IMMUNE2V: IMAGE IMMUNIZATION AGAINST DUAL-STREAM IMAGE-TO-VIDEO GENERATION

Anonymous authors
Paper under double-blind review

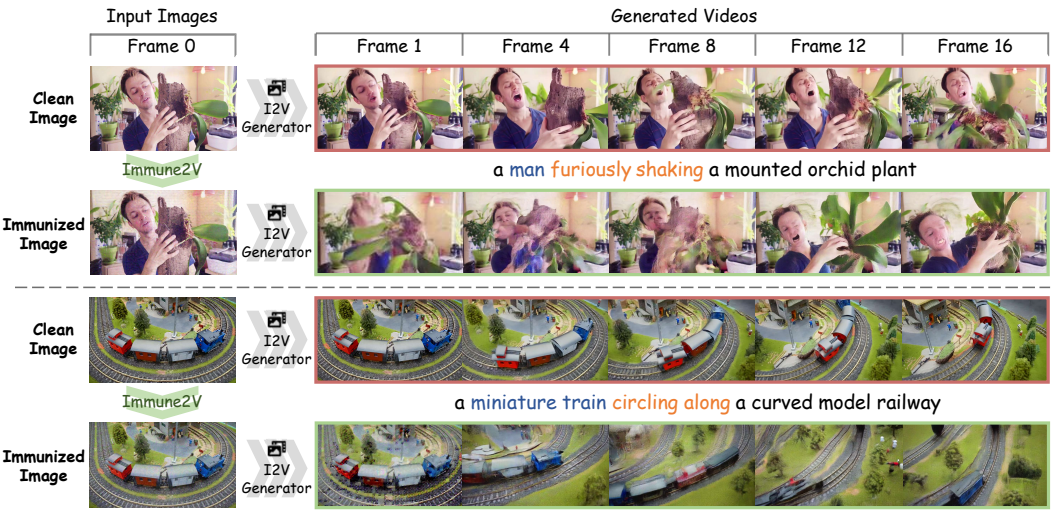


Figure 1: ***Immune2V Protection.*** Immune2V protects images from Image-to-Video (I2V) generation by adding imperceptible immunization noise to produce an immunized image. When processed by an I2V generator, a clean image yields realistic motion (e.g., the human face articulates naturally and the train moves circularly along the track, top row), whereas the immunized image disrupts the generation dynamics, producing highly unrealistic videos (e.g., the human subject is severely distorted and the train structurally collapses, failing to synthesize coherent motion, bottom row).

ABSTRACT

Image-to-video (I2V) generation has the potential for societal harm because it enables the unauthorized animation of static images to create realistic deepfakes. Recent I2V immunization methods protect reference images by introducing imperceptible adversarial perturbations, but remain fragmented across specific architectures and conditioning mechanisms. In this paper, we systematically analyze how reference images participate in modern I2V generation and identify a common dual-stream conditioning paradigm: a *structural stream* that provides a visual anchor, and a *semantic stream* that interacts with other semantic conditions (e.g. text prompt) throughout the generation process. Motivated by the need to jointly disrupt these two streams, we propose **Immune2V** framework which maximizes the deviation between perturbed and clean structural conditioning signals across distinct conditioning mechanisms, while aligning intermediate generative representations with collapse-inducing trajectories across diverse plausible prompts to counteract semantic guidance. Experiments demonstrate that Immune2V provides stronger and more persistent protection than existing baselines, consistently disrupting I2V generation.

1 INTRODUCTION

While the rapid advancement of generative models (Ho et al., 2020; Blattmann et al., 2023a; Ho et al., 2022; Liu et al., 2024b) enables the synthesis of highly realistic content, it also introduces

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

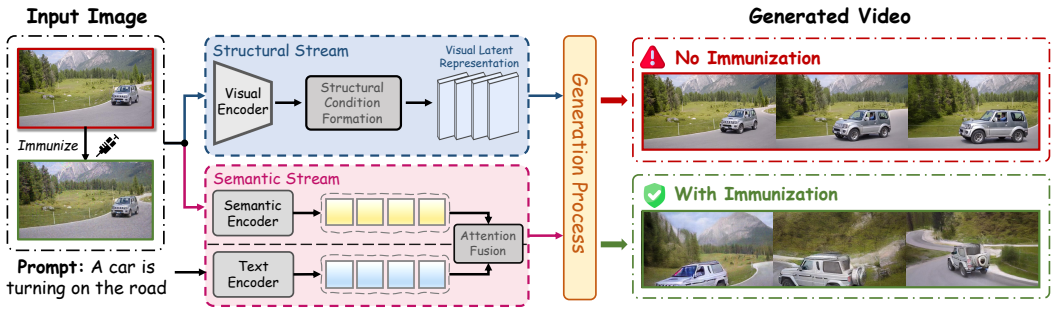


Figure 2: ***Immunization on Dual-Stream I2V Architecture.*** A clean input image guides the generation process through i) a **structural stream**, where the image is encoded and transformed into visual latent representation that structurally conditions video generation, and ii) a **semantic stream**, where image- and text-derived semantic representations provide attention-based guidance. Injecting an adversarial perturbation yields an immunized image that simultaneously disrupts both conditioning streams, forcing the generation process toward structural collapse.

the potential for substantial harm by allowing malicious actors to create convincing deepfakes. A particularly concerning threat is Image-to-Video (I2V) generation, which enables the unauthorized animation of static personal images harvested online into fabricated videos via semantic conditioning (e.g. text prompts) (Kong et al., 2024; Wan et al., 2025; Liu et al., 2024b; Yang et al., 2025). Although significant efforts have been made in deepfake detection (Rana et al., 2022; Rossler et al., 2019; Li et al., 2020; Wang et al., 2023), these methods are inherently reactive. An alternative proactive approach uses adversarial perturbation methods in discriminative vision tasks (Goodfellow et al., 2015; Kurakin et al., 2018; Arnab et al., 2018; Jia et al., 2025) and image diffusion models (Liang et al., 2023; Salman et al., 2023; Liang & Wu, 2023; Xue et al., 2024; Lo et al., 2024; Liu et al., 2024a; Shan et al., 2023; Ozden et al., 2026). This approach adopts the input-level strategy of *immunization*, which injects an optimized perturbation directly into the image before it is shared online. Ideally, if an adversary subsequently attempts to animate such an immunized image, the embedded perturbation would disrupt the generative model's internal representations, causing the synthesized video to become distorted (Fig. 2).

Based on the success of image immunization, a natural question is: ***Why not naively extend existing immunization methods to the video domain?*** While this might appear straightforward, the I2V setting introduces fundamental challenges. First, because I2V models receive strong conditioning via the input image that anchors content generation, an attack cannot simply drive the output toward trivial or unrelated results (e.g., blank frames), but must instead induce generation failures under the constraint of the conditioning signal. Second, analysis of the architectures of modern I2V models reveals an even deeper challenge: State-of-the-art systems adopt a "dual-stream" image conditioning architecture (Fig. 2), in which a single input image provides both structural and semantic conditioning. This makes a direct transfer of image-level immunization objectives inherently insufficient: perturbing the input image alone does not explicitly account for how its information is encoded and propagated through these two distinct streams. Some recent I2V protection methods attempt to address this challenge by designing attack objectives over particular internal representations or architectural components, including temporal attention, selected denoising layers, and text-conditioned interactions (Gui et al., 2025; Vu et al., 2026; Song et al., 2026). However, these methods remain largely tied to specific components of the generation architecture, without explicitly modeling the distinct roles of structural and semantic reference-image conditioning.

In the structural stream (Fig. 2, blue box), the reference image is encoded into visual latent representations that are temporally organized and incorporated into the iterative generation process through mechanisms such as zero-padded encoding (Wan et al., 2025), replicated concatenation (Xing et al., 2024a), or first-frame token replacement (Kong et al., 2024). These representations serve as structural anchors that preserve the appearance and spatial layout of the reference image throughout video generation. In the semantic conditioning stream (Fig. 2, pink box), the reference image is transformed into higher-level semantic representations through visual or multimodal encoders (Radford et al., 2021). These image-derived features then interact with text prompt guidance during denoising, jointly steering the semantic content and motion of the generated video. We show that when prompts are *aligned* with the initial image (i.e., the text prompt matches the image content),

the semantic guidance is reinforced and demonstrates a strong "healing" effect—even if the structural stream is compromised via naive immunization, the continuous semantic pull of a coherent prompt actively overrides the visual corruption and forces the trajectory back to a consistent structure. These two pathways impose different constraints on adversarial protection and motivate the corresponding structural and semantic attacks in our framework.

