-art I2V baselines, I2VGuard [17] and Anti-I2V [45]. Detailed explanations are provided below.

EditShield. EditShiled [9] first introduced an image protection method against instruction-guided image editing models such as InstructPix2Pix [6]. Since such models repeatedly use the latent of the conditioning image during the editing process, EditShield [9] maximizes the ℓ_2 distance between the clean latent and the perturbed latent to disrupt faithful source-image interpretation. To improve robustness, it further incorporates EOT [2] during perturbation optimization. The perturbation is updated by PGD [35], following their original implementation [9].

AdvPaint. To mitigate the risks of misusing image inpainting, AdvPaint [25] proposed immunization method against diffusion-based inpainting. It perturbs attention-related features of the diffusion denoiser so that the model misinterprets the source image during generation. Specifically, we perturb the query, key, and value features of the condition image in CogVideoX [52]; the condition-image self-attention query, key, and value as well as the cross-attention query in LTX-Video [20]; and the query, key, and value features in spatial blocks for Open-Sora [57]. In our implementation, following the original setup, we use a word representing the source image (*e.g.*, *man* or *woman*) as the prompt when computing the cross-attention query and for residual update to deeper layers. The perturbation is updated by PGD [35], following their original implementation [25].

I2VGuard. I2VGuard [17] is one of the first methods to explicitly address image immunization against I2V generation. It optimizes perturbations using multiple objectives, including an encoder attack, alignment of the perturbed attention map with that of a fixed target video (*i.e.*, a black video), and alignment of the denoised video latent with the same target, in order to induce temporal degradation and reduce generation fidelity. At the same time, it employs a contrastive objective that pushes the denoised video latent of the perturbed input away from that of the pre-generated clean outputs. The perturbation is updated by gradient descent algorithm, following their original implementation [17].

Anti-I2V. To prevent malicious I2V generation, Anti-I2V [45] introduced IRC loss, which encourages shallow intermediate representations to collapse toward those of deeper layers, thereby inducing information loss, and IRA loss, which aligns the latent and deep features with those of a target to promote temporal degradation and reduce generation fidelity. To improve robustness against purification, it further proposes dual-space perturbation, which jointly optimizes perturbations in both the Lab and DCT domains. The perturbation is updated by AdamW optimizer, following their original implementation [45].

A.5 Details of Purification

We consider two categories of purification methods. The first consists of transformation-based methods, including JPEG compression [18], Gaussian noise, and Adverse Cleaner [14], which are commonly used to suppress high-frequency adversarial perturbations. The second is diffusion-based purification method that directly target adversarial perturbations in diffusion models. In this category, we use GridPure [55], an effective diffusion-based purification method designed for high-resolution images.

JPEG Compression. JPEG compression [18] is originally designed to reduce image storage and transmission costs through quantization in the DCT domain. This quantization process can cause information loss in carefully crafted perturbations, particularly those concentrated in high-frequency components.

Gaussian Noise Gaussian noise is one of the simplest image degradation methods, but it can still weaken adversarial perturbations by directly corrupting the crafted perturbation pattern with random noise, thereby partially overwriting the perturbation signal embedded in the image.

Adverse Cleaner Adverse cleaner [14] is an effective purifier that removes high-frequency artifacts by repeatedly applying bilateral filtering and then refining the result with guided filtering, while preserving the main visual structure of the original image.

GridPure GridPure [55] is designed to remove adversarial perturbations from high-resolution images by dividing the image into grids, performing slicing and merging, and applying small step DiffPure [38] iteratively to each grid.

A.6 Details of Human Evaluation

We provide additional details for the human evaluation described in Section 5.2. We conduct the evaluation on crowdsourcing with 20 participants following prior works [43, 3]. For each I2V model and each method, we randomly sample 10 generated videos for evaluation. The evaluation is designed to assess not only whether the generated videos preserve the source image, but also whether the resulting failure mode is effective from a protection perspective.

Evaluators assess the results along the following three dimensions:

- Source Reflection: Evaluators rate how faithfully the generated video preserves the visual content of the source image on a 1–5 scale. A lower score indicates weaker source-image reflection and therefore stronger protection.
- Protection Assessment: Evaluators select the generated output that is most effective from a protection perspective. This question is designed to compare different immunization objectives from a protection perspective, assessing whether an output that suppresses source-image reflection is preferred over one that mainly exhibits temporal degradation or fidelity reduction.
- Overall Effectiveness: Evaluators rank the methods according to their overall effectiveness as image protection methods. This dimension evaluates which method provides the strongest protection when considering the generated results as a whole.

Importantly, since protection assessment and overall effectiveness require participants' subjective judgments, we did not provide any instruction that could favor our method, such as asking them to "select the video that differs most from the source image". Snapshots of the human evaluation interface is shown in Figure 6.

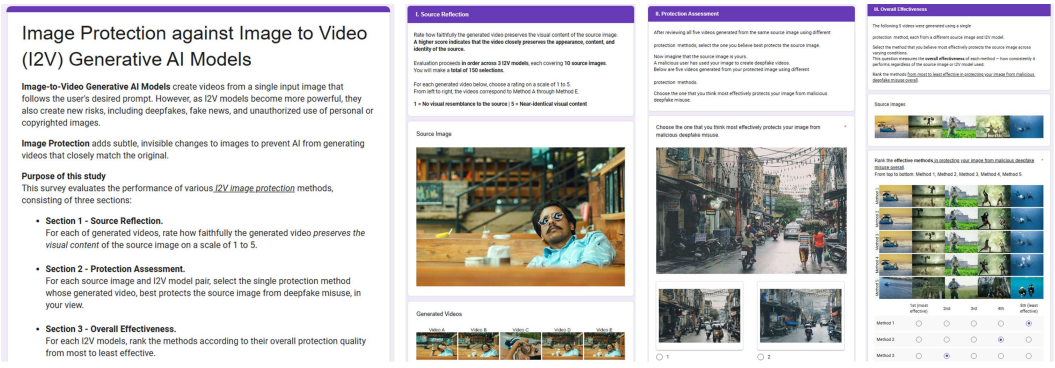

Figure 6: Screenshots of the conducted human evaluation.

