I2VShield: An Efficient Proactive Defense Framework against DiT-based Image-to-Video Models

Yimao Guo Zuomin Qu Wei Lu

School of Computer Science and Engineering, Sun Yat-sen University
Ministry of Education Key Laboratory of Information Technology
Guangdong Province Key Laboratory of Information Security Technology
Guangzhou 510006, China

guoym39@mail2.sysu.edu.cn; quzm@mail2.sysu.edu.cn; luwei3@mail.sysu.edu.cn

Abstract

The rapid advancement of video generation models has led to the increasing misuse of image-to-video (I2V) models. Although substantial progress has been made in detecting AI-generated videos, proactive defenses against I2V models remain underexplored. In particular, current proactive defenses against I2V models predominantly rely on gradient-based adversarial attacks, which require defenders to possess GPUs with substantial memory resources (VRAM) to generate adversarial examples. To address this issue, we propose I2VShield, a privacy protection method based on generative adversarial attacks tailored to Diffusion Transformer (DiT)-based I2V models. The proposed method primarily consists of two components: (1) a text-adaptive perturbation generation framework integrating adversarial learning to mitigate computational overhead while maintaining visual imperceptibility; and (2) an untargeted Multimodal Attention Disruption (MAD) attack that exploits the inherent vulnerabilities of DiT-based I2V models, maximizing the deviation of the internal attention features from their clean states. Extensive experiments demonstrate that our approach achieves highly competitive protection performance across various datasets and mainstream DiT-based I2V models, particularly in disrupting spatiotemporal coherence, while substantially reducing computational costs.

1 Introduction

The rapid evolution of generative artificial intelligence has catalyzed unprecedented advancements in image-to-video (I2V) generation. Driven by the recent paradigm shift toward Diffusion Transformers (DiT), modern I2V models exhibit remarkable capabilities in generating temporally coherent and highly realistic video sequences from static reference images and textual prompts [7]. However, the democratization and widespread availability of these powerful I2V tools have inevitably led to rampant abuse. Malicious actors frequently exploit these models to animate private portraits without authorization, generating deepfakes and malicious content that pose serious threats to personal privacy and public safety.

To mitigate the negative impacts of generative models, extensive research has been dedicated to AI-generated content (AIGC) forgery detection [16, 23, 25]. While these passive detection mechanisms have achieved notable progress, they are inherently reactive; they can only identify malicious content after it has been generated and potentially disseminated. Consequently, there is an urgent need for proactive defense strategies that prevent the unauthorized generation of videos from the outset. Adversarial perturbations have emerged as a promising proactive defense mechanism that

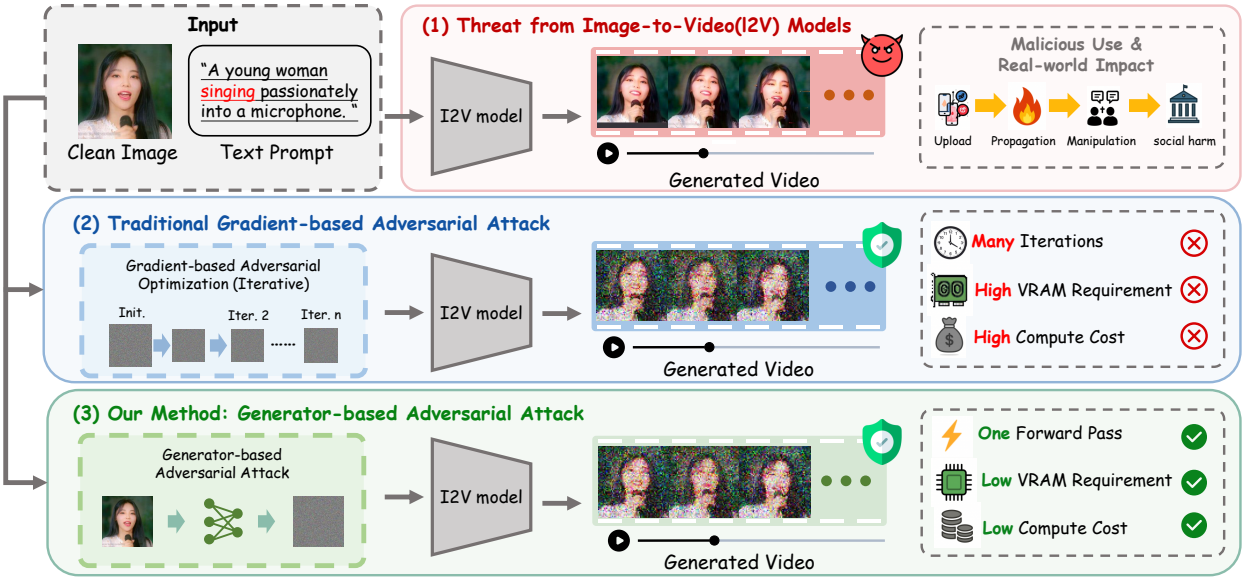

Figure 1: Background and motivation for proactive defense. (1) The abuse of I2V models poses severe privacy threats. (2) Traditional gradient-based adversarial attacks rely on expensive iterative processes, incurring high computational overhead and massive VRAM requirements. (3) Our proposed I2VShield utilizes an efficient generative attack framework to achieve effective defense at a low computational cost.

protects users' privacy by modifying personal images before they are shared online. When an I2V model attempts to process these protected images, the perturbation disrupts the generative process, resulting in severe degradation or semantic collapse in the video.