To this end, we introduce **Immune2V**, a unified dual-stream joint optimization framework that explicitly targets both the structural and semantic streams of I2V models. First, we design a **Structural Attack** that maximizes the divergence between clean and adversarial structural representations, with architecture-specific objectives that account for how the reference image is encoded and injected into the generation process. For temporally distributed representations, we further balance the disruption across temporal latent positions to prevent the adversarial effect from attenuating over time. Second, we design a **Semantic Attack** that draws on empirical distributions of aligned and collapse-inducing prompts. Each collapse-inducing prompt provides a precomputed sampling trajectory as a pseudo-target. We then optimize the adversarial image such that, even when paired with a sampled aligned prompt, its intermediate generative representations are pulled toward the corresponding collapse trajectory. Consequently, during inference time, when an unauthorized user provides a normal prompt, our method hijacks the iterative guidance and forces the video into a state of structural breakdown. By jointly optimizing these two attacks into a single immunization, Immune2V disrupts both forms of reference-image conditioning, inducing severe and long-lasting degradation across the entire generated video. In validating the benefit of Immune2V, we conduct extensive experiments and establish comprehensive quantitative metrics for video quality and consistency. Our main contributions are:

- A systematic analysis of reference-image conditioning in modern I2V models, identifying a common **dual-stream paradigm** and revealing the limitations of direct transferring the image-level immunization.
- **Immune2V**, a unified dual-stream immunization framework that disrupts architecture-specific structural conditioning and steers prompt-conditioned representations toward collapse-inducing trajectories.
- Extensive video-level evaluations demonstrating stronger and more persistent protection than existing baselines.

2 Related Work

Image-To-Video Generation Diffusion- (Ho et al., 2020; Singer et al., 2023; Blattmann et al., 2023b; Bar-Tal et al., 2024; Ho et al., 2022) and flow-based (Liu et al., 2023; Wan et al., 2025; Chen et al., 2025; Jin et al., 2025) text-to-video models synthesize high-quality videos. Building upon these, image-based conditioning guides synthesis via semantic guidance (Xing et al., 2024a; Zhang et al., 2023), visual consistency (Ren et al., 2024; Hu, 2024), and explicit structural control (Yu et al., 2023; Guo et al., 2024; Peng et al., 2024). A widely adopted architecture in I2V generation models—such as Wan 2.1(Wan et al., 2025), HunyuanVideo-I2V (Kong et al., 2024), I2VGen-XL (Zhang et al., 2023), DynamiCrafter (Xing et al., 2024a)—is dual-stream architecture. Specifically, to improve scalability and generalization, Wan 2.1 adopts a rectified-flow formulation and large-scale unified pre-training, while HunyuanVideo-I2V scales a DiT-based video foundation model beyond 13B parameters through progressive training. However, the impact of input-level perturbations on such dual-stream pipelines is underexplored.

Adversarial Attacks on Generative Diffusion Models Adversarial attacks have been extensively studied in diffusion-based systems (e.g., image generation (Salman et al., 2023; Liang et al., 2023; Liang & Wu, 2023; Shen et al., 2025; Xue et al., 2024; Xue & Chen, 2024; Liu et al., 2024a; Ahn et al., 2026), video generation (Gui et al., 2025; Li et al., 2024; 2025; Vu et al., 2026) and policy generation (Chen et al., 2024; Kalra et al., 2025)). For image diffusion, methods like PhotoGuard (Salman et al., 2023), DiffVax (Ozden et al., 2026), Mist (Liang & Wu, 2023) and SDS-Attack (Xue et al., 2024) design perturbations to disrupt encoders and diffusion dynamics, while DAYN (Lo et al., 2024) and DiffusionGuard (Choi et al., 2025) target cross-attention and localized masks. However, extending this success to video diffusion is non-trivial. Some existing studies focus on offensive jailbreaks of video generation models (Li et al., 2025; Wang et al., 2026). For works that study immunization (for I2V generation), I2VGuard (Gui et al., 2025) shows improvements mainly on

SVD (Blattmann et al., 2023a), with only marginal gains on CogVideoX (Yang et al., 2025). More recent defenses further explore model-specific internal representations: Anti-I2V (Vu et al., 2026) combines color- and frequency-domain perturbations with layer-wise objectives targeting VAE and denoiser representations, while Anti-Prompt (Song et al., 2026) suppresses text-conditioned interactions and amplifies visual-only pathways across attention architectures to improve robustness to unseen prompts. These approaches mostly attack the old UNet denoising backbone, and largely target selected internal representations or conditioning interactions, leaving a unified treatment of how reference images condition I2V generation unresolved.

3 MOTIVATION

3.1 PROBLEM SETTING: INPUT-IMAGE IMMUNIZATION FOR I2V MODELS

Image-to-Video (I2V) generation synthesizes a video $\mathcal{V} = G(\cdot\,; I, c)$, using a generative model G conditioned on a clean static image $I \in \mathbb{R}^{C\times H\times W}$ and a semantic condition c (e.g. text prompt). To protect this content against unauthorized generation, input-level "immunization" embeds an imperceptible adversarial perturbation $\delta \in \mathbb{R}^{C\times H\times W}$ directly into the source image, yielding an immunized image $I_{\text{adv}} = I + \delta$. This embedded noise is specifically crafted to disrupt the model's generative process. We achieve this by optimizing the perturbation to minimize the distance between the video generated from the immunized image and a predefined, severely degraded target video $\mathcal{V}_{\text{target}}$ (e.g. a completely zeroed video). This defense is formulated as the following objective:

$$\delta^* = \arg\min_{\|\delta\|_\infty \le \epsilon} \mathcal{L}\big(G(\cdot, ; I_{\text{adv}}, c), ; \mathcal{V}_{\text{target}}\big). \tag{1}$$

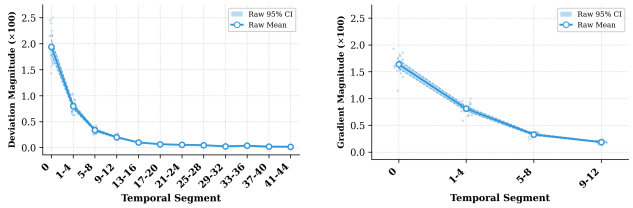
where the loss function $\mathcal{L}$ quantifies the distance between the synthesized output and the reference target $\mathcal{V}_{\text{target}}$. A successful defense must induce severe and persistent degradation across the entire generated video.

3.2 STRUCTURAL CONDITIONING

At the structural level, the reference image anchors the generated video to the source content. It preserves low- and mid-level visual information, including subject appearance, spatial layout, and initial-frame structure, thereby constraining the generation process to remain consistent with the reference image. Although this role is common across modern I2V models, its concrete implementation varies substantially across architectures. For example, Wan 2.1 (Wan et al., 2025) expands the reference image into a pseudo-video and encodes it into a sequence of spatio-temporal latents; HunyuanVideo-I2V (Kong et al., 2024) explicitly injects the encoded visual latents into the denoising process. Other architectures similarly incorporate reference information through latent concatenation, token replacement, cross-frame attention, or reference-guided noise initialization (Xing et al., 2024a; Blattmann et al., 2023a; Ren et al., 2024; Hu, 2024). We collectively denote these structural representations (prior to the iterative generation process) as $c^M_{\text{struct}}(I)$ for an I2V architecture M.

This architectural diversity makes the direct transfer of image-level immunization objectives non-trivial. Existing image attacks typically optimize perturbations against a particular image-level representation. However, such disruption may not survive the architecture-specific encoding and injection process or translate into a persistent deviation in $c^M_{\text{struct}}(I)$.

Wan (Wan et al., 2025) provides a representative example of this mismatch. Its structural stream temporally expands the reference image into a zero-padded pseudo-video $\tilde{I} =$



(a) ***Latent Deviation*** ($d_t = \|z_t^{\text{adv}} - z_t^{\text{clean}}\|_2$). (b) ***Gradient Magnitude*** ($g_t = \|\nabla_I d_t\|_2$).

Figure 3: ***Temporal Attenuation in Wan 2.1.*** Image-level attacks concentrate both forward perturbations and backward gradients in early temporal latents. Gradients are computed over fewer frames due to the high cost of backpropagating through the full VAE architecture

$[I, 0, \ldots, 0]$, which is encoded by a 3D causal VAE $\mathcal{E}_{\text{enc}}$:

$$z = \mathcal{E}_{\text{enc}}(\tilde{I}) = \{z_t\}_{t=1}^{T'}, \qquad z_t \in \mathbb{R}^{C' \times H' \times W'}. \tag{2}$$

The resulting structural condition is therefore distributed across multiple temporal latent slices. When an adversarial perturbation is applied only to the reference image, the corresponding input becomes $\tilde{I}_{\text{adv}} = [I + \delta, 0, \ldots, 0]$. We quantitatively show (in Fig. 3) that this adversarial influence is highly uneven across the temporal axis: early latent slices are substantially perturbed, whereas later slices exhibit much weaker deviations. Specifically, we measure the slice-wise latent deviation $d_t = \left\| z_t^{\text{adv}} - z_t^{\text{clean}} \right\|_2$ together with the corresponding gradient magnitude $g_t = \|\nabla_I d_t\|_2$. Both quantities decrease toward later temporal positions, indicating that the forward adversarial and backward optimization signal are concentrated in the earliest structural latents. Therefore, conventional encoder-level objectives can strongly perturb the initial structural condition while leaving later temporal representations less affected. This temporal attenuation motivates an architecture-aware structural attack that directly disrupts the conditioning representations used by the I2V model and, for Wan 2.1, explicitly balances the adversarial effect across the temporal latent sequence.

