I2VGuard: Safeguarding Images against Misuse in Diffusion-based Image-to-Video Models

Anonymous CVPR submission

Paper ID 4510

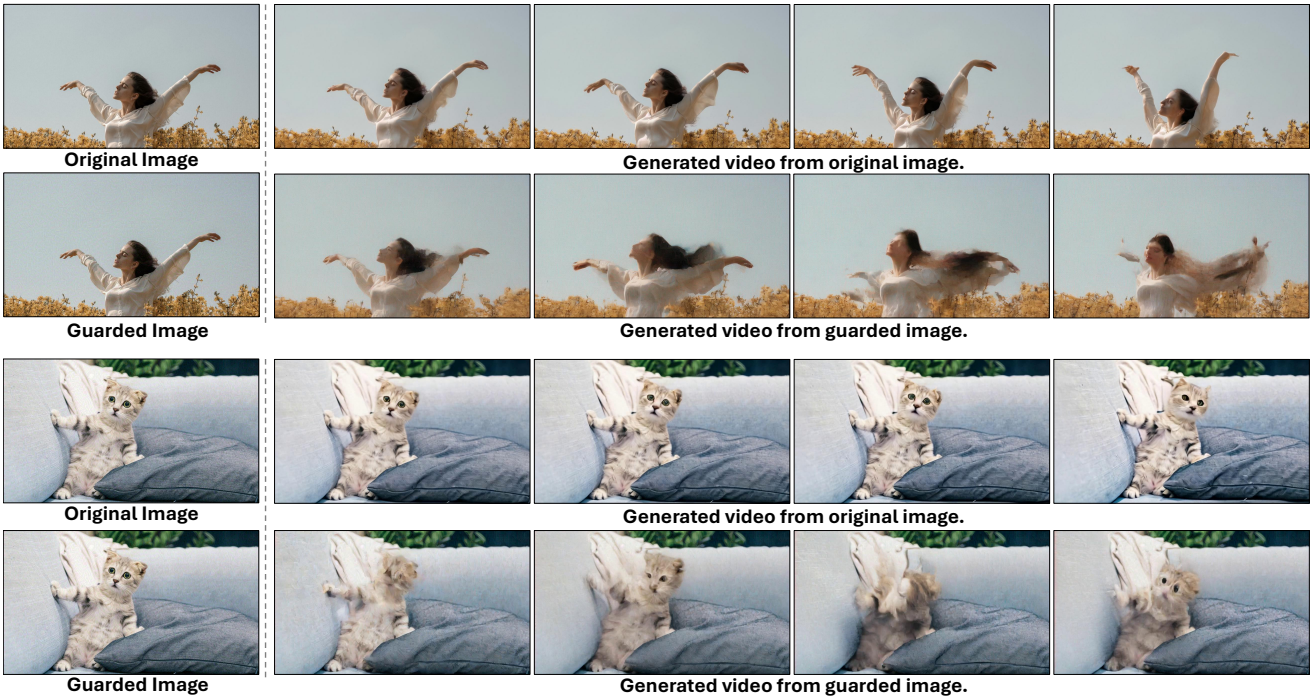


Figure 1. **Results of our I2VGuard**. We present original images, guarded images, and their corresponding SVD-generated videos. All results are generated with the same seed. Our method effectively safeguards images from animation in image-to-video generation.

Abstract

Recent advances in image-to-video generation have enabled animation of still images and offered pixel-level controllability. While these models hold great potential to transform single images into vivid and dynamic videos, they also carry risks of misuse that could impact privacy, security, and copyright protection. This paper proposes a novel approach that applies imperceptible perturbations on images to degrade the quality of the generated videos, thereby protecting images from misuse in white-box image-to-video diffusion models. Specifically, we function our approach as an adversarial attack, incorporating spatial, temporal, and diffusion attack modules. The spatial attack shifts image features from their original distribution to a lower-quality target distribution, reducing visual fidelity. The temporal attack disrupts coherent motion by interfering with temporal attention maps that guide motion generation. To enhance the robustness of our approach across different models, we further propose a diffusion attack module leveraging contrastive loss. Our approach can be easily integrated with mainstream diffusion-based I2V models. Extensive experiments on SVD, CogVideoX, and ControlNeXt demonstrate that our method significantly impairs generation quality in terms of visual clarity and motion consistency, while introducing only minimal artifacts to the images. To the best of our knowledge, we are the first to explore adversarial attacks on image-to-video generation for security purposes.

1. Introduction

Diffusion models [14, 18, 33] have gained significant attention recently, particularly in video generation. Advances in models like Sora [5] and its DiT [31] framework have accel-

erated the development of video diffusion models, leading to numerous academic [4, 21, 25, 51, 53] and commercial [41–43] applications. The primary types of video generation are text-to-video(T2V) and image-to-video(I2V), with the former supporting greater diversity and the latter offering pixel-level control. Leveraging powerful generative models and techniques from text-to-image generation [33], image-to-video models can now produce visually impressive videos. Other conditional I2V models include CogVideoX [51], which generates videos using both text and image inputs, enabling more versatile content creation, and ControlNeXt [32], which focuses on generating videos conditioned on pose information, offering precise control over motion and structure.