Despite the conceptual viability of proactive defenses, applying existing adversarial attack methods to modern DiT-based I2V models presents significant challenges. Current proactive defenses predominantly rely on iterative, gradient-based optimization techniques (e.g., PGD-based attacks) to craft adversarial examples [14]. While effective for image-to-image tasks, extending these gradient-based attacks to the video domain is highly problematic. Generating perturbations for I2V models requires backpropagating gradients through the complex diffusion process across multiple video frames. This optimization process demands prohibitive computational resources and substantial GPU memory, rendering it impractical for everyday users or defenders with limited hardware budgets. Furthermore, existing methods often fail to exploit the specific architectural vulnerabilities inherent to Diffusion Transformers, resulting in limited protection effectiveness against state-of-the-art I2V models.

To address these critical limitations, we propose I2VShield. To the best of our knowledge, this work introduces the first generative adversarial attack framework specifically tailored for DiT-based I2V models. Unlike prior instance-wise optimization methods, I2VShield employs a text-adaptive perturbation generation network that predicts adversarial perturbations in a single forward pass. This design eliminates iterative gradient computation during deployment, substantially reducing computational requirements.

Furthermore, to maximize the disruptive capability of the generated noise, we introduce an untargeted Multimodal Attention Disruption (MAD) attack. By analyzing the conditioning mechanisms of DiT architectures, we find that cross-attention representations provide an effective optimization target for disrupting multimodal conditioning. Our attack explicitly targets these multimodal at-

tention mechanisms within the latent space, maximizing the discrepancy between the original and adversarial features.

The main contributions of this work are summarized as follows:

- We formulate proactive privacy protection for DiT-based I2V generation and propose I2VShield, a generative adversarial framework that protects reference images before they are used for unauthorized video synthesis.
- We develop a text-adaptive perturbation generator together with an untargeted Multimodal Attention Disruption (MAD) objective, enabling prompt-conditioned, ℓ_∞-bounded perturbation generation in a single forward pass while directly disrupting the multimodal conditioning pathway of DiT-based I2V models.
- We conduct extensive experiments on representative portrait and action-video benchmarks across multiple DiT-based I2V models, demonstrating superior protection effectiveness with substantially lower computational cost than gradient-based defenses.

2 Related Work

2.1 I2V Generation Models

Early video generation methods mainly relied on Generative Adversarial Networks (GANs) and autoregressive models, but often struggled with long-term temporal consistency and high-resolution synthesis [20, 26]. Diffusion models (DMs) subsequently improved generation quality and training stability [6, 15]. Models such as Video LDM and AnimateDiff extended pretrained text-to-image U-Nets with temporal attention or convolution layers to promote frame-to-frame coherence [1, 5]. These architectures combine strong image-generation priors with explicit temporal modeling.

More recently, Diffusion Transformers (DiTs) have increasingly been adopted as scalable alternatives to U-Net backbones. DiT-based systems such as Sora and Latte achieve strong physical realism and temporal consistency [2, 11, 13]. Modern I2V models align text prompts with reference images through latent self- and cross-attention. This dependence on multimodal feature alignment also creates a vulnerability: small disruptions to the conditioning pathway can substantially degrade the generated video, which motivates our approach.

2.2 Proactive Defenses and Adversarial Attacks

Proactive defenses protect visual content from unauthorized generative-AI processing by adding imperceptible adversarial perturbations. A series of studies focuse on preventing unauthorized customization. Glaze protects artistic styles from imitation, while Anti-DreamBooth disrupts subject-driven personalization of text-to-image models [18, 21]. Another series of studies target unauthorized generative editing. PhotoGuard protects images against inpainting manipulation, and I2VGuard extends such protection to I2V models by preventing reference images from being maliciously animated [4, 17]. These editing-oriented defenses typically rely on iterative, gradient-based methods such as PGD to optimize perturbations for each input image. For diffusion-based I2V models, repeatedly unrolling the denoising process and backpropagating across multiple video frames incur substantial computation and memory costs. They also require repeated access to the target model and its gradients during deployment.

In contrast, I2VShield learns a text-adaptive perturbation generator that produces protected images in a single forward pass, avoiding instance-specific optimization at inference time. Once trained, it retains prompt awareness without repeatedly accessing the target I2V model.