3.3 Semantic Conditioning in I2V Models

At the semantic level, the reference image provides guidance that shapes the semantic evolution of the generated video. Image-derived semantic representations are obtained through architecture-specific visual or multimodal encoders. During iterative generation, they are typically fused with other semantic conditions, mostly text prompt embeddings, through attention-based mechanisms and jointly injected into the denoising network (Peebles & Xie, 2023; Ronneberger et al., 2015). We denote this conditioning mechanism as $\text{Fuse}^M\left(\mathcal{E}_{\text{img}}^M(I), \mathcal{E}_{\text{text}}^M(c)\right)$.

This persistent semantic guidance introduces a complementary limitation for image-level immunization. Even when the structural condition has been successfully corrupted, the denoising process remains repeatedly guided by semantic information associated with the reference image and the text prompt. In particular, when an aligned prompt describes content and motion consistent with the source image, the resulting semantic guidance can act as a corrective signal, progressively steering the generation back toward a coherent trajectory.

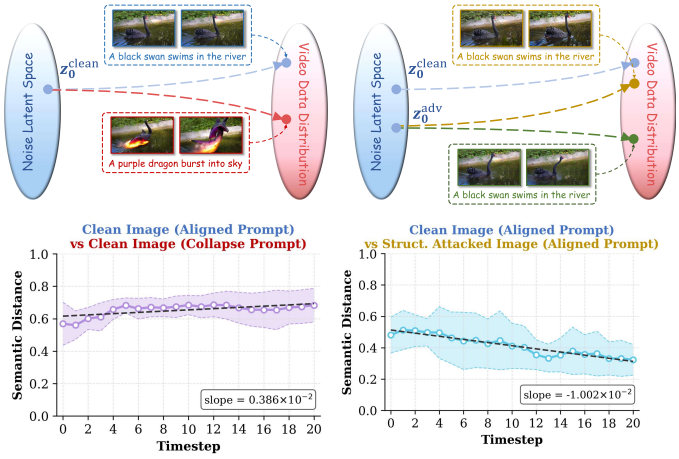


Figure 4: ***Semantic Conditioning Override.*** We analyze generative trajectories in the noise latent space by measuring semantic distance to a clean baseline (Clean Image (Aligned Prompt)) at each timestep. (a) A mismatched prompt (Clean Image (Collapse Prompt)) diverges from the baseline (positive slope); (b) A plausible prompt (Struct. Attacked. Image (Aligned Prompt)) pulls the generation back to normal (negative slope).

We analyze this behavior by comparing denoising trajectories under different image–prompt conditions in Fig. 4. Starting from the same reference image, a semantically conflicting (collapse-inducing) prompt drives the generation away from the clean aligned trajectory, demonstrating that semantic conditioning can substantially alter the evolution of the latent state. In contrast, when a structural attacked image is paired with an aligned prompt, the initially perturbed trajectory progressively approaches the clean trajectory as denoising proceeds (as examined quantitatively in Fig. 4). This *conditioning override* indicates that disrupting the structural stream alone does not prevent the semantic stream from restoring coherent generation.

Therefore, effective I2V immunization must jointly disrupt the structural representation and prevent semantic guidance from correcting the adversarial effect. These observations directly motivate the Structural and Semantic Attacks introduced in Sec. 4.

4 IMMUNE2V FRAMEWORK: DUAL-STREAM JOINT OPTIMIZATION

We propose **Immune2V**, a unified immunization framework designed to disrupt dual-stream I2V models by jointly optimizing two complementary attacks: i) a **Structural Attack** that maximizes the divergence between clean and adversarial structural conditions while accounting for architecture-specific conditioning mechanisms, and ii) a **Semantic Attack** that redirects condition-driven denoising toward collapse-inducing dynamics by manipulating intermediate generative representations.

4.1 STRUCTURAL ATTACK

Given a structural conditioning function $c^M_{\text{struct}}(\cdot)$, our objective is to maximize the discrepancy between the structural conditions induced by the clean and immunized images:

$$\mathcal{L}_{\text{Struct}} = -\mathcal{D}\left(c^M_{\text{struct}}(I_{\text{adv}}), c^M_{\text{struct}}(I)\right), \tag{3}$$

where $\mathcal{D}$ measures representation-level divergence. This formulation directly weakens the structural anchoring provided by the reference image, without the reliance of target images. For most I2V architectures, this objective can be instantiated directly on the reference representations already used by the model, such as encoded image latents or intermediate reference features (Blattmann et al., 2023a; Xing et al., 2024a; Ren et al., 2024). Adapting the structural attack typically only requires identifying the corresponding conditioning representation.

For Wan 2.1 (Wan et al., 2025), the temporal attenuation identified in Sec. 3.2 makes a direct aggregate objective insufficient. We therefore introduce a temporally-balanced objective that explicitly encourages strong deviation at every temporal position. For each latent slice, we measure the normalized deviation between the clean and adversarial representations:

$$d_t = \frac{1}{\sqrt{C'H'W'}} ||z_t^{\text{adv}} - z_t^{\text{clean}}||_F, \tag{4}$$

where $z_t \in \mathbb{R}^{C' \times H' \times W'}$ denotes the t-th latent slice and $||\cdot||_F$ denotes the Frobenius norm. We first maximize the average structural deviation across all temporal slices, and to further strengthen weakly disrupted temporal positions, we introduce a soft coverage objective:

$$\mathcal{J}_{\text{avg}} = \frac{1}{T'} \sum_{t=1}^{T'} d_t, \tag{5}$$

$$\mathcal{J}_{\text{cov}} = -\tau \log \left(\frac{1}{T'} \sum_{t=1}^{T'} \exp\left(-\frac{d_t}{\tau}\right) \right), \tag{6}$$

where $\mathcal{J}_{\text{cov}}$ provides a smooth approximation to the minimum temporal deviation, with τ controlling the emphasis on the least-perturbed slices. The resulting structural loss is

$$\mathcal{L}_{\text{Struct}} = -\mathcal{J}_{\text{avg}} - \lambda_{\text{cov}} \mathcal{J}_{\text{cov}}. \tag{7}$$

Minimizing this objective simultaneously increases the overall structural divergence and promotes balanced corruption across the temporal latent sequence, preventing the immunization effect from concentrating only on the earliest structural representations.

4.2 SEMANTIC ATTACK

In Sec. 3.3, we demonstrated that semantic conditioning continuously steers the denoising trajectory through joint attention over image-derived features and textual guidance. While an aligned prompt exerts a corrective pull that heals structural corruptions, a semantically conflicting (collapse-inducing) prompt naturally drives the trajectory toward incoherent video generation. Our goal is to hijack this mechanism: we force intermediate generative representations under plausible aligned prompts to replicate such collapse-inducing dynamics.

To achieve this, we optimize the immunized image against a distribution of plausible aligned prompts while utilizing collapse-inducing generations as semantic targets. For each reference image I, we construct an aligned-prompt distribution $\mathcal{P}_{\text{align}}(\cdot \mid I)$ capturing plausible motion dynamics and a collapse-prompt distribution $\mathcal{P}_{\text{col}}(\cdot \mid I)$ that triggers generation failures. For each sampled collapse prompt $c^{(k)}_{\text{col}}$, we precompute its denoising trajectory $\mathcal{T}^{(k)}_{\text{col}} = \{x^{(k)}_\tau\}$ using the clean image, yielding prompt–trajectory pairs that serve as intermediate collapse targets.

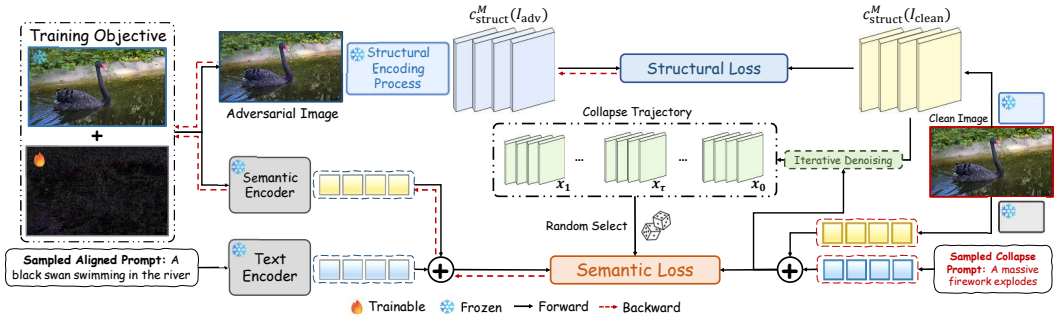


Figure 5: ***Immune2V Framework.*** Our method jointly targets the structural and semantic conditioning streams to ensure persistent disruption throughout video generation. The specific loss formulation of the **Structural Attack** is adapted to each model architecture, but consistently aims to maximize the deviation between clean and perturbed structural conditioning representations. The **Semantic Attack** hijacks the model's guidance by forcing its intermediate representations to mimic a precomputed collapse trajectory.