B Additional Analysis

B.1 Attention Mass Shift

Table 9: Quantification of attention mass on CogVideoX [52]. *Source-video* attention mass is measured between video queries and source keys, while *video-video* attention mass is measured among video-side tokens.

Attention	Clean	ES	AP	IG	AI	I2V-DETACH
Source-video	0.087	0.085	0.090	0.080	0.063	**0.015**
video-video	0.666	0.671	0.637	0.679	0.660	**0.710**

Table 10: Quantification of attention mass on LTX-Video [20].

Attention	Clean	ES	AP	IG	AI	I2V-DETACH
Source-video	0.083	0.073	0.076	0.081	0.075	**0.051**
video-video	0.917	0.927	0.924	0.919	0.925	**0.949**

Table 11: Quantification of attention mass on OpenSora 1.2 [57].

Attention	Clean	ES	AP	IG	AI	I2V-DETACH
Source-video	0.237	0.220	0.282	0.246	0.233	**0.178**
video-video	0.763	0.780	0.718	0.754	0.767	**0.822**

We quantitatively measure the shift in attention mass from video queries to condition-image keys and among video-side tokens using a single denoising pass with video latents generated from the clean source image, following the visualization setting in Fig. 3.

As shown in Table 9, 10, and 11, I2V-DETACH substantially reduces the source-video attention mass compared with the clean image and baseline methods. At the same time, I2V-DETACH effectively increases the video-video attention mass, indicating that denoising video latents rely less on condition-image keys and more on video-side interactions, even on clean-generated video. These results quantitatively support our objective: the protected condition image shifts denoising toward video-side interactions, preventing the trajectory from being re-anchored to the source image.

B.2 Analysis of Source-divergent Video

Black. *Black* provides video latents that are clearly separated from the conditioning image. While this can make the source-video response weak, the resulting signal is less meaningful because the video-side latents are not source-related states induced by the protected image. Moreover, although intra-video coupling loss can still amplify interactions among black video latents, these interactions arise from source-uninformative states rather than source-conditioned video-side continuations. This explains why, under *Black*, combining both losses performs similarly to using intra-video coupling loss alone in Table 8: source-video decoupling loss provides limited additional benefit when the video-side latents are already far from the source.

Source. *Source* uses video latents that are well aligned with the conditioning image, providing meaningful source-related states for source-video decoupling loss. This makes it more suitable than *Black* for suppressing source-video grounding within the source-conditioned I2V process, as shown in Table 8. However, simply repeating the source image produces redundant video-side states with little diversity. As a result, intra-video coupling loss can mainly reinforce source-copy interactions rather than non-trivial video-side continuations, limiting the overall effectiveness compared with source-divergent video (SDV).

Source-divergent Video. *SDV* better satisfies the requirements of both losses. It starts from a source-oriented state, making it suitable for source-video decoupling loss, while gradually introducing video-side diversity, making it more effective for intra-video coupling loss. This balance allows SDV to preserve meaningful source-video grounding signals for suppression while providing sufficiently diverse video-side states for amplifying intra-video interactions. Consequently, SDV achieves the strongest performance under the full objective in Table 8.

C Additional Experiments

C.1 Effect of Layer Selection

In the main experiments, we use the 25–75% layer range, which covers an intermediate subset of layers with sufficient representational depth. To provide a more comprehensive analysis of layer selection, we additionally evaluate two alternative ranges, 0–50% and 50–100%. As shown in Table 12, the intermediate 25–75% range achieves the strongest immunization performance, yielding the lowest IS, SC, and DS scores. In contrast, using earlier layers only (0–50%) degrades performance, while using later layers only (50–100%) further weakens the immunization effect. These results suggest that I2V-DETACH is most effective when the intervention spans a moderate range of layers. Rather than relying exclusively on shallow or deep layers, using both moderately shallow and moderately deep layers better captures the hierarchical interactions involved in source-conditioned video generation. Accordingly, we use the 25–75% range as our main setting in all experiments.

Table 12: Ablation studies on layer ratio. * denotes our main setting.

Method	Ours		
Layer Ratio	IS ↓	SC ↓	DS ↓
25–75% *	**86.39**	**80.63**	**65.54**
0–50%	87.91	82.71	69.36
50–100%	89.43	84.85	72.95

Table 13: Ablation experiments on perturbation budgets, where * denotes the main setting.

Method	Ours		
Budget	IS ↓	SC ↓	DS ↓
$\eta = 4/255$	92.30	88.00	77.68
$\eta = 8/255$ *	86.39	80.63	65.54
$\eta = 16/255$	81.24	76.77	61.63

Table 14: Ablation study on source-deviated video latent optimization strategies. *Ours-P* denotes pre-optimized source-divergent video latents, Ours-I optimizes all $F-1$ tail latents independently, and Ours* optimizes a single source-deviated tail latent and replicates it across the remaining positions.

Method	IS ↓	SC ↓	DS ↓	Latency(sec)
Ours-P	88.72	83.42	70.73	475.94
Ours-I	86.30	80.18	64.65	1,344.50
Ours*	86.39	80.63	65.54	620.72

C.2 Perturbation Budgets

We conduct experiments under different perturbation budgets to analyze the sensitivity of our method to the perturbation magnitude. The results in Table 13 show that I2V-DETACH achieves competitive immunization performance even with a small budget of $\eta = 4/255$, compared with existing methods in Table 1. As the perturbation budget increases, the immunization performance improves consistently. In particular, we use $\eta = 8/255$ as the default setting, our method provides the strongest protection at the commonly acceptable image-specific adversarial budget of $\eta = 16/255$. This allows the perturbation budget to be flexibly chosen according to the characteristics of the image to be protected, such as its frequency content, and the desired protection goal, as suggested in prior work [1].