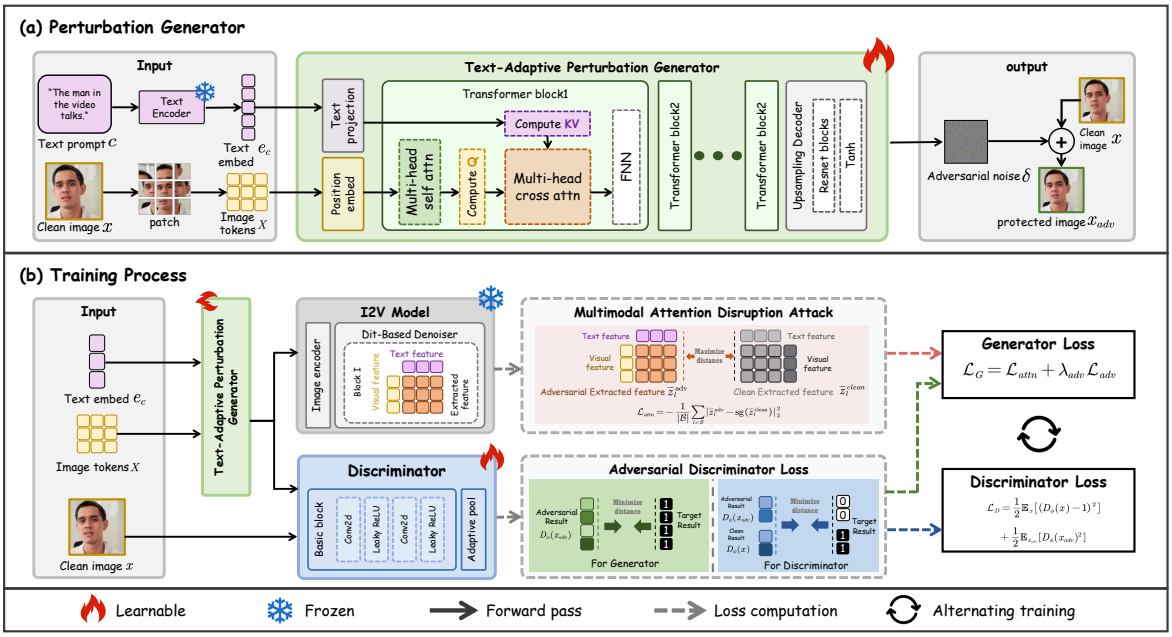

Figure 2: Illustration of the proposed I2VShield framework. (a) Perturbation generation pipeline, highlighting the architecture and key design of the Text-Adaptive Perturbation Generator. (b) Training pipeline of I2VShield, where the perturbation generator and discriminator are alternately optimized using the proposed loss functions.

3 Preliminaries

3.1 Threat Model

Attacker's Goal

We consider a malicious attacker who exploits an I2V generation model $\mathcal{F}$ to synthesize unauthorized content, such as deepfakes. Given a victim's reference image $x \in \mathbb{R}^{H\times W\times 3}$ and a text prompt c, the attacker generates a realistic and temporally coherent video $v \in \mathbb{R}^{F\times H'\times W'\times 3}$, where F denotes the number of frames:

$$v = \mathcal{F}(x, c). \tag{1}$$

Defender's Capabilities

We assume that the defender has white-box access to the target DiT-based I2V model only during offline training. The defender uses its architecture, pretrained weights, and gradients to train the perturbation generator. During online protection, the target I2V model and discriminator are no longer required; the defender encodes the prompt once and generates the protected image through a single forward pass of the perturbation generator, without accessing the target I2V model or computing its gradients.

Defender's Goal

Before publishing an image, the defender injects an imperceptible perturbation δ to obtain $x_{adv} = x + \delta$. When used as the visual condition of the I2V model, x_{adv} should cause spatial degradation,

temporal inconsistency, or semantic collapse in the generated video $v_{adv} = \mathcal{F}(x_{adv}, c)$. The objective is formulated as:

$$\max_{\delta} \mathcal{D}\big(\mathcal{F}(x+\delta, c), \mathcal{F}(x, c)\big) \tag{2}$$

$$\text{s.t.} \quad \|\delta\|_\infty \leq \epsilon, \quad x+\delta \in [0,1]^{H\times W\times 3}, \tag{3}$$

where $\mathcal{D}$ measures the discrepancy between adversarial and clean generations, and the ℓ_∞ constraint limits the perturbation magnitude, and ϵ means the perturbation budget.

3.2 Defense Methods

Traditional Methods

Existing proactive defenses commonly optimize δ through iterative gradient-based methods [10, 12]. For example, Iterative fast gradient sign method (I-FGSM) updates the adversarial image at step t as:

$$x_{adv}^{t+1} = \text{Clip}_{x,\epsilon}\left(x_{adv}^{t} + \alpha \cdot \text{sgn}\left(\nabla_x \mathcal{D}\big(\mathcal{F}(x_{adv}^{t}, c), \mathcal{F}(x, c)\big)\right)\right), \tag{4}$$

where α is the step size and $\text{Clip}_{x,\epsilon}$ constrains the result within the ϵ-ball around x. For I2V models, computing the input gradient requires backpropagating through the diffusion denoising process over multiple frames, resulting in prohibitive computation and GPU memory consumption.

Generative Defense Methods

To avoid instance-specific optimization at inference time, we introduce a parameterized perturbation generator G_θ. Instead of directly optimizing δ for each input, the generator parameters θ are learned over a training dataset:

$$\max_{\theta} \mathbb{E}_{x,c}\left[\mathcal{D}\big(\mathcal{F}(x + G_\theta(x, e_c), c), \mathcal{F}(x, c)\big)\right], \tag{5}$$

where e_c denotes the embedding of the text prompt c.

After offline training, G_θ produces an input-specific perturbation in a single forward pass. This transfers the computational cost to training and enables efficient $O(1)$ protection at inference. The key challenge is therefore to design an effective perturbation generator and suitable training objectives, as discussed in the following section.

4 Method

4.1 Overview

An overview of I2VShield is shown in Figure 2. Existing proactive defenses iteratively optimize image-specific perturbations, incurring substantial computational overhead. Moreover, their image-only objectives overlook the joint visual-text conditioning of modern DiT-based I2V models, making prompt-agnostic perturbations potentially ineffective.

I2VShield addresses these limitations by training a lightweight text-adaptive perturbation generator offline, enabling protection with a single forward pass. It combines an untargeted Multimodal Attention Disruption attack, which disrupts image-prompt interactions related to subject identity, semantic alignment, and temporal coherence, with discriminator-based regularization for visual fidelity. During training, the frozen I2V model processes clean and protected image-prompt pairs, while the generator maximizes their internal multimodal feature discrepancy under fidelity constraints. After training, the I2V model and discriminator are discarded, retaining only the perturbation generator and text encoder for online protection.