During adversarial optimization, we sample an aligned prompt $c_{\text{align}} \sim \mathcal{P}_{\text{align}}$, a collapse prompt–trajectory pair $(c_{\text{col}}, \mathcal{T}_{\text{col}}) \sim \mathcal{P}_{\text{col}}$, and a high-noise timestep τ' with corresponding latent $x_{\tau'}$ from $\mathcal{T}_{\text{col}}$. Starting from this shared noisy state $x_{\tau'}$, we compute the representations from the first K generative blocks under the adversarial aligned condition, $h_{\tau'}^{\text{adv}} = f_\theta^{(K)}(x_{\tau'}, I_{\text{adv}}, c_{\text{align}})$, and the collapse target condition, $h_{\tau'}^{\text{col}} = f_\theta^{(K)}(x_{\tau'}, I, c_{\text{col}})$. We then align the adversarial representation with the collapse target by minimizing:

$$\mathcal{L}_{\text{Sem}} = \mathbb{E}_{c_{\text{align}},\, c_{\text{col}},\, \tau'} \left[\left\| h_{\tau'}^{\text{adv}} - h_{\tau'}^{\text{col}} \right\|_2 \right]. \tag{8}$$

Minimizing $\mathcal{L}_{\text{Sem}}$ forces the immunized image to emulate collapse-oriented dynamics even when conditioned on normal, plausible prompts, effectively nullifying the corrective pull of semantic guidance (Fig. 5).

Joint Optimization with Dual-Stream Objectives. We combine the Structural and Semantic Attacks into a unified objective $\mathcal{L}_{\text{total}} = \lambda_{\text{Struct}}\mathcal{L}_{\text{Struct}} + \lambda_{\text{Sem}}\mathcal{L}_{\text{Sem}}$, where λ_{Struct} and λ_{Sem} control the relative strength of disruption in each stream. To preserve the visual appearance of the protected image, we constrain $\|\delta\|_\infty \leq \epsilon$ and optimize the perturbation using Projected Gradient Descent (PGD) (Madry et al., 2018), $\delta \leftarrow \Pi_{\|\delta\|_\infty \leq \epsilon}(\delta - \alpha \nabla_\delta \mathcal{L}_{\text{total}})$. By jointly optimizing these complementary objectives, Immune2V disrupts both the structural anchoring and semantic guidance induced by the reference image, producing severe and persistent degradation throughout the generated video.

5 Experiments

5.1 Evaluation Setup

Dataset. Lacking a public benchmark, we construct an evaluation set of 200 images from three sources: 100 images from DAVIS dataset (Perazzi et al., 2016), 50 images from VBench (Huang et al., 2025), and 50 manually collected / curated human-face images. For DAVIS, we extract the first frame of each selected video. We generate a diverse prompt bank for each image, consisting of 10 aligned prompts describing plausible natural motions (6 used for training, and 4 used for testing), and 4 collapse prompts describing semantically conflicting scenes and dynamics.

Baselines and Implementations. We evaluate against four baselines: **I2VGuard** (Gui et al., 2025), **Anti-I2V** (Vu et al., 2026), **Anti-Prompt** (Song et al., 2026), and **PhotoGuard** (Salman et al., 2023), faithfully reimplementing them following their original papers as some official codebases are unavailable. Clean, unperturbed inputs serve as the quality upper bound. All baselines are evaluated on frozen Wan 2.1 (I2V-14B-480P) (Wan et al., 2025) and HunyuanVideo-I2V (Kong et al., 2024) pipelines. We set the perturbation budget to $\epsilon = 16/255$ and generate 33 frames per video on an NVIDIA A100 (80GB) GPU. Additional dataset, baseline, and implementation details are provided in the Supplementary Material.



Figure 6: ***Qualitative Comparisons.*** A qualitative comparison across baselines, showing the input frame and subsequent video frames.

Automated Quantitative Metrics. Successful immunization degrades conditioning fidelity and structural quality, evaluated across two categories: i) **Condition Preservation** assesses fidelity to conditioning signals. *Perceptual Image Consistency (PIC, ↓)* (Xing et al., 2024a) averages DreamSim (Fu et al., 2023) similarity between input and generated frames to capture source divergence. *Text Alignment (TA, ↓)* computes the average cosine similarity between CLIP (Radford et al., 2021) embeddings of the prompt and each frame for semantic alignment. ii) **Video Structural Quality** evaluates temporal coherence and visual plausibility. Following VBench (Huang et al., 2024), we adopt *Subject Consistency (SC, ↓)* and *Motion Smoothness (MS, ↓)*. SC computes pairwise DINO (Caron et al., 2021) similarities between consecutive frames to measure appearance stability, whereas MS assesses motion regularity based on the reconstruction error of a VFI-based interpolation model that predicts intermediate frames. *Aesthetic Score (AS, ↓)* averages per-frame LAION (Schuhmann et al., 2022) scores to estimate overall visual quality.

5.2 Analysis and Discussion

Table 1: **Quantitative comparison.** **Bold** / underline: best / second-best. All metrics ↓, lower is better.

	Wan 2.1					HunyuanVideo I2V				
Method	PIC	TA	SC	MS	AS	PIC	TA	SC	MS	AS
Clean	85.26	30.14	95.85	96.71	5.13	86.63	29.58	93.91	94.50	5.12
I2VGuard	83.95	29.61	93.56	95.33	5.01	81.74	29.17	93.48	94.22	4.48
Anti-I2V	83.66	29.89	92.83	95.89	4.74	81.93	29.02	92.76	93.68	4.72
Anti-Prompt	83.54	29.55	91.21	**94.01**	4.93	80.68	28.61	91.84	92.07	4.59
PhotoGuard	84.88	30.05	95.60	96.52	5.07	83.56	29.36	94.21	95.14	4.96
Ours	**80.26**	**28.87**	**90.34**	94.42	**4.55**	**79.17**	**27.79**	**90.11**	**91.26**	**4.31**

Automated Quantitative Evaluation. Table 1 compares our method against existing immunization methods. Results are averaged over videos generated using four test prompts. Our method's capability to target both structural and semantic conditioning enables it to produce the strongest and most consistent degradation across both categories of metrics. In contrast, existing baselines primarily perturb image-space representations or peripheral components of the conditioning pipeline, resulting in only limited degradation of the generated videos. The degradation they induce is less consistent across the two modern DiT-based pipelines. Anti-Prompt generally performs better, as it explicitly attenuates text-conditioned interactions during denoising while reinforcing visual-only pathways, thereby weakening prompt guidance. PhotoGuard is particularly ineffective in this setting, producing results close to those obtained with clean inputs across all metrics. This further suggests that protection methods designed for image generation or editing models do not directly transfer to I2V generation. Overall, the consistent reductions in both conditioning-preservation and video-structure metrics demonstrate that our method more effectively disrupts the generated videos.

Qualitative Comparisons. Fig. 6 visualizes the comparisons. Clean inputs and baseline-protected inputs yield coherent videos with plausible motion; our approach induces progressive structural collapse that persists across frames (e.g., the BMX rider becomes geometrically distorted midway



Figure 7: ***Additional Qualitative Results.*** Sampled frames from generated videos. Left: general subjects; right: human faces. All videos are AI-generated solely for research purposes; the depicted motions do not represent real events or actions performed by the individuals shown.

Table 2: ***Ablation on loss components.*** All metrics ↓, lower is better.

$\mathcal{L}_{\text{Struct}}$	$\mathcal{L}_{\text{Sem}}$	PIC	TA	SC	MS	AS
✓	✗	82.71	29.73	91.96	95.04	4.91
✗	✓	84.08	29.46	92.83	95.61	5.07
✓	✓	**80.26**	**28.87**	**90.34**	**94.42**	**4.55**

Table 3: ***Purif.*** Transformations applied before generation. All metrics ↓, lower is better.

Setting	PIC	TA	SC	MS	AS
No protection	85.26	30.14	95.85	96.71	5.13
Ours	**80.26**	**28.87**	**90.34**	**94.42**	**4.55**
+ Gaussian	83.12	29.41	93.18	95.36	4.74
+ JPEG	82.47	29.82	94.05	94.97	4.89
+ Resize	81.68	29.63	93.76	95.81	4.96

Table 4: ***Cross-model Generalization.*** All metrics ↓, lower is better.

Method	PIC	TA	SC	MS	AS
CogVideoX					
No protection	82.12	29.31	92.33	92.05	4.68
Ours	**78.64**	**27.16**	**89.72**	**91.08**	**4.07**
DynamiCrafter					
No protection	80.33	29.35	92.62	93.77	4.11
Ours	**68.71**	**27.42**	**84.66**	**86.97**	**3.11**

through the sequence, while the horse and its rider lose anatomical coherence). Fig. 7 presents additional examples under our immunization. Across diverse scenes involving both general subjects and human faces, the generated videos exhibit progressive structural breakdown, with subject geometry and fine-grained details deteriorating into unstable textures and pronounced temporal flickering emerging across frames. For more results and video versions, please see Supplementary Material.