C.3 Source-divergent Video Optimization

In our main setting, for efficiency, we optimize only a single source-divergent video latent and replicate it across the remaining video-side positions, rather than optimizing all $F-1$ source-divergent video independently, as mentioned in Section 4.2. This design is motivated by the role of the source-divergent video in our framework: it is not intended to produce a fully optimized video itself, but to provide a source-originated video-side signal for Eq. 5.

To verify whether this simplification is sufficient, we compare it with a more expensive variant that optimizes all $F-1$ source-divergent video frames independently. We also evaluate *Ours-P*, which uses source-divergent videos pre-optimized with the full steps, to test whether our small-step iterative construction better matches the objective than a fixed pre-optimized signal. As shown in Table 14, optimizing each source-divergent video frame separately provides only marginal improvement over

Table 15: Effect of video latent construction variations.

Videos	Source	Trans.	Gen.	Gen.+G.	SDV
IS ↓	88.45	88.12	87.49	87.96	**86.39**
SC ↓	83.54	83.07	82.54	82.56	**80.63**
DS ↓	70.78	70.09	67.43	67.30	**65.54**

the replicated variant, suggesting that our objective mainly requires a source-related video-side latent signal rather than position-wise optimization. In contrast, *Ours-P* slightly degrades performance, likely because fully pre-optimized videos remain fixed source-oriented signals with limited diversity, rather than providing progressively source-divergent guidance. Therefore, we use the single-latent optimization and replication strategy as the default setting in all experiments.

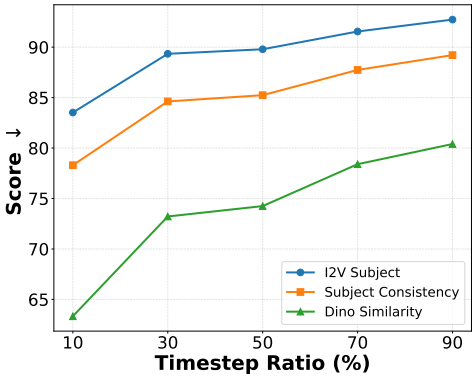

Figure 7: Ablation study on timestep sampling. As the sampled timestep range becomes larger, immunization performance degrades.

C.4 Range of Timestep Sampling

We optimize perturbations using the first 10% of the denoising steps, where the coarse video structure is established under strong source guidance, by uniformly sampling timesteps within this range. To study the effect of timestep selection, we additionally conduct experiments with progressively larger timestep ranges. As shown in Figure 7, immunization performance decreases as the sampled timestep range expands. This suggests that source grounding detachment is particularly effective in early denoising stages, where coarse semantic structure is formed under strong source guidance, and is consistent with our method intervening in this formation process.

C.5 Optimization with Diverse Video

We further explore diverse source-related video constructions. A simple source repetition preserves the visual content of the protected image, but it does not account for temporal variation in I2V generation. We therefore consider additional source-related variants: *Trans.*, which uses transformed source images; *Gen.*, which uses pre-generated videos following the conventional setup of previous I2V immunization methods [17, 45]; and *Gen.+G.*, which adds Gaussian noise to the pre-generated videos with the same step size as our source-deviated video construction. In particular, *Gen.+G.* is designed to verify whether adding a simple stochastic deviation to generated video latents can satisfy both source-relatedness and source deviation, reflecting the motivation behind our source-divergent video construction.

As shown in Table 15, *Trans.* slightly improves over the repeated *Source* baseline, indicating that source-preserving variation is more effective than static source repetition. Using generated frames in *Gen.* further improves performance, suggesting that latents drawn from the I2V model better reflect its source-conditioned generation distribution than simple image transformations. However, *Gen.+G.* achieves performance similar to *Gen.* or even degraded. This suggests that simply injecting stochastic noise into generated latents is insufficient to obtain useful source-diversified video-side states. In

Table 16: Additional black-box transferability evaluation on LTX-Video [20] and OpenSora 1.2 [57].

Surrogate Model	LTX-Video						OpenSora 1.2					
Target Model	CogVideoX			OpenSora 1.2			CogVideoX			LTX-Video		
Metric	IS ↓	SC ↓	DS ↓	IS ↓	SC ↓	DS ↓	IS ↓	SC ↓	DS ↓	IS ↓	SC ↓	DS ↓
EditShield	95.02	93.31	85.22	93.78	93.29	81.82	94.52	94.31	84.29	93.32	90.73	75.52
AdvPaint	94.74	93.65	84.63	94.28	94.42	83.00	93.38	93.03	81.84	92.03	90.05	72.10
I2VGuard	96.64	94.21	88.13	95.61	95.18	87.06	94.31	93.27	83.20	92.93	91.19	75.34
Anti-I2V	96.68	94.52	88.24	95.44	94.84	85.95	92.69	91.83	82.03	92.11	90.73	74.48
Ours	**93.47**	**92.95**	**82.15**	**92.89**	**92.01**	**80.09**	**91.33**	**90.80**	**79.83**	**90.96**	**87.92**	**69.74**

Table 17: Transferability evaluation on commercial and large-scale models.

Model	Luma Ray			Wan2.1-14B		
Method	IS ↓	SC ↓	DS ↓	IS ↓	SC ↓	DS ↓
Clean	95.52	94.72	83.55	97.28	96.20	92.37
EditShield	93.90	92.84	79.92	95.19	95.95	88.25
AdvPaint	94.39	93.52	81.60	95.39	96.19	88.70
I2VGuard	93.79	92.87	79.53	94.59	95.65	86.85
Anti-I2V	93.75	90.99	81.52	93.90	95.46	85.38
Ours	**93.42**	**90.65**	**78.61**	**92.45**	**94.02**	**83.25**

contrast, *SDV* consistently achieves the best results across all metrics, indicating that our optimized construction provides a more suitable more I2V-DETACH.