4.2 Framework of I2VShield

Existing proactive defenses typically rely on instance-wise iterative optimization, which incurs substantial computational and memory costs for I2V models with high-dimensional spatiotemporal latents. To improve deployment efficiency, I2VShield amortizes this process into a lightweight generator that produces adversarial perturbations in a single forward pass.

The generator is additionally conditioned on the text prompt because different prompts may activate distinct image–text interactions for the same reference image. Incorporating prompt embeddings therefore enables the generator to produce context-adaptive perturbations that more effectively disrupt prompt-specific generation.

Let E_{txt} denote the frozen text encoder associated with the target I2V model. For a prompt c, it produces

$$e_c = E_{\text{txt}}(c), \qquad e_c \in \mathbb{R}^{L_t \times d_e}, \tag{6}$$

where L_t is the number of text tokens and d_e is the text-embedding dimension.

At deployment time, given a clean reference image x and a text prompt c, the prompt is first encoded into a text embedding e_c. The trained generator G_θ then predicts the perturbation and produces the protected image:

$$\delta = G_\theta(x, e_c), \qquad x_{adv} = x + \delta. \tag{7}$$

The generator contains three lightweight components: image tokenization, text conditioning, and multimodal fusion with perturbation decoding.

Image tokenization

The input image is divided into non-overlapping patches through a convolutional patch embedding layer [3]:

$$X_0 = \text{PatchEmbed}(x) + P_{pos}, \qquad X_0 \in \mathbb{R}^{N_p \times d}, \tag{8}$$

where $N_p = (H/p)(W/p)$ is the number of patches, p is the patch size, d is the token dimension, and P_{pos} is a learnable positional embedding. This tokenized representation allows the generator to model long-range visual dependencies while keeping the architecture compact.

Text conditioning

The prompt embedding is projected into the same latent dimension as the image tokens:

$$T = \text{MLP}(e_c), \qquad T \in \mathbb{R}^{L_t \times d}, \tag{9}$$

where L_t denotes the number of text tokens, and MLP denotes the multilayer perceptron. This projection transforms the textual condition into a form that can be effectively fused with visual tokens.

Multimodal fusion and perturbation decoding

The visual tokens are processed by a stack of lightweight transformer blocks. Within each block, self-attention first captures global visual context by enabling interactions between distant image regions. Cross-modal interaction is then introduced by using the projected text tokens as semantic guidance, enabling the visual representation to adapt to the prompt-dependent generation scenario. Finally, feed-forward layers refine the fused representation and produce feature patterns that are effective for disrupting the I2V model.

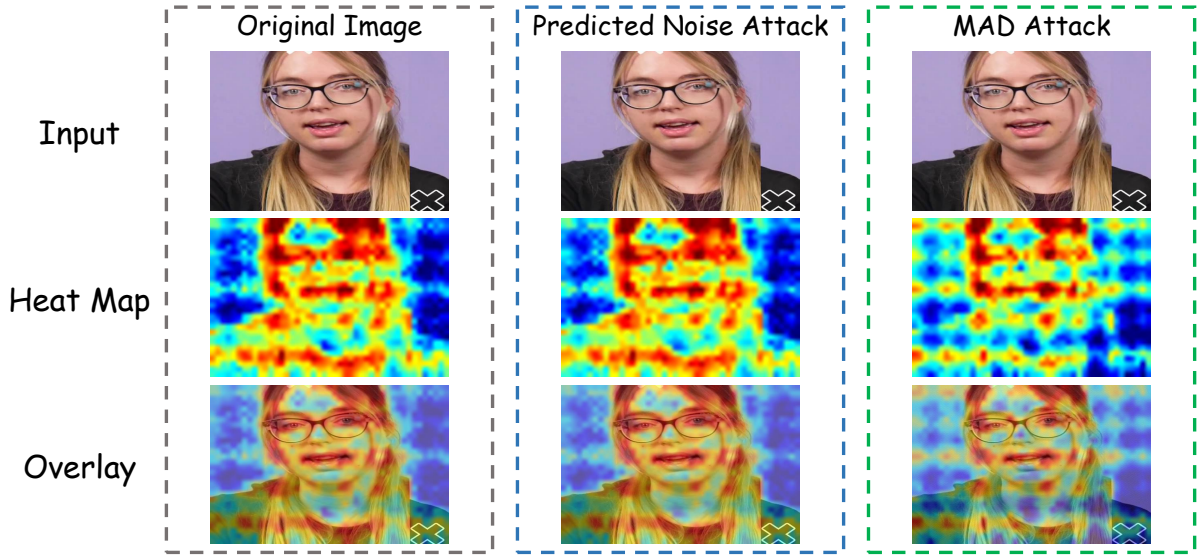

Figure 3: Multimodal attention maps obtained from a clean image and images protected using the predicted noise attack and the Multimodal Attention Disruption (MAD) attack.

After the final transformer block, the fused tokens are reshaped into a spatial feature map and passed through a cascaded upsampling decoder. The decoder predicts a dense perturbation map with the same spatial resolution as the input image. Because the entire process is feed-forward, I2VShield avoids iterative gradient computation during deployment and provides efficient proactive protection.