Ablation Study. We ablate the two components of our proposed loss in Table 2. Applying only the semantic loss yields the weakest degradation, as its perturbed conditional guidance diminishes along the denoising trajectory due to temporal attenuation (Sec.3.2). The structural loss alone yields stronger degradation, while combining both produces the most consistent structural collapse, proving joint optimization is most effective.

Robustness to Input Transformations. We further evaluate whether common image transformations can remove the adversarial protection. We apply Gaussian filtering, JPEG compression, and resizing to the immunized images before I2V generation. The results were evaluated on Wan 2.1 to establish a stronger baseline. As shown in Tab.3, these transformations only moderately weaken the attack, and Immune2V continues to substantially degrade the generated videos, demonstrating robustness against common perturbation-removal (defense) operations.

Cross-Model Generalization. Our generalization experiments cover CogVideoX and DynamiCrafter. Although CogVideoX does not strictly follow the dual-stream design, perturbations optimized solely on Wan 2.1 remain effective when directly transferred without re-optimization (Tab. 4). We deduce that this transferability arises from Immune2V's ability to exploit a common property shared across modern I2V architectures: despite differences in their conditioning designs, the input image participates throughout video generation as a persistent conditioning signal. DynamiCrafter provides a weaker target dual-stream model; perturbations optimized on the stronger model also transfer effectively.

6 Conclusion

We provided a systematic analysis of how reference images condition I2V generation through structural and semantic streams, and examined why naive image-level attacks fail to provide effective immunization against modern I2V models. We then present **Immune2V**, a unified dual-stream immunization framework that jointly disrupts structural conditioning and redirects semantic representations toward collapse-inducing dynamics. Experiments demonstrate effective protection across multiple modern I2V models. We hope our findings can motivate further research on robust and generalizable image immunization for increasingly capable video generation systems.

7 STATEMENTS

AI USAGE STATEMENT

In this work, we used generative AI tools to assist with the construction of our evaluation data. Specifically, generative AI (LLM) was used to generate prompts paired with selected input images (as described in the main paper, the aligned and collapse prompts) and to assist with checking and cleaning these image–prompt pairs. The source images themselves were obtained from existing datasets and open resources and were not generated by AI. We did not use generative AI to generate synthetic image datasets, design the methodology or experiments, interpret results, or perform other core research tasks.

Additionally, we used generative AI tools to polish the writing of the manuscript and to assist with code debugging. All AI-assisted dataset construction and curation, code changes, and writing polish were reviewed and verified by the authors before being used in the final work. We take responsibility for the final content of this work.

ETHICS STATEMENT

This work promotes the safety and responsible deployment of generative AI by developing proactive defenses to protect visual content against unauthorized image-to-video manipulation and potential deepfake misuse. Our work focuses on immunization and protection rather than the creation or facilitation of harmful content.

All videos presented in this work, including videos involving human faces, are generated by AI (video) models for research and evaluation purposes and do not correspond to real recorded events. Facial examples are included solely to better illustrate potential misuse scenarios and to evaluate the effectiveness of the proposed protection. We do not generate or present offensive, harmful, or otherwise inappropriate content.

The source face images used in our experiments were obtained from publicly available resources, primarily Wikimedia Commons, under their respective licenses. These images are used only for research evaluation and demonstration of the proposed protection mechanism.

REPRODUCIBILITY STATEMENT

Implementation details, experimental settings, and hyperparameters can refer to section A and B.1. We also provide generated videos used in the paper in the Supplementary Material.

Upon acceptance, we will release the implementation code and dataset together with the corresponding source, licensing, and attribution information.

REFERENCES

Namhyuk Ahn, Wonhyuk Ahn, KiYoon Yoo, Daesik Kim, and Seung-Hun Nam. Imperceptible protection against style imitation from diffusion models. *IEEE Transactions on Multimedia*, 2026.

Anurag Arnab, Ondrej Miksik, and Philip HS Torr. On the robustness of semantic segmentation models to adversarial attacks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 888–897, 2018.

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 Conference Papers*, pp. 1–11, 2024.

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*, 2023a.

Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22563–22575, 2023b.

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

Shoufa Chen, Chongjian Ge, Yuqi Zhang, Yida Zhang, Fengda Zhu, Hao Yang, Hongxiang Hao, Hui Wu, Zhichao Lai, Yifei Hu, et al. Goku: Flow based video generative foundation models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 23516–23527, 2025.

Yipu Chen, Haotian Xue, and Yongxin Chen. Diffusion policy attacker: Crafting adversarial attacks for diffusion-based policies. *Advances in Neural Information Processing Systems*, 37:119614–119637, 2024.

June Suk Choi, Kyungmin Lee, Jongheon Jeong, Saining Xie, Jinwoo Shin, and Kimin Lee. Diffusionguard: A robust defense against malicious diffusion-based image editing. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=9OfKxKoYNw.

Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. *Advances in Neural Information Processing Systems*, 36:50742–50768, 2023.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *3rd International Conference on Learning Representations*, 2015.

Dongnan Gui, Xun Guo, Wengang Zhou, and Yan Lu. I2vguard: Safeguarding images against misuse in diffusion-based image-to-video models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 12595–12604, 2025.

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 *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Fx2SbBgcte.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022.

Li Hu. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8153–8163, 2024.

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*, pp. 21807–21818, 2024.

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*, 2025.

Redwan Hussain, Mizanur Rahman, and Prithwiraj Bhattacharjee. Toward generalized detection of synthetic media: Limitations, challenges, and the path to multimodal solutions. *arXiv preprint arXiv:2511.11116*, 2025.

Xiaojun Jia, Sensen Gao, Simeng Qin, Tianyu Pang, Chao Du, Yihao Huang, Xinfeng Li, Yiming Li, Bo Li, and Yang Liu. Adversarial attacks against closed-source MLLMs via feature optimal alignment. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=ktC3cDu320.

Yang Jin, Zhicheng Sun, Ningyuan Li, Kun Xu, Kun Xu, Hao Jiang, Nan Zhuang, Quzhe Huang, Yang Song, Yadong MU, and Zhouchen Lin. Pyramidal flow matching for efficient video generative modeling. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=66NzcRQuOq.

Akansha Kalra, Basavasagar Patil, Guanhong Tao, and Daniel S Brown. How vulnerable is my learned policy? universal adversarial perturbation attacks on modern behavior cloning policies. *arXiv preprint arXiv:2502.03698*, 2025.

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.

Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *Artificial intelligence safety and security*, pp. 99–112. Chapman and Hall/CRC, 2018.

Changzhen Li, Yuecong Min, Jie Zhang, Zheng Yuan, Shiguang Shan, and Xilin Chen. T2vattack: Adversarial attack on text-to-video diffusion models. *arXiv preprint arXiv:2512.23953*, 2025.

Guanlin Li, Shuai Yang, Jie Zhang, and Tianwei Zhang. Prime: Protect your videos from malicious editing. *arXiv preprint arXiv:2402.01239*, 2024.

Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. Face x-ray for more general face forgery detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5000–5009. IEEE, 2020.

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

Chumeng Liang, Xiaoyu Wu, Yang Hua, Jiaru Zhang, Yiming Xue, Tao Song, Zhengui Xue, Ruhui Ma, and Haibing Guan. Adversarial example does good: Preventing painting imitation from diffusion models via adversarial examples. In *International Conference on Machine Learning*, pp. 20763–20786. PMLR, 2023.

Xingchao Liu, Chengyue Gong, and qiang liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=XVjTT1nw5z.

Yixin Liu, Chenrui Fan, Yutong Dai, Xun Chen, Pan Zhou, and Lichao Sun. Metacloak: Preventing unauthorized subject-driven text-to-image diffusion-based synthesis via meta-learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24219–24228, 2024a.

Yixin Liu, Kai Zhang, Yuan Li, Zhiling Yan, Chujie Gao, Ruoxi Chen, Zhengqing Yuan, Yue Huang, Hanchi Sun, Jianfeng Gao, et al. Sora: A review on background, technology, limitations, and opportunities of large vision models. *arXiv preprint arXiv:2402.17177*, 2024b.

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

Yue Ma, Kunyu Feng, Zhongyuan Hu, Xinyu Wang, Yucheng Wang, Mingzhe Zheng, Bingyuan Wang, Qinghe Wang, Xuanhua He, Hongfa Wang, et al. Controllable video generation: A survey. *arXiv preprint arXiv:2507.16869*, 2025.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

Tarik Can Ozden, Ozgur Kara, Oguzhan Akcin, Kerem Zaman, Shashank Srivastava, Sandeep P. Chinchali, and James Matthew Rehg. Diffvax: Optimization-free image immunization against diffusion-based editing. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=QEJaKJYOIn.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

Bohao Peng, Jian Wang, Yuechen Zhang, Wenbo Li, Ming-Chang Yang, and Jiaya Jia. Controlnext: Powerful and efficient control for image and video generation. *arXiv preprint arXiv:2408.06070*, 2024.

Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 724–732, 2016.

Claudiu Popa, Rex Pallath, Liam Cunningham, Hewad Tahiri, Abiram Kesavarajah, and Tao Wu. Deepfake technology unveiled: the commoditization of ai and its impact on digital trust. *arXiv preprint arXiv:2506.07363*, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

Md Shohel Rana, Mohammad Nur Nobi, Beddhu Murali, and Andrew H Sung. Deepfake detection: A systematic literature review. *IEEE access*, 10:25494–25513, 2022.

Weiming Ren, Huan Yang, Ge Zhang, Cong Wei, Xinrun Du, Wenhao Huang, and Wenhu Chen. Consisti2v: Enhancing visual consistency for image-to-video generation. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=vqniLmUDvj.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1–11, 2019.

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

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022.

Shawn Shan, Jenna Cryan, Emily Wenger, Haitao Zheng, Rana Hanocka, and Ben Y Zhao. Glaze: Protecting artists from style mimicry by {Text-to-Image} models. In *32nd USENIX Security Symposium (USENIX Security 23)*, pp. 2187–2204, 2023.

Linghui Shen, Mingyue Cui, and Xingyi Yang. Decontext as defense: Safe image editing in diffusion transformers. *arXiv preprint arXiv:2512.16625*, 2025.

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-a-video: Text-to-video generation without text-video data. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=nJfylDvgzlq.

Yeonghwan Song, Chanhui Lee, Jinsoo Park, and Jeany Son. Anti-prompt: Image protection against text-guided image-to-video generation. *arXiv preprint arXiv:2607.01499*, 2026.

Duc Vu, Anh Nguyen, Chi Tran, and Anh Tran. Anti-i2v: Safeguarding your photos from malicious image-to-video generation. *arXiv preprint arXiv:2603.24570*, 2026.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models, 2025. URL https://arxiv.org/abs/2503.20314.

Songping Wang, Rufan Qian, Yueming Lyu, Qinglong Liu, Linzhuang Zou, Jie Qin, Songhua Liu, and Caifeng Shan. Runawayevil: Jailbreaking the image-to-video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9296–9305, 2026.

Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, and Houqiang Li. Altfreezing for more general video face forgery detection. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4129–4138. IEEE, 2023.

Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Wangbo Yu, Hanyuan Liu, Gongye Liu, Xintao Wang, Ying Shan, and Tien-Tsin Wong. Dynamicrafter: Animating open-domain images with video diffusion priors. In *European Conference on Computer Vision*, pp. 399–417. Springer, 2024a.

Zhen Xing, Qijun Feng, Haoran Chen, Qi Dai, Han Hu, Hang Xu, Zuxuan Wu, and Yu-Gang Jiang. A survey on video diffusion models. *ACM Computing Surveys*, 57(2):1–42, 2024b.

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

Haotian Xue, Chumeng Liang, Xiaoyu Wu, and Yongxin Chen. Toward effective protection against diffusion-based mimicry through score distillation. In *International Conference on Learning Representations*, volume 2024, pp. 25734–25758, 2024.

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

Jiwen Yu, Xiaodong Cun, Chenyang Qi, Yong Zhang, Xintao Wang, Ying Shan, and Jian Zhang. Animatezero: Video diffusion models are zero-shot image animators. *arXiv preprint arXiv:2312.03793*, 2023.

Shiwei Zhang, Jiayu Wang, Yingya Zhang, Kang Zhao, Hangjie Yuan, Zhiwu Qin, Xiang Wang, Deli Zhao, and Jingren Zhou. I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models. *arXiv preprint arXiv:2311.04145*, 2023.

APPENDIX OVERVIEW

A Immune2V Algorithm Details **17**

A.1 Method Overview 17

A.2 Model-Specific Implementation Notes 18

B Additional Experiment Details **19**

B.1 Attack Optimization Details & Hyperparameters 19

B.2 Baselines 20

C Dataset and Prompts **21**

C.1 Dataset Construction 21

C.2 Aligned and Collapse Prompts. 22

D Additional Results on other I2V Models **23**

E Challenging Cases **24**

F Broader Impact **25**

F.1 Positive Societal Impact 25

F.2 Potential Misuse and Responsible Release 25

F.3 Identity and Face-Specific Considerations 25

A IMMUNE2V ALGORITHM DETAILS

A.1 METHOD OVERVIEW

Immune2V is designed for image immunization in Image-to-Video (I2V) generation, where the input image influences the video generation process through structural and semantic conditioning. By disrupting both pathways, the generation dynamics can be effectively destabilized. The goal is to optimize a small perturbation on the input image that simultaneously attacks both streams. The **Structural Attack** maximizes the discrepancy between clean and adversarial structural conditions, with its loss formulation adapted to the model's conditioning architecture. The **Semantic Attack** aligns intermediate representations under plausible aligned prompts with collapse-inducing representations obtained from precomputed trajectories, counteracting semantic guidance toward coherent generation. Together, these two attacks induce persistent structural breakdown throughout the generated video.

Algorithm 1 summarizes the overall optimization procedure. The same framework can be readily extended to other dual-stream I2V architectures by instantiating the structural divergence objective according to their corresponding conditioning representations. Here, $\mathcal{D}$ denotes the architecture-specific structural divergence objective. At each iteration, the semantic loss is estimated using a sampled aligned prompt, a cached collapse prompt–trajectory pair, and a high-noise timestep from that trajectory.

Algorithm 1 Immune2V: Joint Structural and Semantic Immunization for I2V Models

Input: Clean image I; aligned- and collapse-prompt distributions $\mathcal{P}_{\text{align}}(\cdot \mid I)$ and $\mathcal{P}_{\text{col}}(\cdot \mid I)$; structural conditioning function c^{M}_{struct}; structural divergence objective $\mathcal{D}$; generative backbone f_θ; number of generative blocks K; number of cached collapse trajectories B; step size α; perturbation budget ϵ; loss weights $\lambda_{\text{Struct}}, \lambda_{\text{Sem}}$; optimization steps N.

Initialize $\delta \sim \mathcal{U}(-\epsilon, \epsilon)$
$s^{\text{clean}} \leftarrow c^{M}_{\text{struct}}(I)$
$\mathcal{C}_{\text{col}} \leftarrow \varnothing$
for $k = 1$ to B **do**
  Sample $c^{(k)}_{\text{col}} \sim \mathcal{P}_{\text{col}}(\cdot \mid I)$
  Run denoising with $(I, c^{(k)}_{\text{col}})$ and cache trajectory $\mathcal{T}^{(k)}_{\text{col}} = \{x^{(k)}_\tau\}$
  Add $(c^{(k)}_{\text{col}}, \mathcal{T}^{(k)}_{\text{col}})$ to $\mathcal{C}_{\text{col}}$
end for
for $i = 1$ to N **do**
  $I_{\text{adv}} \leftarrow I + \delta$
  Structural Attack
  $s^{\text{adv}} \leftarrow c^{M}_{\text{struct}}(I_{\text{adv}})$
  $\mathcal{L}_{\text{Struct}} \leftarrow -\mathcal{D}(s^{\text{adv}}, s^{\text{clean}})$
  Semantic Attack
  Sample $c_{\text{align}} \sim \mathcal{P}_{\text{align}}(\cdot \mid I)$
  Sample a pair $(c_{\text{col}}, \mathcal{T}_{\text{col}})$ from $\mathcal{C}_{\text{col}}$
  Sample a high-noise timestep τ' from $\mathcal{T}_{\text{col}}$ and retrieve $x_{\tau'}$
  $h^{\text{adv}}_{\tau'} \leftarrow f^{(K)}_\theta(x_{\tau'}, I_{\text{adv}}, c_{\text{align}})$
  $h^{\text{col}}_{\tau'} \leftarrow f^{(K)}_\theta(x_{\tau'}, I, c_{\text{col}})$
  Treat $h^{\text{col}}_{\tau'}$ as a fixed target
  $\mathcal{L}_{\text{Sem}} \leftarrow \|h^{\text{adv}}_{\tau'} - h^{\text{col}}_{\tau'}\|_2$
  Joint Optimization
  $\mathcal{L}_{\text{total}} \leftarrow \lambda_{\text{Struct}}\mathcal{L}_{\text{Struct}} + \lambda_{\text{Sem}}\mathcal{L}_{\text{Sem}}$
  $\delta \leftarrow \Pi_{\|\delta\|_\infty \le \epsilon}\left(\delta - \alpha \nabla_\delta \mathcal{L}_{\text{total}}\right)$
end for
return $I_{\text{adv}} = I + \delta$

A.2 Model-Specific Implementation Notes

While the formulation of Immune2V is intentionally general, modern I2V systems may instantiate the dual-stream conditioning structure with slightly different architectural components. In practice, structural conditioning is commonly derived from encoded reference-image representations and injected directly into the denoising process, for example through latent concatenation or feature-level conditioning. Some implementations employ spatio-temporal VAEs that compress the temporal dimension during encoding (Wan et al., 2025), while others encode frames independently and preserve the full temporal sequence (Blattmann et al., 2023a; Xing et al., 2024a). Other models incorporate such information through first-frame features or reference-aware attention (Zhang et al., 2023; Ren et al., 2024). Our structural attack objective is compatible with these designs, as it operates directly on the corresponding reference-conditioned representations. In Wan 2.1, the encoded latent is additionally combined with Gaussian noise through a small MLP to produce the initial latent for the diffusion process. This mainly serves as a specific initialization strategy and does not change the fundamental role of the structural conditioning pathway.