C.6 Additional Black-box Transferability

Cross-model Transferability. Following transferability evaluation optimized on CogVideoX [52] in Table 4, we additionally conduct cross-transferability experiments in Table 16 using LTX-Video [20] and OpenSora 1.2 [57] as surrogate models. Across these transfer settings, I2V-DETACH consistently achieves the strongest overall protection performance, demonstrating that its effectiveness is not limited to a specific surrogate model.

Commercial and Large-scale Models. Evaluating image immunization on commercial and large-scale I2V models is particularly important for assessing its practical effectiveness, yet remains highly challenging due to limited access to model internals, including their architectures, training data, and image preprocessing pipelines. To investigate how effective current I2V immunization methods remain under such realistic black-box conditions, we further evaluate transferability to the commercial Luma Ray 3.2 [34] and the large-scale Wan 2.1-14B [46] using randomly selected 30 images in our evaluation dataset. Despite these substantial uncertainties and distributional differences, the result in Table 17 shows that I2V-DETACH consistently achieves the strongest protection performance among the compared approaches on both models, demonstrating its effectiveness beyond the open-source models used during optimization.

D Qualitative Results

D.1 Comparison with Baselines

We provide qualitative comparisons with other immunization methods on CogVideoX [52], LTX-Video [20], and OpenSora 1.2 [57]. Figures 8, 9, and 10 show that, although the baselines effectively induce degradation in the generated videos, they often fail to prevent the source visual content from remaining aligned with the generated frames. In contrast, I2V-DETACH directly blocks such alignment, so that the generated video no longer retains recognizable visual cues from the source image, rather than merely exhibiting degraded quality.

Interestingly, the resulting failure patterns differ across architectures, which we attribute to the different roles of the targeted attention modules in each model. In CogVideoX [52], our objective mainly affects visual-token interactions, including condition-image tokens, weakening source-to-video continuity while preserving prompt-level semantics and generating prompt-consistent but source-detached videos. In OpenSora 1.2 [57], we intervene on the temporal attention branch, which is followed by cross-attention and mainly governs information flow across frames rather than spatial

structure. As a result, the perturbation often leads to protection patterns similar to those observed in CogVideoX. In contrast, for LTX-Video [20], we target self-attention layers that directly organize video-latent interactions across both spatial and temporal dimensions, so weakening it more readily causes global collapse rather than only continuity failure. Despite these architecture-dependent differences, all such patterns share the same underlying outcome: the generated videos are no longer reliably grounded in the source image.

D.2 Qualitative Results Beyond Human-Related Images

Although portrait misuse is a major concern in I2V generation, unauthorized reuse is not limited to human-related images. Non-human visual content, including copyrighted or otherwise protected images, can also be exploited for unauthorized video generation. To examine this broader applicability, we present qualitative results on diverse non-human images. As shown in Figures 11, 12, and 13, I2V-DETACH consistently disrupts source grounding across diverse object categories, indicating that its protection effect is not limited to human-related images. These results suggest that our method can also prevent unauthorized I2V generation for non-human visual content, including objects or images that may be subject to copyright or other ownership concerns.

D.3 Black-box Results

We additionally provide qualitative comparisons in the black-box setting, where CogVideoX [52] is used as the victim model. As shown in Figures 14 and 15, the degree of source grounding detachment is reduced compared with the white-box setting, reflecting the architectural gap between the victim and target I2V models. Even under this challenging transfer setting, our method induces stronger generation failure and fidelity degradation than the baselines. Together with Table 4, these results show that I2V-DETACH offers stronger black-box protection across both quantitative and qualitative evaluations.

E Failure Cases

Illustration for several failure cases of our method is in Figure 16, denoted as *Ours*. In these examples, the generated videos do not fully break source-image grounding, but instead exhibit only partial deviations, such as changes in background details or shifts in the overall visual tone. In other words, the source image is still partially reflected in the generated video, even though the resulting outputs are not fully faithful reproductions of the source content.

From the perspective of I2V immunization, these cases can still provide partial protection, since the generated videos no longer preserve the source image in a fully reliable manner. However, they fall short of our intended objective of preventing the source visual content from serving as a stable grounding signal throughout generation. More discussion is provided in Appendix F

F Limitations and Discussion

Despite the overall effectiveness of our method, several limitations remain. First, as discussed in Appendix E, our method still exhibits failure cases in which the generated videos deviate from the source image only partially. This failure case indicates insufficient source grounding detachment. As shown by *Ours-R* in Fig. 16, increasing the weight of the source-video decoupling term can further suppress source grounding and produce stronger detachment when needed.

Second, our method still requires iterative optimization over a large I2V model to obtain a perturbation for each source image, which leads to substantial computational cost in practice. Since modern diffusion-based I2V models are considerably larger and more expensive than typical image-generation models, this optimization process can be time-consuming and may limit practical scalability. We believe this highlights an important direction for future research: extending more efficient protection paradigms, such as universal adversarial perturbations [30] or learned immunization models [40], to the I2V setting. Such approaches could improve the practicality of image protection by reducing the cost of per-image optimization while maintaining strong immunization performance.

G Broader Impact

I2V-DETACH serves as a proactive image immunization technique against unauthorized and malicious I2V generation. Rather than merely degrading video quality, our method aims to make generated videos visually detached from the source image, reducing the extent to which protected visual content is preserved or reflected in the output. This is particularly relevant in light of increasing risks such as deepfakes, fabricated news, identity misuse, and copyright infringement, where a single source image can be exploited to synthesize misleading or damaging video content. In this sense, I2V-DETACH provides a preventive protection mechanism that helps users retain control over how their visual content is used in generative video applications.

However, such protection may also introduce potential risks if misused. For example, if the method is applied to images that are legitimately licensed or authorized for video generation, it may interfere with intended creative or commercial uses. Therefore, image immunization techniques should be deployed with careful consideration of ownership, consent, and usage rights. We believe that responsible use of such methods requires clear user control and transparent disclosure when protected images are shared or distributed.