4.3 Multimodal Attention Disruption Attack

DiT-based I2V models use multimodal attention to fuse reference-image identity and appearance with textual semantic and motion cues. As shown in Figure 3, we compare the attention maps of the I2V model for adversarial images generated by attacking the predicted noise and cross-attention features, respectively. Attacking the predicted noise only slightly changes the attention distribution, whereas attacking cross-attention features substantially redistributes attention away from identity-relevant regions. This suggests that directly disrupting attention features destabilizes multimodal conditioning more effectively than perturbing noise prediction.

Motivated by this observation, we propose the untargeted Multimodal Attention Disruption (MAD) attack that maximizes the discrepancy between clean and adversarial cross-attention features. By corrupting cross-modal associations, MAD degrades subject fidelity, prompt alignment, and temporal coherence without requiring a predefined adversarial target concept or target video.

Formally, let $\mathcal{F}$ denote the frozen DiT-based I2V model. During training, we sample a diffusion timestep $\tau \sim \mathcal{U}\{1, \ldots, T_d\}$ and construct the same noisy latent $\xi \sim \mathcal{N}(0, I)$ for the clean and adversarial branches. Given a selected set of transformer blocks $\mathcal{B}$, we extract the multimodal attention features z_l from each block. The clean and adversarial feature states are denoted as:

$$z_l^{\text{clean}} = \Phi_l(x, c, \tau, \xi), \qquad z_l^{\text{adv}} = \Phi_l(x_{adv}, c, \tau, \xi), \quad l \in \mathcal{B}, \tag{10}$$

where $\Phi_l(\cdot)$ denotes the multimodal attention feature extractor at the l-th block of $\mathcal{F}$. Both branches share the same prompt c, timestep τ, and diffusion noise state ξ, ensuring that the measured discrepancy is caused by the injected perturbation rather than by stochastic variation.

To reduce the influence of layer-wise scale differences, we normalize the extracted features:

$$\bar{z}_l^b = \frac{z_l^b}{\|z_l^b\|_2 + \eta}, \qquad b \in \{\text{clean}, \text{adv}\}, \tag{11}$$

where η is a small constant for numerical stability. The MAD loss is then defined as:

$$\mathcal{L}_{attn} = -\frac{1}{|\mathcal{B}|} \sum_{l \in \mathcal{B}} \left\| \bar{z}_l^{\text{adv}} - \text{sg}\left(\bar{z}_l^{\text{clean}}\right) \right\|_2^2, \tag{12}$$

where $\text{sg}(\cdot)$ denotes the stop-gradient operation. Minimizing $\mathcal{L}_{attn}$ maximizes the feature discrepancy between the clean and protected branches. Since the clean branch is detached, the generator is encouraged to move the adversarial representation away from the original multimodal attention state without altering the I2V model itself.

This untargeted objective has two advantages. First, it does not require a target video or a predefined adversarial semantic concept, which makes the training process simple and broadly applicable. Second, it attacks the feature-level alignment mechanism that is essential for DiT-based I2V generation, thereby inducing degradation in subject consistency, prompt consistency, and temporal coherence simultaneously.

4.4 Discriminator and Visual Fidelity Regularization

Strong feature disruption may introduce structured or high-frequency artifacts that remain noticeable even under an ℓ_∞ constraint, reducing the practical usability of protected images. We therefore introduce a PatchGAN-style discriminator D_ϕ to regularize local texture patterns and encourage the protected images to remain perceptually natural [9].

This adversarial regularization complements the norm constraint: the latter limits perturbation magnitude, while the discriminator constrains its spatial and perceptual distribution, thereby balancing attack effectiveness and visual fidelity.

We adopt the Least Squares GAN objective for stable optimization. The discriminator loss is defined as:

$$\mathcal{L}_D = \frac{1}{2}\mathbb{E}_x \left[(D_\phi(x) - 1)^2 \right] + \frac{1}{2}\mathbb{E}_{x_{adv}} \left[D_\phi(x_{adv})^2 \right]. \tag{13}$$

The generator-side adversarial loss is:

$$\mathcal{L}_{adv} = \mathbb{E}_{x_{adv}} \left[(D_\phi(x_{adv}) - 1)^2 \right]. \tag{14}$$

The final generator objective combines the MAD loss with the visual fidelity regularization:

$$\mathcal{L}_G = \mathcal{L}_{attn} + \lambda_{adv}\mathcal{L}_{adv}, \tag{15}$$

where λ_{adv} controls the trade-off between adversarial disruption and visual fidelity.

During training, D_ϕ and G_θ are optimized alternately. The discriminator learns to distinguish clean images from protected images, while the generator learns to both fool the discriminator and disrupt the multimodal attention features of the frozen I2V model. After training converges, the discriminator and target I2V model are removed, while the trained G_θ and the frozen E_{txt} are retained for online image protection.

Table 1: Average comparative experimental results on CelebV-Text and UCF101. "-" denotes not applicable.