For the semantic conditioning pathway, the input image is usually encoded by a pretrained vision-language encoder. In many pipelines, this is a CLIP-based image encoder (Radford et al., 2021), which provides rich semantic features for conditioning the generative process. The text prompt is commonly processed by a large language encoder such as T5 (Raffel et al., 2020) or a similar transformer-based text model. Other models also explore multimodal large language models (MLLMs) for richer semantic understanding (Kong et al., 2024). HunyuanVideo-I2V employs a pretrained decoder-only MLLM to jointly process the reference image and textual prompt, producing semantic image tokens that are concatenated with the video latent tokens and fused through full-attention computation. The resulting visual and textual representations are then fused within the denoising backbone, which may be implemented using U-Net (Ronneberger et al., 2015) or DiT-based (Peebles & Xie, 2023) architectures.

Again, Immune2V does not rely on the specific choice of encoders or backbone architecture. As a result, the proposed optimization procedure can be applied to a wide range of dual-stream I2V generation frameworks with minimal modification.

B Additional Experiment Details

B.1 Attack Optimization Details & Hyperparameters

For the Wan 2.1 (I2V-14B-480P) pipeline (Wan et al., 2025) implementation, the input resolution is set to 480×832, and the number of generated frames is fixed to 33. We optimize input-space perturbations using Projected Gradient Descent (PGD) for 500 iterations under an ℓ_∞ constraint of $\epsilon = 16/255$ and a step size $\alpha = 0.001$. PGD is initialized with a random perturbation uniformly sampled within the ℓ_∞ constraint.

During optimization, collapse trajectories need to be generated. We set the number of denoising steps to 25, which defines the trajectory length. The CFG scale can be adjusted (e.g., 1, 3, or 5) to control the strength of semantic conditioning.

To target the early stage of video layout formation, the attack is restricted to the first 10 timesteps out of the 25 total inference steps, corresponding to the region closest to the pure-noise initialization stage. For latent feature extraction, we use the outputs of the first $K = 4$ DiT blocks to compute the semantic attack objective. At each optimization step, one aligned prompt, one collapse trajectory, and one timestep are uniformly sampled. We use 6 aligned prompts during optimization and 4 held-out prompts for evaluation, together with 4 collapse prompts, each associated with a precomputed trajectory.

The coefficients λ_{Struct} and λ_{Sem} control the relative strengths of the structural and semantic losses. Since the semantic loss typically has a larger magnitude and higher variance during optimization, we set $\lambda_{\text{Struct}} = 1$ and $\lambda_{\text{Sem}} = 0.125$ to balance the two objectives. For the temporal coverage term in the structural objective, we set $\lambda_{\text{cov}} = 0.5$ and $\tau = 0.1$. These weights can be adjusted to emphasize either structural collapse or semantic interference depending on the desired attack behavior.

For HunyuanVideo-I2V, the image resolution is 352×608 on 360p generation bucket. We generate 33 frames and optimize the perturbation for 200 PGD iterations with $\epsilon = 16/255$ and $\alpha = 1/255$. The collapse trajectory contains 50 denoising steps, and the semantic attack samples from the first 20 timesteps. We use the first $K = 4$ double-stream transformer blocks for semantic feature extraction, with $\lambda_{\text{Struct}} = 1$ and $\lambda_{\text{Sem}} = 0.001$. The CFG scale is 1.0, while the model additionally receives an embedded guidance scale of 6.0. Since the structural conditioning latent contains a single temporal slice in this configuration, the structural objective uses a global latent RMS deviation without the temporal coverage term used for Wan 2.1.

With gradient checkpointing, truncating the DiT to the first $K = 4$ blocks, freezing all model weights, and computing the target branch without gradient tracking during the backward pass, immunizing a single image requires approximately 20 minutes on one NVIDIA A100 (80GB) GPU. Video generation during testing takes approximately 1 minute per sample. With larger GPU memory, the optimization can be extended to longer videos with more frames.

All hyperparameters used in our experiments are summarized in Tables 5 and 6.

Table 5: Hyperparameters for Immune2V optimization on Wan 2.1.

Resolution	**Frames**	**PGD Iters**	**Budget** ϵ	**Step** α	**Traj. Length**	**Timestep Pool**	**DiT** (K)	λ_{Struct}	λ_{Sem}	λ_{cov}	τ
480×832	33	500	$16/255$	0.001	25	First 10	4	1	0.125	0.5	0.1

Table 6: Hyperparameters for Immune2V optimization on HunyuanVideo-I2V.

Resolution	**Frames**	**PGD Iters**	**Budget** ϵ	**Step** α	**Traj. Length**	**Timestep Pool**	**DiT** (K)	λ_{Struct}	λ_{Sem}
352×608	33	200	$16/255$	$1/255$	50	First 20	4	1	0.001

B.2 BASELINES

In both quantitative and qualitative experiments, we compare our method with I2VGuard, Anti-I2V, Anti-Prompt, and PhotoGuard. Their respective attack objectives can be summarized as follows:

$$\mathcal{L}_{\text{I2VGuard}} = \|I_{\text{adv}} - I\|_2^2 + \alpha\mathcal{L}_{\text{enc}} + \beta\mathcal{L}_{\text{con}} + \gamma\mathcal{L}_{\text{att}}, \tag{9}$$

$$\mathcal{L}_{\text{Anti-I2V}} = \mathcal{L}_{\text{IRC}} + \mathcal{L}_{\text{IRA}} + \mathcal{L}_{\text{auxiliary}} - \mathcal{L}_{\text{DM}}, \tag{10}$$

$$\mathcal{L}_{\text{Anti-Prompt}} = \mathcal{L}_{\text{sup}} + \lambda_1\mathcal{L}_{\text{vis}} + \lambda_2\mathcal{L}_{\text{enc}}. \tag{11}$$

For I2VGuard, $\mathcal{L}_{\text{enc}}$, $\mathcal{L}_{\text{con}}$, and $\mathcal{L}_{\text{att}}$ denote its encoder, contrastive diffusion, and temporal-attention attacks, respectively. For Anti-I2V, $\mathcal{L}_{\text{IRC}}$ and $\mathcal{L}_{\text{IRA}}$ denote Internal Representation Collapse and Internal Representation Anchor, $\mathcal{L}_{\text{auxiliary}}$ combines its CLIP and LPIPS objectives, and $\mathcal{L}_{\text{DM}}$ is the diffusion denoising loss. For Anti-Prompt, $\mathcal{L}_{\text{sup}}$, $\mathcal{L}_{\text{vis}}$, and $\mathcal{L}_{\text{enc}}$ correspond to text suppression, visual dominance, and encoder attack, respectively.

For PhotoGuard (Salman et al., 2023), we combine its encoder-level and diffusion-level objectives:

$$\mathcal{L}_{\text{PG}} = \lambda_{\text{enc}}\mathcal{L}_{\text{enc}}^{\text{PG}} + \lambda_{\text{diff}}\mathcal{L}_{\text{diff}}^{\text{PG}}, \tag{12}$$

$$\mathcal{L}_{\text{enc}}^{\text{PG}} = \left\|\mathcal{E}(I_{\text{adv}}) - z^{\text{tgt}}\right\|_2^2, \tag{13}$$

$$\mathcal{L}_{\text{diff}}^{\text{PG}} = \|f(I_{\text{adv}}) - I_{\text{tgt}}\|_2^2. \tag{14}$$

As most baselines are close-sourced, we reproduce them following their publicly available papers and released implementations whenever available. For methods without complete official implementations, we follow the reported objectives and hyperparameters as closely as possible and use consistent settings across methods where implementation details are unspecified.

C Dataset and Prompts

C.1 Dataset Construction

We construct a diverse evaluation set containing a total of 200 source images from three complementary sources: 100 images from DAVIS (Perazzi et al., 2016), 50 images from the I2V Image Suite of VBench++ (Huang et al., 2025), and 50 human-face images collected from Wikimedia Commons. The three sources are selected to provide complementary coverage of common objects and animals, diverse I2V benchmark content, and human-centric facial content.