However, the rapid advancement of generative models has also heightened risks related to privacy, security, and copyright [2]. For instance, an adversary with access to a person's image could use image-to-video models to generate misleading or harmful videos, depicting actions the individual never performed, thus compromising individual's privacy and security. Additionally, the unauthorized use of copyrighted images to create animations poses significant risks to intellectual property rights. To address these risks, this work introduces an adversarial attack method designed to prevent the misuse of images in diffusion-based I2V models. Prior research has primarily focused on protecting images from malicious edits [35, 52], which are often viewed as adversarial attacks on image-editing models. To our knowledge, this is the first adversarial attack method aimed at securing images against misuse in I2V generation.

The main challenges in attacking I2V models come from two aspects: first, disrupting temporal consistency with perturbations applied only to individual images; and second, achieving robustness across diverse diffusion models, including UNet and DiT architectures, even under strong conditioning such as poses and text prompts. To address these challenges, our method targets different modules within I2V models and affects various attributes of the generated video. Specifically, we divide our approach into spatial, temporal, and diffusion attacks. The spatial attack focuses on altering the spatial features of the I2V models by targeting the VAE encoder in latent diffusion models. Here, we manipulate image latents, shifting them toward suboptimal values. The temporal attack aims to perturb the temporal features, which are crucial for video motion and consistency. By extracting the temporal attention map, we intentionally disrupt it by pushing it away from its original attention map. This creates a chaotic attention map, leading to temporal inconsistencies in the generated video. The diffusion attack focuses on the diffusion process and targets the output of the denoising module, which includes the UNet in SVD and the transformer in CogVideoX. Unlike a standard denoising process, our approach aims to push the predicted outputs away from the original frames, guiding them closer to target frames. To achieve this, we design a contrastive loss that degrades the quality of the predicted frames by treating the target frames as positive samples and the original predicted frames as negative samples.

We conduct a comprehensive qualitative and quantitative analysis of the attack results. Qualitative analysis demonstrates that our method effectively disrupts the original image, leading to low temporal consistency and reduced aesthetic quality in both simple and conditioned I2V models. Quantitative analysis confirms a significant degradation in both spatial quality and temporal consistency of the generated videos.

2. Related Work

Image-to-Video Generation With the great advancements in image generation [14, 18, 29, 33, 38, 39], video generation [3, 5, 19, 21, 37, 41–43, 50, 51] has gathered significant attention and is developing rapidly [17, 46, 47, 49]. Recently, following the success of diffusion models [14, 18, 33, 38] in image generation, diffusion models integrated with UNet [3, 13, 20, 34] or transformers (DiT) [5, 31, 45, 53] have been adopted for video generation. In the field of image-to-video generation, AnimateDiff [17] animates personalized text-to-image diffusion models by inserting a temporal motion module, Stable Video Diffusion(SVD) [4] demonstrates strong performance, benefiting from intensive training on large amounts of high-quality data, CogVideoX [51] supports text-image jointly conditional video generation, Animate-Anyone [22] and ControlNeXt [32] leverage pose information for controllable image-to-video generation. In this work, we focus on I2V models, which are susceptible to misuse and therefore require robust protection.

Adversarial Attacks on Diffusion Models Current attacks on diffusion models, which can be broadly categorized [44] into backdoor [8, 11] and adversarial attacks, primarily target image-based diffusion models [12, 15, 16, 48]. In adversarial attacks, the input image or text prompt is perturbed to generate adversarial examples. For instance, AdvDM [28] and Mist [27] utilize adversarial examples to protect human-created artworks, PhotoGuard [35] alters the image by modifying either the encoder or the diffusion outputs against malicious editing, while Glaze [36] perturbs the image to prevent style mimicry, [52] shifts the image away from its original distribution, [54] embeds personal watermarks in the generation of adversarial examples, and DiffusionGuard [10] generates adversarial noise targeting the early stage of the diffusion process. For video-based attacks, PRIME [26] modifies videos frame-by-frame to prevent malicious video editing, while [30] disrupts style mimicry attacks from video imagery. There are also works protect the image against other generation, such as text-to-3D generation [40]. In this work, we focus on perturbing the