Method	VRAM (GB)	TFLOPs	VBench				Q-Align	Gemini-3.1-flash-lite			
			Sub. Cons.	Bg. Cons.	Mot. Smooth	Img. Quality	Vis. Score	Prompt Cons.	Temp. Cons.	Mot. Plaus.	Frame Qual.
CogVideoX-5B											
Clean	-	-	0.9105	0.9343	0.9786	0.4981	0.5409	4.9400	3.9700	4.8000	3.7800
Random Noise	-	-	0.9093	0.9329	0.9782	0.4885	0.5299	4.9400	3.9300	4.7900	3.7000
PhotoGuard	22.42	1438.02	0.9016	0.9253	0.9778	**0.4865**	**0.5171**	4.8900	**3.8400**	**4.6900**	**3.6400**
I2VShield (Ours)	**9.49**	**2.23**	**0.8962**	**0.9248**	**0.9756**	0.4978	0.5216	**4.8700**	3.9200	4.7800	3.7450
Wan2.1-14B											
Clean	-	-	0.8781	0.9195	0.9673	0.5224	0.5196	**4.8600**	3.9300	4.8000	3.6100
Random Noise	-	-	0.8653	0.9170	0.9683	0.5281	0.5186	4.9400	3.9600	4.8700	3.6500
PhotoGuard	40.92	2133.51	0.8379	0.9013	0.9601	0.4957	0.4372	4.9200	3.8600	4.7500	3.5200
I2VShield (Ours)	**12.27**	**4.95**	**0.8292**	**0.8968**	**0.9556**	**0.4895**	**0.4304**	4.9300	**3.8300**	**4.6800**	**3.3600**
OpenSora-V2-11B											
Clean	-	-	0.9237	0.9495	0.9905	0.5428	0.5242	4.8800	3.9800	4.8600	3.7400
Random Noise	-	-	0.9241	0.9498	0.9896	0.5396	0.5205	4.8800	**3.9600**	4.8800	3.7100
PhotoGuard	34.73	377.50	0.9267	0.9496	0.9888	**0.5343**	**0.4901**	4.9000	4.0300	4.8500	**3.6700**
I2VShield (Ours)	**9.79**	**14.59**	**0.9209**	**0.9474**	**0.9883**	0.5434	0.5033	**4.8500**	4.0100	**4.8000**	**3.6700**

5 Experiments

5.1 Experimental Setup

Datasets

We evaluate I2VShield on two datasets: UCF101 [19] for complex human action generation, and CelebV-Text [27] for facial portrait animation. For both datasets, we randomly sample 900, 100, and 50 instances for training, validation, and testing, respectively.

Target Models and Baselines

We assess the defensive capability of I2VShield against three state-of-the-art DiT-based I2V models: CogVideoX-5B [7], OpenSora-V2-11B [28], and Wan2.1-14B [22]. Consistent with our threat model, we assume a white-box setting strictly during the offline training phase. We compare I2VShield against three baselines: (1) *Clean* (original, unperturbed images); (2) *Random Noise* (uniformly distributed noise under the same ℓ_∞ budget); and (3) *PhotoGuard*, a representative gradient-based defense.

Evaluation Metrics

Because more effective protection should cause greater degradation in unauthorized generations, lower scores generally indicate stronger protection. We employ VBench [8] to objectively assess spatial fidelity and temporal coherence (specifically subject/background consistency, motion smoothness, and image quality); Q-Align [24] for blind visual quality assessment; and a Vision-Language Model (VLM)[1] to evaluate semantic dimensions, including prompt consistency, temporal consistency, motion plausibility, and frame quality.

5.2 Main Results

Quantitative Results

As shown in Table 1, when averaged across CelebV-Text and UCF101, I2VShield provides highly competitive protection across all evaluated DiT-based I2V models. Compared with PhotoGuard,

[1]We use Gemini-3.1-flash-lite as an automated judge for scalable, zero-shot video assessment.

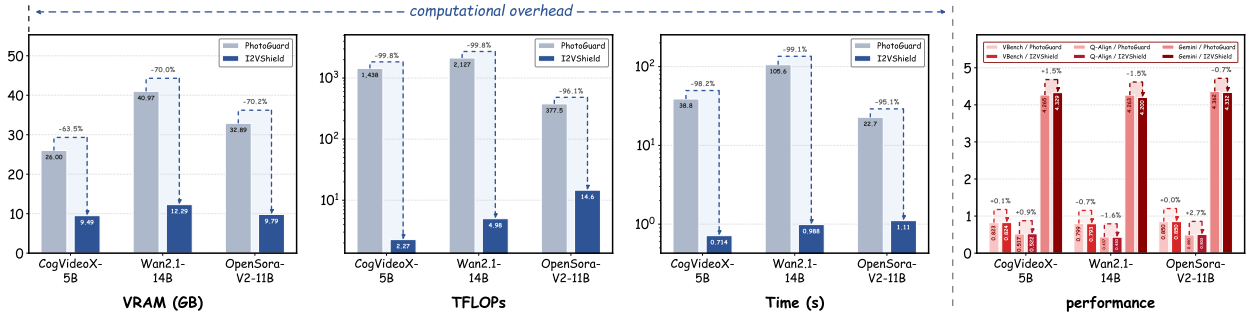

Figure 4: Comparison of the computational overhead and performance of PhotoGuard and I2VShield. The proposed I2VShield significantly reduces computational overhead while achieving defense performance comparable to the baseline.