DAVIS. The DAVIS dataset (Perazzi et al., 2016) contains high-quality videos with clearly defined foreground objects. We select the first frame from 100 DAVIS videos where the main foreground object is clearly visible and can be unambiguously identified. The selected frames cover common semantic categories such as humans, animals, vehicles, and everyday objects, together with diverse backgrounds and scene layouts. Using the first frame provides a natural single-image input for I2V generation while retaining the realistic visual statistics and foreground–background composition of the original videos.

VBench++. We include 50 images from the official I2V Image Suite of VBench++ (Huang et al., 2025). The VBench++ I2V benchmark is specifically constructed for evaluating image-to-video generation and contains both subject-centric and background-centric images. Its subject categories include single-human, multiple-human, animal, transportation, food, plant, and other objects, while its background categories include architecture, scenery, indoor scenes, and abstract content. We select images across these categories to complement the predominantly foreground-centric examples obtained from DAVIS. The original VBench++ images are high-resolution images, with the majority of the image suite provided at approximately 4K resolution or above. It additionally provides an adaptive cropping pipeline that preserves the primary visual content while producing several common aspect ratios with different native input resolutions. We use the VBench++ images as source visual content, while the textual conditioning used in our experiments is constructed independently as described below rather than directly adopting the original prompts.

Human-Face Images. We additionally collect 50 human-face images from Wikimedia Commons. We select images in which the primary face is clearly visible, while retaining diversity in appearance, pose, expression, background, and photographic conditions. These examples provide a dedicated human-centric subset for evaluating whether image immunization remains effective when preservation of facial appearance and identity is important. For each Wikimedia Commons image, we retain its original source URL together with the available author, licensing or public-domain status, and attribution information.

Together, these three subsets form a 200-image evaluation dataset with complementary content distributions. DAVIS contributes realistic foreground objects extracted from natural videos, VBench++ broadens the coverage of object and scene categories specifically used in I2V evaluation, and the Wikimedia Commons subset provides additional coverage of human faces. Images are subsequently resized or cropped according to the native input configuration of each evaluated I2V model, as described in Sec. B.1.

C.2 ALIGNED AND COLLAPSE PROMPTS.

For each source image, we construct two types of textual conditions, referred to as *Aligned Prompts* and *Collapse Prompts*. All prompts are generated using Claude Opus 5.

The *Aligned Prompts* are designed to remain semantically consistent with the visual content of the source image while covering diverse, plausible motion patterns. Specifically, they span (i) ordinary subject motion, such as walking, turning, or subtle body movement; (ii) more dynamic or large-scale actions, such as running, jumping, or rapid object displacement; (iii) natural interactions between subjects and surrounding objects; (iv) pronounced facial expression changes and expressive body motions, such as sudden head, arm, or upper-body movements; and (v) camera motion, including panning, zooming, and moderate viewpoint changes. All prompts remain compatible with the depicted content and exclude offensive, harmful, or otherwise inappropriate scenarios. For each image, we construct 6 aligned prompts for attack optimization and an additional 4 held-out aligned prompts used only for evaluation, allowing us to assess generalization to unseen but semantically consistent textual conditions. Prompts are shuffled.

The *Collapse Prompts* are designed to introduce strong semantic mismatch with the source image. They describe unrelated subjects, actions, or scenes that are deliberately incompatible with the depicted content, thereby providing conflicting textual guidance. We construct 4 collapse prompts for each source image, covering diverse types of semantic deviation. These prompts are used only to define the semantic attack objective.

D ADDITIONAL RESULTS ON OTHER I2V MODELS



Figure 8: Clean and attacked video results on DynamiCrafter using Immune2V.

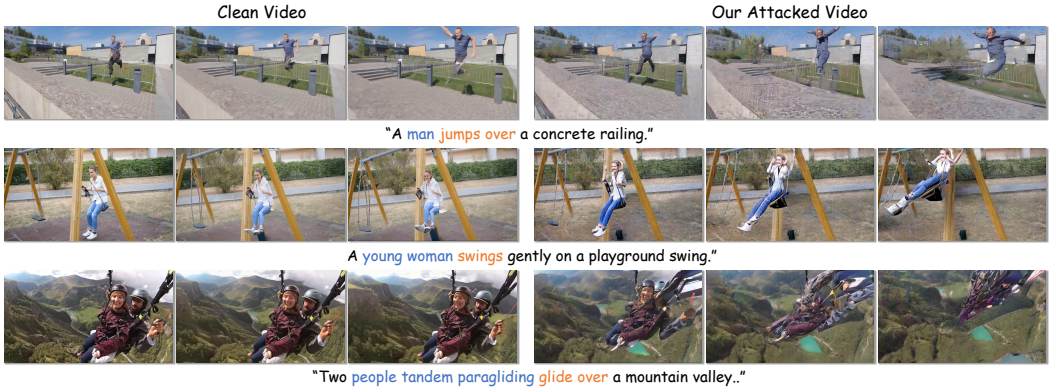


Figure 9: Clean and attacked video results on I2VGen-XL using Immune2V.

We provide qualitative examples on two earlier dual-stream I2V models, DynamiCrafter (Xing et al., 2024a) and I2VGen-XL (Zhang et al., 2023), in Figs. 8 and 9. In both cases, the attacks easily lead to visible distortions and motion inconsistencies, further confirming that our approach generalizes to dual-stream I2V models.

Attacking these earlier models appears **significantly easier** than attacking Wan 2.1 and HunyuanVideo-I2V. We conjecture that this is largely due to architectural and training differences, such as older VAE designs and more limited training data. While strong degradation can be readily achieved on these earlier models, this does not necessarily imply that the I2V immunization task itself is trivial. Instead, stronger and more modern models provide a more challenging and informative testbed for systematically analyzing the structure of dual-stream I2V pipelines and designing targeted attack strategies.

E CHALLENGING CASES

Image immunization for I2V generation remains a challenging problem due to the strong conditioning imposed by the input image and the complex temporal dynamics involved in video generation. In our experiments, Immune2V generally produces stronger and more consistent disruption when the input admits clear and meaningful motion, such as human actions, animal movements, or object interactions. In these cases, the generation process needs to establish coherent temporal dynamics from the reference image, providing richer structural and semantic representations for the attack to disrupt. In contrast, images with weak, ambiguous, or nearly static motion cues may admit many plausible generation trajectories, making the attack effect less predictable.

Another limitation is the coverage of the prompt distribution used during optimization. Although our aligned and collapse prompt sets are constructed to cover diverse plausible motions and semantically conflicting scenarios, they still represent only a subset of the much broader prompt distribution encountered in real-world I2V usage. The current prompt banks can be naturally expanded by introducing more aligned prompts, collapse prompts, and motion patterns for each source image. With additional computational resources, scaling these prompt distributions provides a direct way to approximate increasingly diverse real-world generation conditions and potentially improve robustness to unseen prompts.

F BROADER IMPACT

F.1 POSITIVE SOCIETAL IMPACT

The proposed Immune2V framework aims to improve the safety of image-to-video generation systems by enabling proactive protection of visual content. As generative video models become increasingly capable (Xing et al., 2024b; Ma et al., 2025), concerns about unauthorized manipulation and synthetic media creation have also grown (Hussain et al., 2025; Popa et al., 2025). By introducing an image immunization mechanism, our method allows creators and platform providers to reduce the risk firsthand that images are later used to generate misleading or manipulated videos. Such protection mechanism is particularly valuable for journalists, public figures, and individuals whose images are frequently shared online. Our work therefore contributes to the growing line of research on defensive technologies that mitigate potential harms of generative media.

F.2 POTENTIAL MISUSE AND RESPONSIBLE RELEASE

Although Immune2V is designed as a defensive mechanism, adversarial perturbation techniques may potentially be repurposed in unintended ways. For example, similar optimization procedures could be adapted to manipulate generative systems or interfere with other machine learning models. As with many security-related research areas, there exists a dual-use risk where defensive insights could inspire new forms of attacks. To mitigate such risks, we frame our work primarily as a protection mechanism for visual content rather than a tool for exploiting generative systems. Our experimental evaluation focuses on understanding vulnerabilities in current I2V pipelines and developing countermeasures against misuse.

F.3 IDENTITY AND FACE-SPECIFIC CONSIDERATIONS

Image-to-video generation systems raise important concerns when applied to identity-related visual content. Synthetic videos generated from personal images may contribute to misinformation, reputational harm, or non-consensual media creation. In this context, protective methods may serve as a potential safeguard against the unauthorized reuse of visual content in generative pipelines.

Identity preservation in generative video systems involves not only facial images but also broader identity-related visual signals. Distinctive clothing, personal belongings, private environments, or contextual cues may all be used to falsely suggest that a person appeared in a particular place, participated in a certain event, or engaged in an action that never occurred. These risks highlight the importance of considering identity-related information beyond facial appearance when developing protection mechanisms against generative misuse.

At the same time, building reliable safeguards for identity-related content remains a long-term challenge. Future work may further investigate how protection mechanisms can generalize across different forms of personal visual information while maintaining the responsible use of generative technologies.