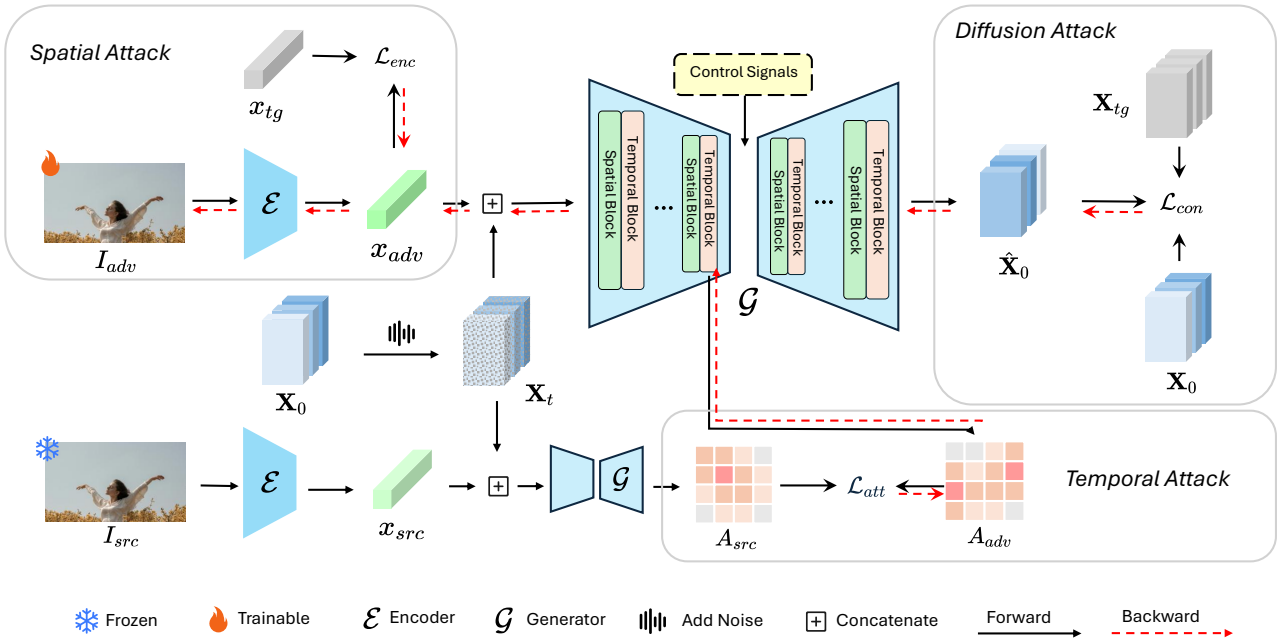


Figure 2. **Overview of our proposed method**. The training process begins with a trainable copy I_{adv} of the original image I_{src}. First, we perform inference to obtain the original video V_0 and get the latent frames $\mathbf{X}_0$. Noisy latent frames $\mathbf{X}_t$ and the latent image x_{adv}, x_{src} are then processed by the denoising model to predict the original frames. The encoded latent image x_{adv} is utilized for the spatial encoder attack, while the predicted original frames $\hat{\mathbf{X}}_0$ serve in the diffusion attack to compute contrastive loss. Within the denoising module, we hook into the temporal attention module to extract the temporal attention map. By altering the attacked attention map A_{adv} to diverge from the original A_{src}, we implement the temporal attack.

input image in image-to-video diffusion models to perform adversarial attacks.

3. Method

In this section, we first provide preliminaries on latent diffusion models and problem settings in Sec. 3.1. We then present our attack method, detailing the spatial attack in Sec. 3.2, the temporal attack in Sec. 3.3, and the diffusion attack in Sec. 3.4. An overview of our method is presented in in Fig. 2.

3.1. Preliminary

Latent Diffusion Models In diffusion models, realistic image or video sampling from a target distribution $q(\cdot)$ is achieved via a sequence of denoising steps in a *diffusion process*. This approach involves iteratively adding and removing noise to transform an initial sample $\mathbf{x}_0 \sim q(\cdot)$. Noise is incrementally introduced across steps, creating intermediate states defined by $\mathbf{x}_{t+1} = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon_t$, where ϵ_t is Gaussian noise. The parameters α_t control the noise transition, leading to a final noisy state with a standard Gaussian distribution. By reversing this process and iteratively denoising each state, the model learns to predict the noise added at each step, ultimately enabling thr recovery of samples from $q(\cdot)$ starting from $\mathbf{x}_T$. Latent diffusion models (LDMs), a subclass of diffusion models, operate in a compressed latent space rather than directly in the image space. In LDMs, each input image or video $\mathbf{x}_0$ is first mapped to a latent representation $\mathcal{E}(\mathbf{x}_0)$ via an encoder $\mathcal{E}$. The diffusion process then applies in this latent space, producing a series of noisy representations. Training in this latent space enables LDMs to achieve efficient training and faster inference while retaining high-quality outputs.

Problem Settings Adversarial attacks involve adding subtle, nearly undetectable perturbations δ_{adv} to the input data to influence a model's output. In computer vision, adversarial attacks have been widely explored in classifications and image synthesis. However, in image-to-video generation tasks, adversarial attack remains a relatively unexplored area. In this work, we aim to develop adversarial attacks to safeguard images from misused by image-to-video latent diffusion models. To distinguish between the latent and pixel spaces, we represent images and videos in pixel space as I and V, and in latent space as x and $\mathbf{X}$, respectively.