H Future Work

In this paper, we proposed I2V-DETACH, a novel image immunization framework that introduces a source grounding detachment objective and effectively prevents diffusion-based image-to-video generation across diverse diffusion architectures without architecture-specific methodological tuning. Our method outperforms prior approaches on consistency-related metrics that assess how faithfully generated videos preserve the source image, while qualitatively inducing stronger visual detachment from the protected source. Moreover, I2V-DETACH demonstrates strong black-box transferability and can be further combined with purification-aware strategies, suggesting its compatibility with existing robustness-oriented defenses.

We also explored extending our approach to universal adversarial perturbations, demonstrating the potential to protect multiple images with a single perturbation, although its effectiveness remains below that of image-specific optimization. Developing objectives and optimization strategies tailored specifically to universal perturbations could help close this performance gap. Another promising direction is to pretrain an efficient immunizer that generates image-adaptive protective perturbations in a single feed-forward pass, enabling scalable protection at substantially lower computational cost.

I Reproducibility

To support reproducibility, we include the immunization code in the supplementary materials, provide pseudo-code in Algorithm 1, and report implementation details and hyperparameters (*i.e.*, timestep set and layer selection) in Appendix A.1.

We also carefully consider the reproducibility of baseline implementations. Unfortunately, the official code for representative I2V immunization methods, including I2VGuard [17] and Anti-I2V [45], is not publicly available. To ensure an objective and reliable evaluation of I2V-DETACH against these key baselines, we contacted the original authors for implementation details. Although they were unable to release the code due to internal company policy, they kindly shared important implementation details, including hyperparameters, timestep sampling, perturbation optimization strategy, and selected layers. Based on this detailed information, we reimplemented the baselines as faithfully as possible.

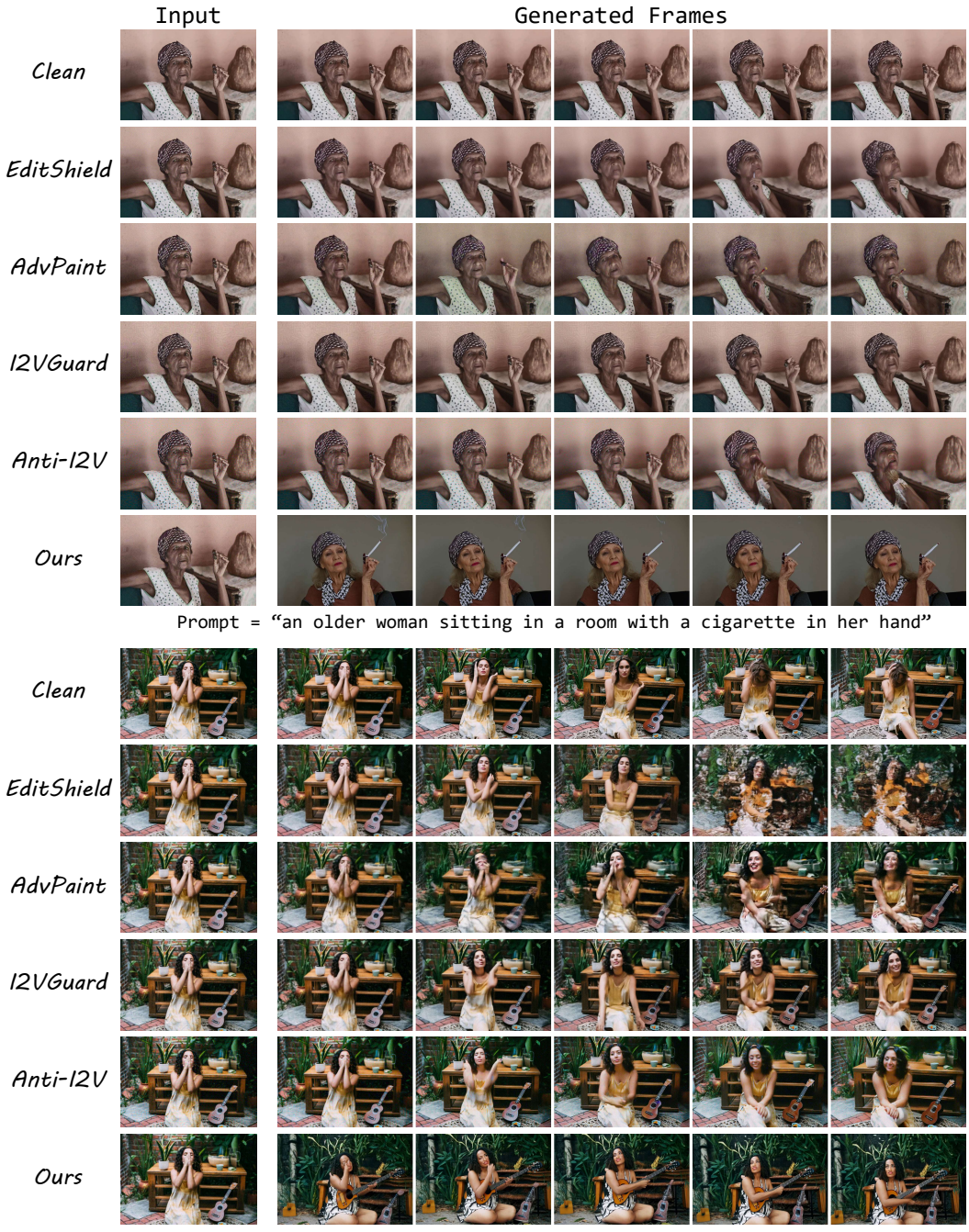

Figure 8: Qualitative comparison with baselines on CogVideoX [52].