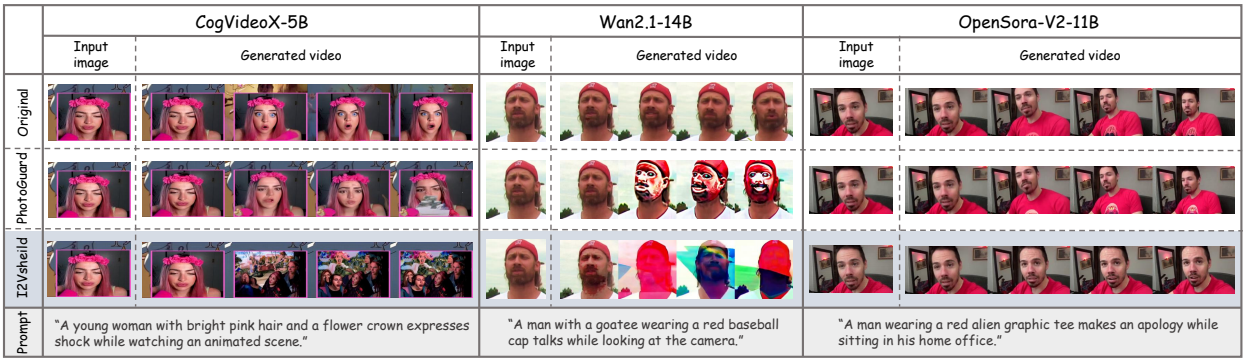

Figure 5: Qualitative comparison between the proposed method and the baseline. Our I2VShield produces more severe degradation in both visual appearance and temporal consistency.

I2VShield consistently achieves lower Subject Consistency, Background Consistency, and Motion Smoothness scores on CogVideoX-5B, Wan2.1-14B, and OpenSora-V2-11B, demonstrating stronger disruption of the generated videos' spatiotemporal coherence. The Gemini-based VLM evaluation reveals complementary, model-dependent effects: I2VShield obtains lower Prompt Consistency on CogVideoX-5B and OpenSora-V2-11B, while achieving lower Motion Plausibility on Wan2.1-14B and OpenSora-V2-11B. Moreover, I2VShield achieves these competitive protection effects with substantially lower VRAM consumption and computational cost than PhotoGuard across all three models, highlighting its efficiency and practical scalability.

Figure 4 further compares the per-image online protection costs and performance of I2VShield and PhotoGuard. The reported results exclude the offline training cost of I2VShield. We train a separate perturbation generator for each target model to accommodate its input resolution and text-conditioning interface, resulting in different online costs across models. PhotoGuard requires iterative forward and backward passes through the target I2V model for every input image, whereas I2VShield requires only text encoding and a single forward pass through the corresponding perturbation generator. Consequently, I2VShield substantially reduces peak GPU memory consumption, floating-point operations, and protection time across the evaluated settings. These results demonstrate the deployment-time efficiency of I2VShield after the perturbation generator has been trained.

Table 2: Ablation results of text-conditioned perturbation generation.

Perturbation Generator Input		Video Quality Degradation (↓)		
Visual Feature	Text Embedding	VBench-IQ	Q-Align	Gemini-MP
✓		0.5755	**0.5524**	4.6600
✓	✓	**0.5746**	0.5585	**4.5200**

Qualitative Results

Figure 5 presents a qualitative comparison of videos generated from clean reference images and images protected by PhotoGuard or I2VShield. When conditioned on clean images, all target I2V models consistently synthesize temporally coherent videos while preserving high subject fidelity.

Multimodal Attention Disruption (MAD), employed by I2VShield, disrupts both frame-level appearance and the temporal propagation of identity- and prompt-relevant cues. Consequently, all three evaluated models exhibit persistent subject degradation and temporal inconsistencies, although the specific manifestations vary across architectures. For CogVideoX-5B, the generated sequence progressively deviates from the reference subject and eventually drifts toward unrelated visual content. For Wan2.1-14B, facial structures become distorted and color fidelity deteriorates across consecutive frames, accompanied by abrupt inter-frame variations. For OpenSora-V2-11B, the intended facial motion is markedly suppressed, resulting in repetitive frames with limited semantic progression. In comparison, PhotoGuard only partially disrupts the generated content: although localized facial artifacts emerge, the overall identity, background, and motion trajectory remain recognizable in several cases. The consistent failure patterns induced by MAD across all three architectures demonstrate that it generalizes beyond model-specific artifact patterns and provides similarly effective protection against the evaluated I2V models.

5.3 Ablation Study

We evaluate three key designs of I2VShield on CelebV-Text: (1) text-adaptive perturbation generation, which incorporates the text embedding as an additional input to the perturbation generator, (2) the optimization target (our MAD attack vs. standard predicted noise attack), and (3) the PatchGAN discriminator ($\mathcal{L}_{adv}$) for fidelity regularization.

Effectiveness of Text-adaptive Perturbation Generation

As shown in Table 2, VBench-IQ and Q-Align measure visual quality using VBench and Q-Align, respectively, while Gemini-MP evaluates motion smoothness with Gemini; lower scores indicate stronger degradation. Adding text embeddings reduces VBench-IQ and Gemini-MP, demonstrating improved degradation of both visual quality and motion smoothness. Although Q-Align slightly increases slightly, the two variants remain broadly comparable on this metric. Overall, text conditioning provides complementary semantic guidance for generating more effective prompt-aware perturbations.

Table 3: Ablation results of attention disruption

Attack Target	Video Quality Degradation (↓)		
	VBench-IQ	Q-Align	Gemini-MP
Predicted Noise	0.5749	0.5677	4.5600
Attention Disruption	**0.5746**	**0.5585**	**4.5200**

Table 4: Fidelity comparison of protected reference images.