Specifically, an adversary can generate a modified image $I_{adv} = I_{src} + \delta$ from source image I_{src}, crafted to disrupt the generative model $\mathcal{G}$, resulting in poor-quality or distorted

Figure 3. **Temporal self-attention map visualizations**. Difference between generated frames from the original image (left) and the guarded image (right) in Fig. 1.

outputs. This adversarial image is optimized by maximizing the loss between the videos generated from the original and modified images, represented as follows:

$$\delta_{adv} = \arg\max_{\|\delta\|_p \leq \epsilon} \mathcal{L}(\mathcal{G}(I_{src} + \delta, \mathbf{X}_t, c), \mathcal{G}(I_{src}, \mathbf{X}_t, c))$$

where δ_{adv} represents the computed perturbation, constrained by $\|\delta\|_p \leq \epsilon$ to ensure minimal visibility, and c represents other conditions such as text or pose in the image-to-video generation process. The constraint on δ_{adv} is maintained by a regularization term $||I_{adv} - I_{src}||^2$.

3.2. Spatial Attack

The generated video has two main attributes: spatial features and temporal features. In this section, we focus on attacking the spatial features within individual frames. Previous work [35, 52] has applied adversarial attacks to image editing tasks, focusing solely on spatial features. Our work extends these methods to image-to-video generation task.

In Latent Diffusion Models (LDMs), an encoder, denoted as $\mathcal{E}$, compresses the original image or frames into a latent space representation. An encoder attack aims to compromise this encoder, reducing the effectiveness of generative models.

$$\mathcal{L}_{enc} = \|\mathcal{E}(I_{adv}) - \mathcal{E}(I_{tg})\|^2 \quad (1)$$

Specifically, as shown in Eq. (1), the attack transforms the latent representation $\mathcal{E}(I_{adv})$ into a poor one $\mathcal{E}(I_{tg})$, resulting in poor-quality final videos.

3.3. Temporal Attack

Unlike attacks on image editing tasks, where no temporal features are involved, attacks on video generation rely heavily on temporal information. In this task, disrupting temporal consistency through adversarial attacks is both essential and challenging. To reduce temporal consistency and induce chaotic behavior in the outputs, we target the temporal attention maps within the temporal transformer blocks in video diffusion generator, $\mathcal{G}$. Specifically, during training, we insert a hook, denoted as $\underset{A_{adv}}{\mathcal{G}}$, within the attention block to extract the self-attention map A_{adv}. We then apply a targeted attack to this attention map, pushing it away from the original map, A_{src}, thereby reducing temporal consistency in the generated outputs. To achieve this, we introduce a temporal attention loss, denoted as $\mathcal{L}_{att}$, as shown in Eq. (2), which induces motion-related perturbations and promotes temporal chaos.

$$\mathcal{L}_{att} = \tau_2 - ||A_{adv} - A_{src}||^2 \quad (2)$$

As shown in Fig. 3, the visualization of the temporal self-attention map on both the original and modified frames highlights this impact: the manipulated attention map misaligns with the original, directly leading to inconsistencies in motion across frames. This visual evidence underscores the effectiveness of our method in disrupting temporal coherence in video generation by distorting the temporal attention mechanism within the model.

Algorithm 1 Adversarial Attack on Image-to-Video Generation

Input: Source Image I_{src} and Trainable Copy I_{adv}, Target Image I_{tg}, Target Video V_{tg}, Image-to-Video Pipeline $\mathcal{P}$, Generator $\mathcal{G}$, Encoder $\mathcal{E}$, Diffusion Scheduler $\mathcal{S}$
(Optional:Condition c)
Hyperparameters $\tau_1, \tau_2, \alpha, \beta, \gamma, \lambda$
Encode image $x_{src}, x_{tg} = \mathcal{E}(I_{src}), \mathcal{E}(I_{tg})$
Generate original video $V_0 = \mathcal{P}(I_{src})$
Encode frames $\mathbf{X}_0, \mathbf{X}_{tg} = \mathcal{E}(V_0), \mathcal{E}(V_{tg})$
for each iteration **do**
  Encode attacked image $x_{adv} = \mathcal{E}(I_{adv})$
  Compute encoder loss: $\mathcal{L}_{enc} = ||x_{adv} - x_{tg}||^2$
  Generate noisy frames $\mathbf{X}_t = \mathcal{S}(\mathbf{X}_0)$
  Predict $\mathbf{X}_0$ $\begin{cases} \hat{\mathbf{X}}_{0,adv} = \underset{A_{adv}}{\mathcal{G}}(\mathbf{X}_t, x_{adv}, c) \\ \hat{\mathbf{X}}_{0,src} = \underset{A_{src}}{\mathcal{G}}(\mathbf{X}_t, x_{src}, c) \end{cases}$
  Compute spatial contrastive loss

$$\mathcal{L}_{con} = ||\hat{\mathbf{X}}_{0,adv} - \mathbf{X}_{tg}||^2 + \max\left(0, \tau_1 - ||\hat{\mathbf{X}}_{0,adv} - \mathbf{X}_0||^2\right)$$