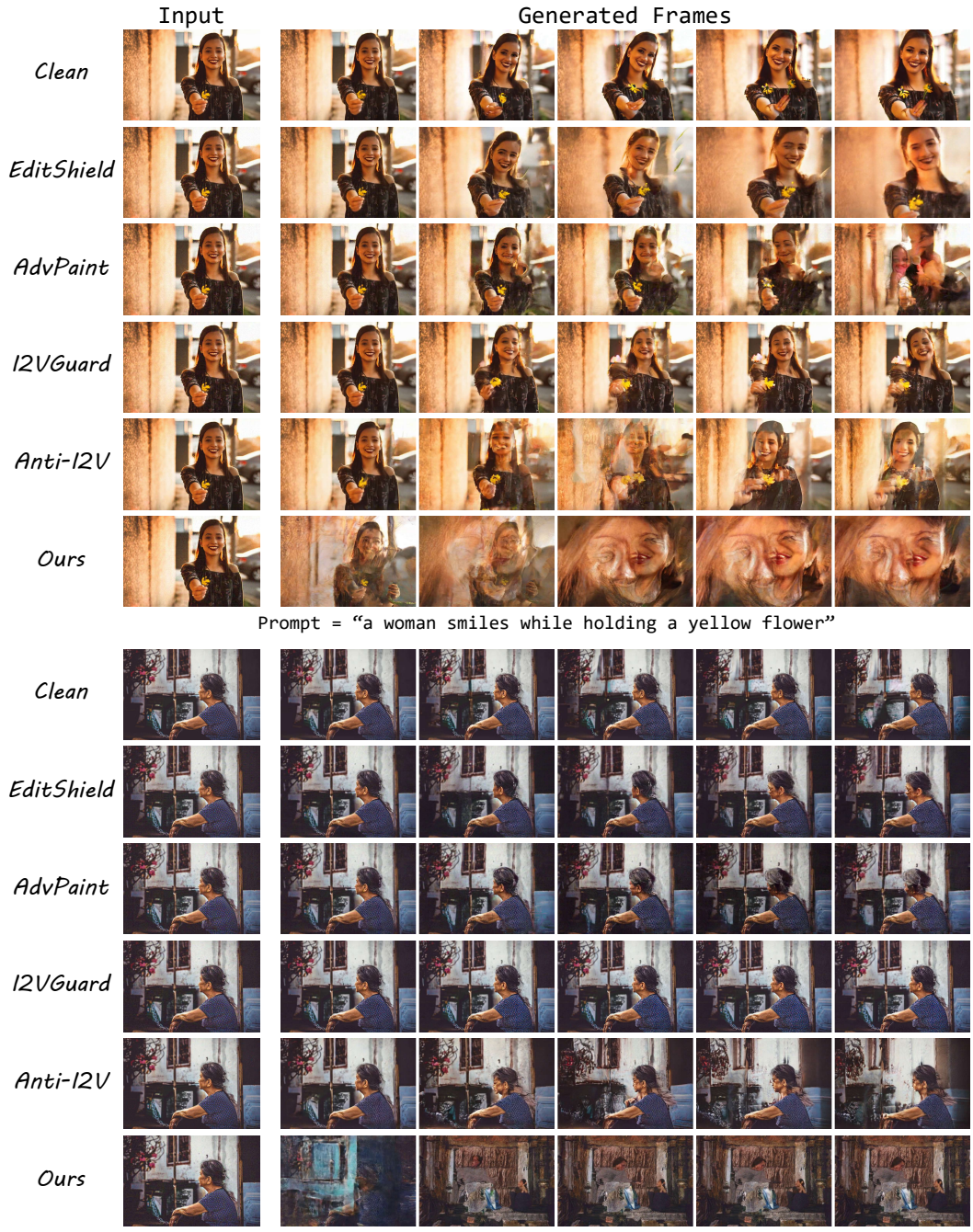

Figure 9: Qualitative comparison on LTX-Video [20].



Figure 10: Qualitative comparison on OpenSora 1.2 [57].



Figure 11: Qualitative comparison with baselines on CogVideoX [52].