Loss Function	Image Fidelity Metrics		
	LPIPS (↓)	PSNR (↑)	SSIM (↑)
w/o $\mathcal{L}_{adv}$	0.3278	30.5159	0.6814
w/ $\mathcal{L}_{adv}$	**0.3191**	**32.7229**	**0.7885**

Effectiveness of MAD Attack

As demonstrated in Table 3, targeting internal cross-attention features achieves more effective video degradation than perturbing the predicted noise space. Our full framework consistently outperforms the predicted noise attack across all evaluation metrics, with a particularly notable improvement on Q-Align, whose visual-quality score decreases from 0.5677 to 0.5585. These results suggest that perturbing cross-attention features more effectively disrupts multimodal feature alignment, leading to accumulated distortions in both visual appearance and temporal consistency throughout the video generation process.

Contribution of Discriminator Regularization

Table 4 and Figure 6 illustrate the impact of the PatchGAN discriminator. Without $\mathcal{L}_{adv}$, the MAD-only variant introduces noticeable structural and high-frequency artifacts. Incorporating $\mathcal{L}_{adv}$ improves PSNR by 7.2%, reduces the LPIPS score by 2.7%, and yields a more pronounced 15.7% gain in SSIM, indicating that the discriminator is particularly effective in preserving structural information. These improvements suggest that adversarial supervision regularizes perturbation generation, encouraging texture-adaptive perturbations that better align with the underlying image distribution while maintaining the protection capability.

6 Conclusions

We propose I2VShield, an efficient proactive privacy defense against DiT-based I2V models. Unlike costly iterative gradient-based methods, I2VShield uses a lightweight generative framework to produce text-adaptive adversarial perturbations in a single forward pass. Its untargeted MAD

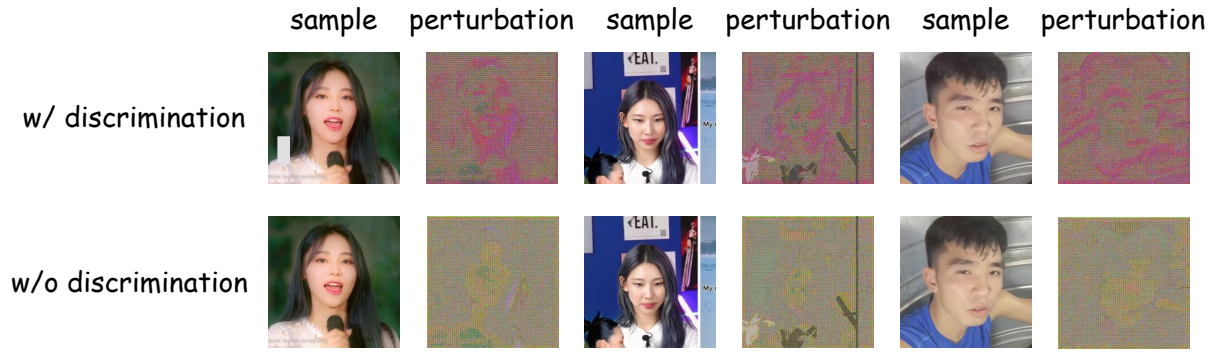

Figure 6: Effect of discriminator regularization on protected image fidelity.

attack maximizes feature deviation within DiT cross-attention layers, weakening image-text conditioning. Experiments across multiple I2V models demonstrate competitive protection performance with substantially lower inference time and memory consumption, supporting practical large-scale protection.

References

[1] 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*, pages 22563–22575, 2023.

[2] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Leo Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. Video generation models as world simulators. *OpenAI Blog*, 1(8):1, 2024.

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[4] 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*, pages 12595–12604, 2025.

[5] 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. *arXiv preprint arXiv:2307.04725*, 2023.

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

[7] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022.

[8] 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*, pages 21807–21818, 2024.

[9] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[10] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*, 2016.

[11] Xin Ma, Yaohui Wang, Xinyuan Chen, Gengyun Jia, Ziwei Liu, Yuan-Fang Li, Cunjian Chen, and Yu Qiao. Latte: Latent diffusion transformer for video generation. *arXiv preprint arXiv:2401.03048*, 2024.

[12] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

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

[14] Zuomin Qu, Wei Lu, Xiangyang Luo, Qian Wang, and Xiaochun Cao. Id-guard: A universal framework for combating facial manipulation via breaking identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(2):1720–1735, 2026.

[15] 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*, pages 10684–10695, 2022.

[16] 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*, pages 1–11, 2019.

[17] Hadi Salman, Alaa Khaddaj, Guillaume Leclerc, Andrew Ilyas, and Aleksander Madry. Raising the cost of malicious ai-powered image editing. *arXiv preprint arXiv:2302.06588*, 2023.

[18] 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)*, pages 2187–2204, 2023.

[19] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[20] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. Mocogan: Decomposing motion and content for video generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1526–1535, 2018.

[21] Thanh Van Le, Hao Phung, Thuan Hoang Nguyen, Quan Dao, Ngoc N Tran, and Anh Tran. Anti-dreambooth: Protecting users from personalized text-to-image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2116–2127, 2023.

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

[23] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8695–8704, 2020.

[24] Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, et al. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090*, 2023.

[25] Wenbo Xu, Junyan Wu, Wei Lu, Xiangyang Luo, and Qian Wang. A multimodal deviation perceiving framework for weakly-supervised temporal forgery localization. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 11581–11589, 2025.

[26] Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. Videogpt: Video generation using vq-vae and transformers. *arXiv preprint arXiv:2104.10157*, 2021.

[27] Jianhui Yu, Hao Zhu, Liming Jiang, Chen Change Loy, Weidong Cai, and Wayne Wu. Celebv-text: A large-scale facial text-video dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14805–14814, 2023.

[28] 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.