  Compute temporal attention loss

$$\mathcal{L}_{att} = \tau_2 - ||A_{adv} - A_{src}||^2$$

  Compute the final loss
  $\mathcal{L} = ||I_{adv} - I_{src}||^2 + \alpha \cdot \mathcal{L}_{enc} + \beta \cdot \mathcal{L}_{con} + \gamma \cdot \mathcal{L}_{att}$
  Update parameters $I_{adv} \leftarrow I_{adv} - \lambda \cdot \nabla_I \mathcal{L}$
end for

3.4. Diffusion Attack

Although spatial and temporal attacks can disrupt structural and motion information in generated videos, their effectiveness may vary significantly across different models [4, 32, 51], especially when other conditions, *e.g.*, text prompts, are involved. To enhance the robustness and generalization of the perturbation across diffusion models, we introduce a latent space perturbation that uses contrastive loss to shift the predicted noise closer to the target frames

while pushing it away from the original frames. Specifically, given the latent representation of the original frames, $\mathbf{X}_0$, and the target latent frames, $\mathbf{X}_{tg}$, we construct a contrastive objective [7, 9], as shown in Eq. (3).

$$\mathcal{L}_{con} = ||\hat{\mathbf{X}}_{0,adv} - \mathbf{X}_{tg}||^2 + \max\left(0, \tau_1 - ||\hat{\mathbf{X}}_{0,adv} - \mathbf{X}_0||^2\right) \quad (3)$$

The positive pair consists of the predicted frames $\hat{\mathbf{X}}_{0,adv}$ and the target frames $\mathbf{X}_{tg}$, which we aim to bring closer together, while the negative pair is composed of the predicted frames $\hat{\mathbf{X}}_{0,adv}$ and the original frames $\mathbf{X}_0$, which we aim to push apart.

With the hyperparameters α, β, γ, we define the final training objective as:

$$\mathcal{L} = ||I_{adv} - I_{src}||^2 + \alpha \cdot \mathcal{L}_{enc} + \beta \cdot \mathcal{L}_{con} + \gamma \cdot \mathcal{L}_{att} \quad (4)$$

The complete algorithm is presented in Algorithm 1.

It is noteworthy that our method supports both global and local attacks, allowing perturbations to be applied across the entire image or to specific objects within the image. This feature can be implemented by applying a mask to the noise during the attack process.

4. Experiments

4.1. Experimental Settings

Models In the field of video generation, there are few open-source video diffusion models available for experimentation. Stable Video Diffusion [4] (SVD) is a widely used and effective image-to-video diffusion model, which we adopt as the baseline model for our attack and experimental analysis. For other-conditioned image-to-video models, we utilize CogVideoX-5B-I2V [51] for text-image joint conditional video generation. The open-sourced CogVideoX, leverages DiT rather than the UNet architecture used in SVD, providing enhanced capabilities. We also select ControlNeXt-SVD-v2 [32] for pose-guided image-to-video generation, which is built upon SVD and allows for more control in character animation.

Data To our knowledge, there are currently no established benchmarks or datasets for image-based adversarial attacks in video generation. In previous research, PhotoGuard [35] reported attack results on over 60 images, while PRIME [26] used an internal dataset containing 35 video clips. For our experiments, we collect a dataset of 300 images and their corresponding generated frames. The selected images and frames depict people and animals with dynamic actions, which could potentially be misused. We exclude images of scenery without main subjects or images that do not yield noticeable temporal movement, as these typically do not produce meaningful motion, rely mainly on camera movements, and may even remain static.

Metrics We validate our experiments from two perspectives: image perturbation and generation effects. Our goal is to introduce minimal perturbations to the images that significantly degrade the generated results. We use peak signal-to-noise ratio (PSNR), structural similarity (SSIM) metrics, and Fréchet Inception Distance (FID) to evaluate the similarity between the attacked and original images. For evaluating video generation results, we use the metrics *Subject Consistency*, *Motion Smoothness*, *Aesthetic Quality*, and *Image Quality* from VBench [23]. Subject Consistency measures the average cosine similarity of Dino [6] features between each frame and the first frame, assessing temporal coherence. Motion Smoothness calculates the mean absolute error between interpolated and dropped frames. Aesthetic Quality is assessed using the LAION aesthetic predictor [1], and Image Quality evaluates the presence of low-level distortions in generated video frames with MUSIQ [24].

Setup We control the added noise strength by adjusting the number of training steps and the perturbation threshold at each step. For the loss function, we scale the hyperparameters α, β, γ to ensure that each loss component remains on a similar scale. During training, we use a black image as the target image I_{tg} and black frames as the target video V_{tg}. For the attention hooks, we focus on the temporal attention block of the last downsampling blocks in the UNet and the middle blocks in the transformer, separately.