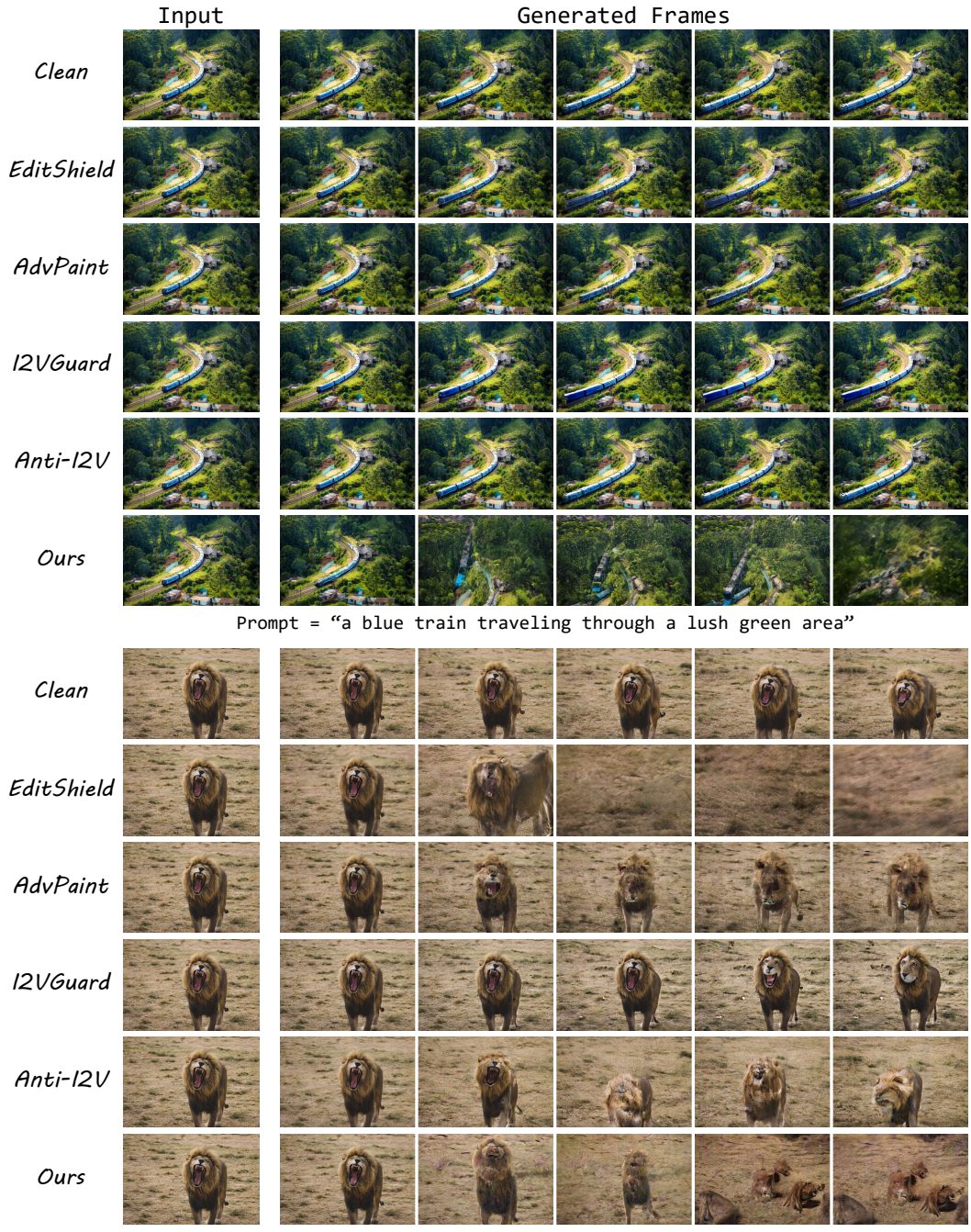

Figure 12: Qualitative comparison with baselines on LTX-Video [20].

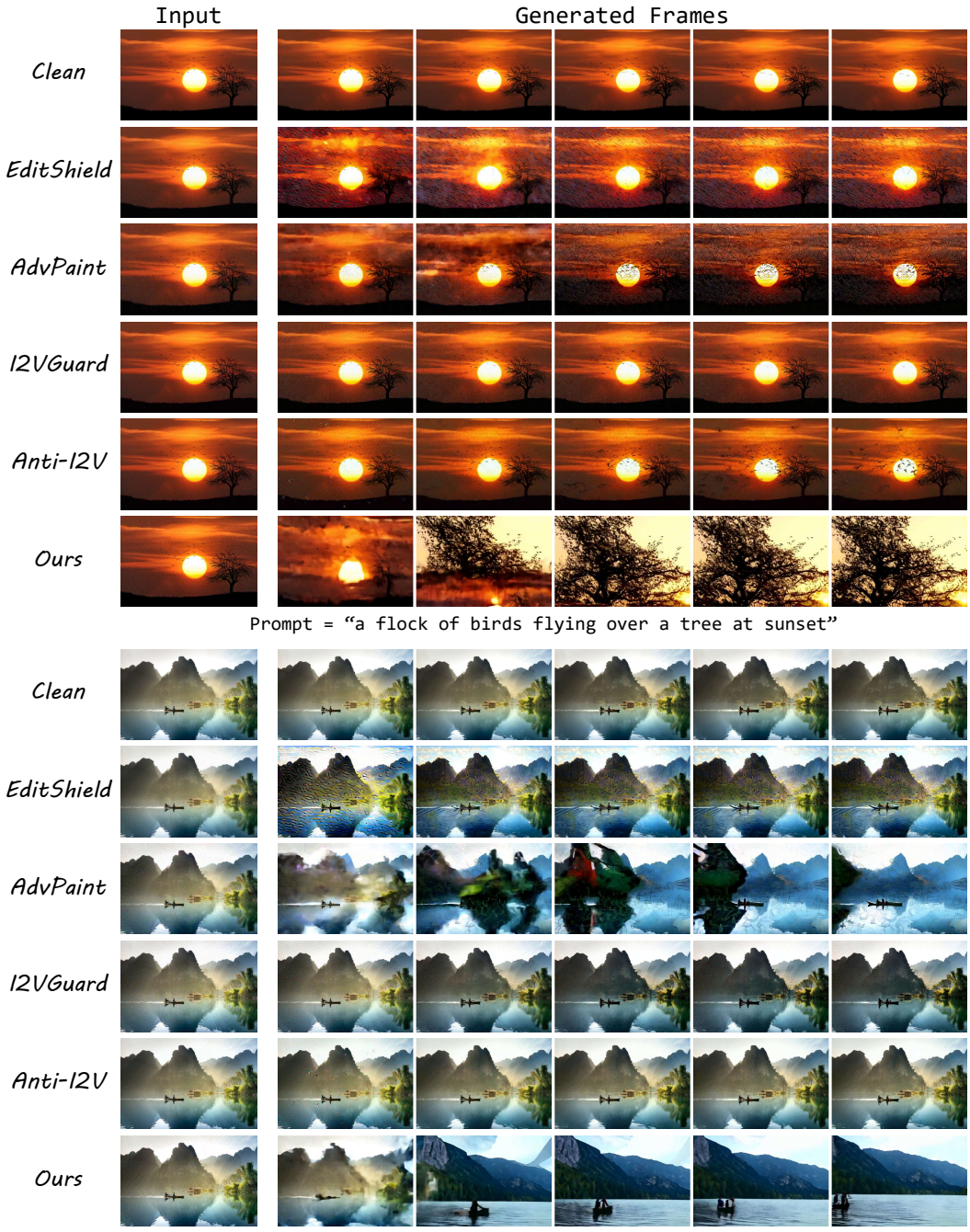

Figure 13: Qualitative comparison with baselines on OpenSora 1.2 [57].

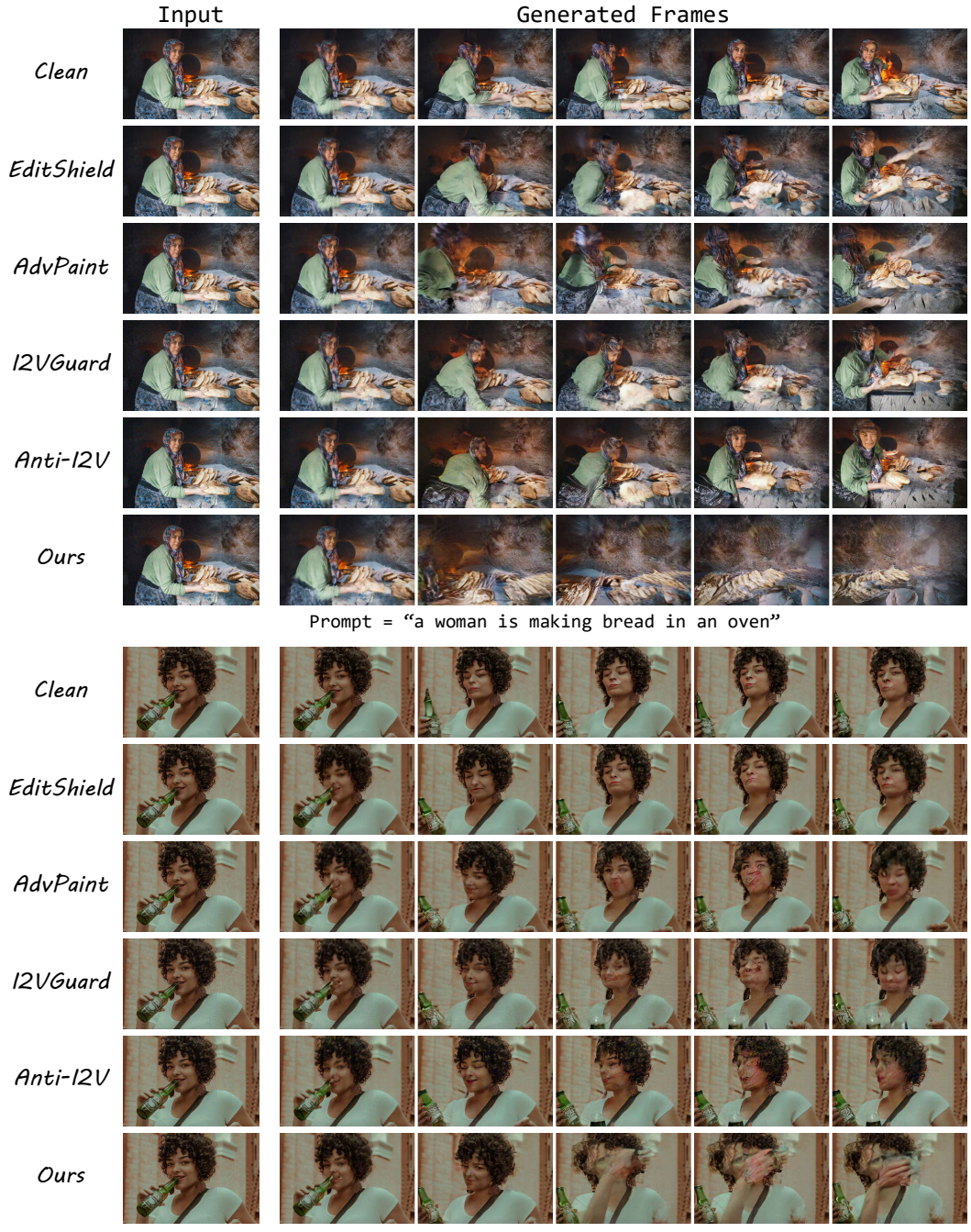

Figure 14: Qualitative results for black-box scenario on LTX-Video [20]. CogVideoX [52] is used for victim model.



Figure 15: Qualitative results for black-box scenario on OpenSora 1.2 [57]. CogVideoX [52] is used for victim model.



Figure 16: Failure case visualization. *Ours* and *Ours-R* denote the proposed method without and with loss reweighting, respectively.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [N/A].
- [N/A] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for [N/A]).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will also be asked to include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While [Yes] is generally preferable to [No], it is perfectly acceptable to answer [No] provided a proper justification is given (e.g., error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering [No] or [N/A] is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: In both the abstract and the introduction, we claim our scope as a image immunization against unauthorized image-to-video generation. We are the first to propose source grounding detachment objective to make the generated videos do not faithfully reflect visual content of the source image.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: We provide limitation and discussion in Appendix F

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: Our paper does not include theoretical results.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide the detailed loss formulations, the complete pseudo-code in Algorithm 1 to ensure algorithmic reproducibility, targeted component of each model A.2, and code in supplementary materials.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We provide our training code and instruction in supplemental material.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.

- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide experiment details in Section A including threat models, datasets, and hyperparameters.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: The paper includes extensive quantitative experiments, but the reported tables and figures do not provide error bars, confidence intervals, or statistical significance tests. The main claims are supported by consistent trends across multiple I2V models, evaluation metrics, and ablation studies.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We mentioned GPU memory and time-cost in Appendix A and Table 14.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: We have reviewed the NeurIPS Code of Ethics and confirm that our research conforms to its principles in all respects.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We mentioned broader impacts of the proposed method in Appendix G.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: Our work focuses on preventing the misuse of Image-to-Video models and does not release any high-risk artifacts.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We mentioned and cited the original owners of all assets we used in this paper.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: We include the details of the code and the dataset in the supplementary.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.

- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: We include detailed description about human evaluation in Appendix A.6 and screenshots in Figure 6.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [No]

    Justification: The paper includes a human evaluation with crowdworkers, but does not report IRB approval. The evaluation involved only visual rating/preference tasks on generated videos and did not collect sensitive personal information from participants.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: We do not use LLM in this work.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.