4.2. Qualitative Analysis

Attack on Image-to-Video Generation We analyze the generation results using two image-to-video generation models: Stable Video Diffusion (SVD) and CogVideoX, as shown in Fig. 4. We observe that the noise added to the guarded image is nearly invisible, yet it causes significant variations in the generated results. In the example shown in Fig. 4a, we additionally illustrate the generated results for both a random noisy image and an image protected by PhotoGuard [35]. The random noise does not affect generation quality, preserving both temporal consistency and normal spatial appearance, likely because the VAE filters out such random noise. The PhotoGuard-protected image, however, introduces noise specifically on the person, which slightly affects appearance without compromising temporal consistency. In contrast, our protected image disrupts both spatial content and temporal consistency. Specifically, the woman's head and hair movement appear unnatural, chaotic frames are generated, and unusual textures emerge in the background across all frames. Fig. 4b shows results from CogVideoX under an image-only setting, where no prompt is applied during training and inference. CogVideoX, which is relatively more powerful than SVD, requires relatively stronger noise to affect generation. It can be observed that motion distortions still occur, particularly in areas where motion takes place, such as the blurred

(a) SVD generation results on original image, random noisy image, PhotoGuard image and our I2VGuard image.

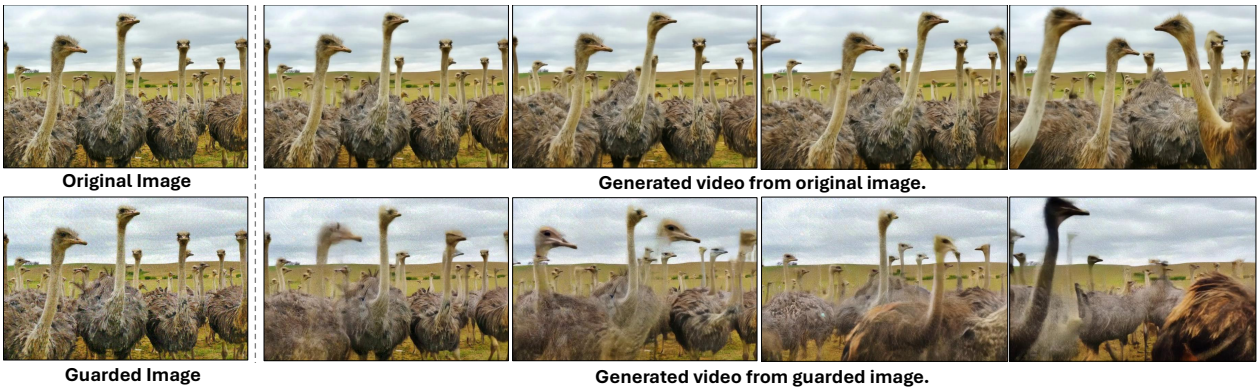
(b) CogVideoX generation results, where we do not set the prompt in the process of video generation.

Figure 4. **Qualitative results of adversarial attacks on I2V models SVD and CogVideoX**. We also include generation results of SVD with random noise and PhotoGuard [35] perturbations for comparison. All generation results are using the same seed.

head movement of the man, with the noise causing the surrounding regions to fail in generation, resulting in a blur. In summary, our method effectively disrupts both spatial content and temporal consistency in generated videos.

Attack on Conditional Image-to-Video Generation We present an attack on conditional image-to-video models in Fig. 5, specifically examining pose-conditioned ControlNeXt-SVD and text-guided video generation with CogVideoX. In the first row of Fig. 5a, the ControlNeXt model effectively controls the character's motion while maintaining consistent character appearance and background features. However, in the second row, where we guard the original image, the model fails to preserve character's appearance and fidelity in the background. Noticeable differences emerge: for instance, the fruits in the background exhibit blurring, and the character appears deformed. These spatial distortions illustrate the effectiveness of our attack method. Despite this, the pose control remains robust, likely due to the intensive training of the original SVD UNET, which results in strong overfitting to the specific style and hands shape. In the example of CogVideoX shown in Fig. 5b, we observe that adding special texture noise leads to the generation of temporal inconsistency and blurred content compared with the original ones. Specifically, in this case, texture noise is applied around the animals, causing distortions in the generated output, such as headless and transparent ostriches. In our experiments, CogVideoX exhibits strong prompt-following

(a) ControlNeXT-SVD-v2 generation results.



(b) CogVideoX generation results. The prompt is "*Many ostriches are running*".

Figure 5. **Qualitative results of adversarial attacks on conditional I2V models ControlNeXt and CogVideoX**. It can be observed that visual features are not controlled effectively, leading to unreasonable generation results. All generation results are using the same seed.

Figure 6. **Detailed visualization**. Comparison between generated results from the original and attacked images from Fig. 1 shows that some unreasonable textures are generated in the attacked frames.

ability, and by slightly increasing the noise strength, we are able to successfully execute the attack.

Details of Generation Results In Fig. 6, we provide a detailed examination of the distortions in generated frames compared to the frames generated from the original image. As shown in the right image, abnormal fragmented textures appear, which noticeably degrade the quality of the generated content. This result demonstrates that our method effectively disrupts the low-level visual quality in the generated frames.

4.3. Quantitative Analysis

Image Perturbation Analysis We present the modified image quality results in Tab. 2. Random noise was added to the original images to establish a baseline for comparison. The mean and variance of the added noise in this baseline are generally comparable to the modifications introduced in our attacked images. The results demonstrate that, despite similar perturbation strength, our method achieves significantly higher image quality, as indicated by higher score in both peak signal-to-noise ratio (PSNR) and structural similarity

Generated Video Source	Subject Consistency(%, ↑)	Motion Smoothness(%, ↑)	Aesthetic Quality (%, ↑)	Image Quality(↑)
Original Image	95.86±2.62	97.90±1.43	56.76±4.75	67.28±6.18
Random Noise Image	94.93±3.58	97.69±1.32	56.48±5.02	67.31±6.52
Our Guarded Image	**91.57**±3.95	**97.18**±1.21	**53.42**±4.93	**64.38**±8.23
w.o. Spatial Attack	94.72±3.67	97.62±1.35	56.28±5.31	66.68±6.90
w.o. Temporal Attack	93.74±3.87	97.56±1.37	54.80±5.35	65.98±7.70
w.o. Diffusion Attack	93.43±4.30	97.53±1.39	55.44±5.65	67.05±7.33

Table 1. Analysis of video generation effects of SVD from original images, images with random noise, and images guarded by our method. An ablation study on the different attack methods is also presented. Mean and variance of evaluations are reported. We exclude results with extremely high subject consistency and motion smoothness, as these indicate static frames, which are outside the scope of this evaluation.

Image Type	PSNR(↑)	SSIM(↑)	FID(↓)
Random Noise	26.22±0.04	0.774±0.003	14.31
Guarded Image	31.83±0.19	0.868±0.016	16.78

Table 2. Analysis of perturbations between original and modified images. Random noise is introduced to the original images to establish a baseline, and the mean and variance of the perturbations are reported.

index measure (SSIM). A qualitative comparison reveals that the noise added to the image possesses its own texture, which is less visible than the random noise added. In terms of image quality and diversity, our approach yields a higher Fréchet Inception Distance (FID). This is expected because the specific texture noise we add aligns with the original image texture, resulting in lower diversity compared to images with added random noise.

Video Generation Analysis Evaluation results for generated videos are shown in Tab. 1. First, we assess the performance of the SVD model on original images, finding that it demonstrates strong subject consistency, smooth motion transitions, high aesthetic quality, and well-preserved low-level image quality. This suggests that the SVD model is capable of generating visually coherent and temporally stable content. Next, we evaluate the baseline, which introduces random noise to the generation process. With the addition of random noise, the generated results display a minor decrease in all evaluation metrics. This modest drop indicates that the SVD model has robust generalization abilities, managing to preserve video quality despite slight perturbations. In contrast, our method, which uses a similar noise strength, significantly impacts all evaluated metrics. To be precise, temporal assessments reveal declines of 4.7% in subject consistency and 0.7% in motion smoothness, showing that our method effectively disrupts the model's temporal coherence. Specifically, the decline in subject consistency is reflected in irregular movement of the main object's location or the sporadic appearance and disappearance of the object. The decline in motion smoothness is reflected in abrupt, discontinuous frame-to-frame movements. Spatial evaluations also show that our method considerably diminishes aesthetic quality and low-level image quality by 6.2% and 4.5%, respectively. The decrease in aesthetic quality is evident in the appearance of unusual textures or object deformations that detract from the visual appeal. These reductions reflect a successful attack on the image-to-video model.

4.4. Ablation Study

We conduct an ablation study, as illustrated in Tab. 1, to examine the impact of omitting each of the three attack methods. The results indicate that without the spatial attack, the protection effect is relatively weak, as spatial-temporal performance remains high. In that case, fewer distinct textures are generated, and the added noise tends to be more uniform. This suggests that, without the spatial attack, added noise is partially filtered out by the encoder. Additionally, we observe that when both spatial and temporal attacks are applied, the protected results exhibit improved temporal consistency but reduced spatial quality compared to results with only spatial and diffusion attacks. This occurs because the temporal attack focuses exclusively on temporal features, while the diffusion attack affects both spatial and temporal aspects simultaneously. The exclusion of any one of the three attacks leads to reduced protection effectiveness, underscoring the importance of all three attacks in achieving optimal protection.

5. Conclusion

In this paper, we presented I2VGuard, a novel adversarial approach designed to apply imperceptible perturbations on images to degrade the quality of videos generated by diffusion-based image-to-video models, thereby protecting images from misuse. Concretely, we designed three attack modules, *i.e.* spatial attack, temporal attack and diffusion attack to disrupt visual fidelity, temporal consistency, and cross-model robustness, respectively. Extensive experiments on state-of-the-art generative models demonstrated the effectiveness of our approach.

References

[1] LAION AI. Laion-aesthetics. 5

[2] Ali Asghar, Amna Shifa, and Mamoona Naveed Asghar. Survey on video security: Examining threats, challenges, and future trends. *Computers, Materials & Continua*, 80(3), 2024. 2

[3] Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Yuanzhen Li, Tomer Michaeli, et al. Lumiere: A space-time diffusion model for video generation. *arXiv preprint arXiv:2401.12945*, 2024. 2

[4] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 2, 4, 5

[5] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024. 1, 2

[6] 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*, pages 9650–9660, 2021. 5

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 5

[8] Weixin Chen, Dawn Song, and Bo Li. Trojdiff: Trojan attacks on diffusion models with diverse targets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4035–4044, 2023. 2

[9] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *IEEE Conference on Computer Vision and Pattern Recognition, 2021*. Computer Vision Foundation / IEEE, 2021. 5

[10] June Suk Choi, Kyungmin Lee, Jongheon Jeong, Saining Xie, Jinwoo Shin, and Kimin Lee. Diffusionguard: A robust defense against malicious diffusion-based image editing. In *ICML 2024 Next Generation of AI Safety Workshop*. 2

[11] Sheng-Yen Chou, Pin-Yu Chen, and Tsung-Yi Ho. How to backdoor diffusion models? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4015–4024, 2023. 2

[12] Hai Ci, Yiren Song, Pei Yang, Jinheng Xie, and Mike Zheng Shou. Wmadapter: Adding watermark control to latent diffusion models. *arXiv preprint arXiv:2406.08337*, 2024. 2

[13] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19*, pages 424–432. Springer, 2016. 2

[14] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 1, 2

[15] Weitao Feng, Wenbo Zhou, Jiyan He, Jie Zhang, Tianyi Wei, Guanlin Li, Tianwei Zhang, Weiming Zhang, and Nenghai Yu. Aqualora: Toward white-box protection for customized stable diffusion models via watermark lora. *arXiv preprint arXiv:2405.11135*, 2024. 2

[16] Pierre Fernandez, Guillaume Couairon, Hervé Jégou, Matthijs Douze, and Teddy Furon. The stable signature: Rooting watermarks in latent diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22466–22477, 2023. 2

[17] 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. 2

[18] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1, 2

[19] 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. 2

[20] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in Neural Information Processing Systems*, 35:8633–8646, 2022. 2

[21] 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. 2

[22] 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*, pages 8153–8163, 2024. 2

[23] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818, 2024. 5

[24] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5148–5157, 2021. 5

[25] PKU-Yuan Lab and Tuzhan AI etc. Open-sora-plan, 2024. 2

[26] Guanlin Li, Shuai Yang, Jie Zhang, and Tianwei Zhang. Prime: Protect your videos from malicious editing. *arXiv preprint arXiv:2402.01239*, 2024. 2, 5

[27] Chumeng Liang and Xiaoyu Wu. Mist: Towards improved adversarial examples for diffusion models. *arXiv preprint arXiv:2305.12683*, 2023. 2

[28] 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. *arXiv preprint arXiv:2302.04578*, 2023. 2

[29] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pages 8162–8171. PMLR, 2021. 2

[30] Josephine Passananti, Stanley Wu, Shawn Shan, Haitao Zheng, and Ben Y Zhao. Disrupting style mimicry attacks on video imagery. *arXiv preprint arXiv:2405.06865*, 2024. 2

[31] 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. 1, 2

[32] 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. 2, 4, 5

[33] 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. 1, 2

[34] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015. 2

[35] Hadi Salman, Alaa Khaddaj, Guillaume Leclerc, Andrew Ilyas, and Aleksander Madry. Raising the cost of malicious ai-powered image editing. In *Proceedings of the 40th International Conference on Machine Learning*, pages 29894–29918, 2023. 2, 4, 5, 6

[36] 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. 2

[37] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. 2

[38] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015. 2

[39] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2

[40] Jingwei Sun, Xuchong Zhang, Changfeng Sun, Qicheng Bai, and Hongbin Sun. Latent feature and attention dual erasure attack against multi-view diffusion models for 3d assets protection. *arXiv preprint arXiv:2408.11408*, 2024. 2

[41] Kling AI Team. Kling ai, 2024. 2

[42] Luma AI Team. Luma ai, 2024.

[43] The Movie Gen team @ Meta. Movie gen: A cast of media foundation models. 2024. 2

[44] Vu Tuan Truong, Luan Ba Dang, and Long Bao Le. Attacks and defenses for generative diffusion models: A comprehensive survey. *arXiv preprint arXiv:2408.03400*, 2024. 2

[45] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 2

[46] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023. 2

[47] Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, et al. Lavie: High-quality video generation with cascaded latent diffusion models. *arXiv preprint arXiv:2309.15103*, 2023. 2

[48] Yuxin Wen, John Kirchenbauer, Jonas Geiping, and Tom Goldstein. Tree-ring watermarks: Fingerprints for diffusion images that are invisible and robust. *arXiv preprint arXiv:2305.20030*, 2023. 2

[49] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7623–7633, 2023. 2

[50] 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*, 2023. 2

[51] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 2, 4, 5

[52] Jianping Zhang, Zhuoer Xu, Shiwen Cui, Changhua Meng, Weibin Wu, and Michael R Lyu. On the robustness of latent diffusion models. *arXiv preprint arXiv:2306.08257*, 2023. 2, 4

[53] 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, 2024. 2

[54] Peifei Zhu, Tsubasa Takahashi, and Hirokatsu Kataoka. Watermark-embedded adversarial examples for copyright protection against diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24420–24430, 2024. 2