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

Supplementary Material

A. Robustness

In this section, we present the results of counterattacking our safeguarded image to evaluate the robustness of our proposed method, as illustrated in Fig. A. We begin by presenting the baseline results using both the original image and the safeguarded image. The findings indicate that our approach effectively disrupts the generated motion. In the third row, we examine the impact of JPEG compression, a common factor in practical applications. Our results demonstrate that even at a compression quality of 60%, the protective effect of our method remains substantial, continuing to disrupt motion generation. The fourth row presents a simple counterattack baseline, where Gaussian noise with a standard deviation of $\sigma = 10$ is added at the pixel level. The results confirm that despite the introduction of noisy blur, our approach continues to achieve effective motion disruption. Finally, in the last row, we incorporate NAFNet into our training pipeline, and training the adversarial image with denoising module and video diffusion model jointly. The results demonstrate that even in the presence of a denoising module, our method remains effective in taking it account and further impacting video generation models.

B. Details of Generation Results

In Fig. B, we provide a detailed examination of the distortions in generated frames compared to the frames generated from the original image. As shown in the generated frame from guarded 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.

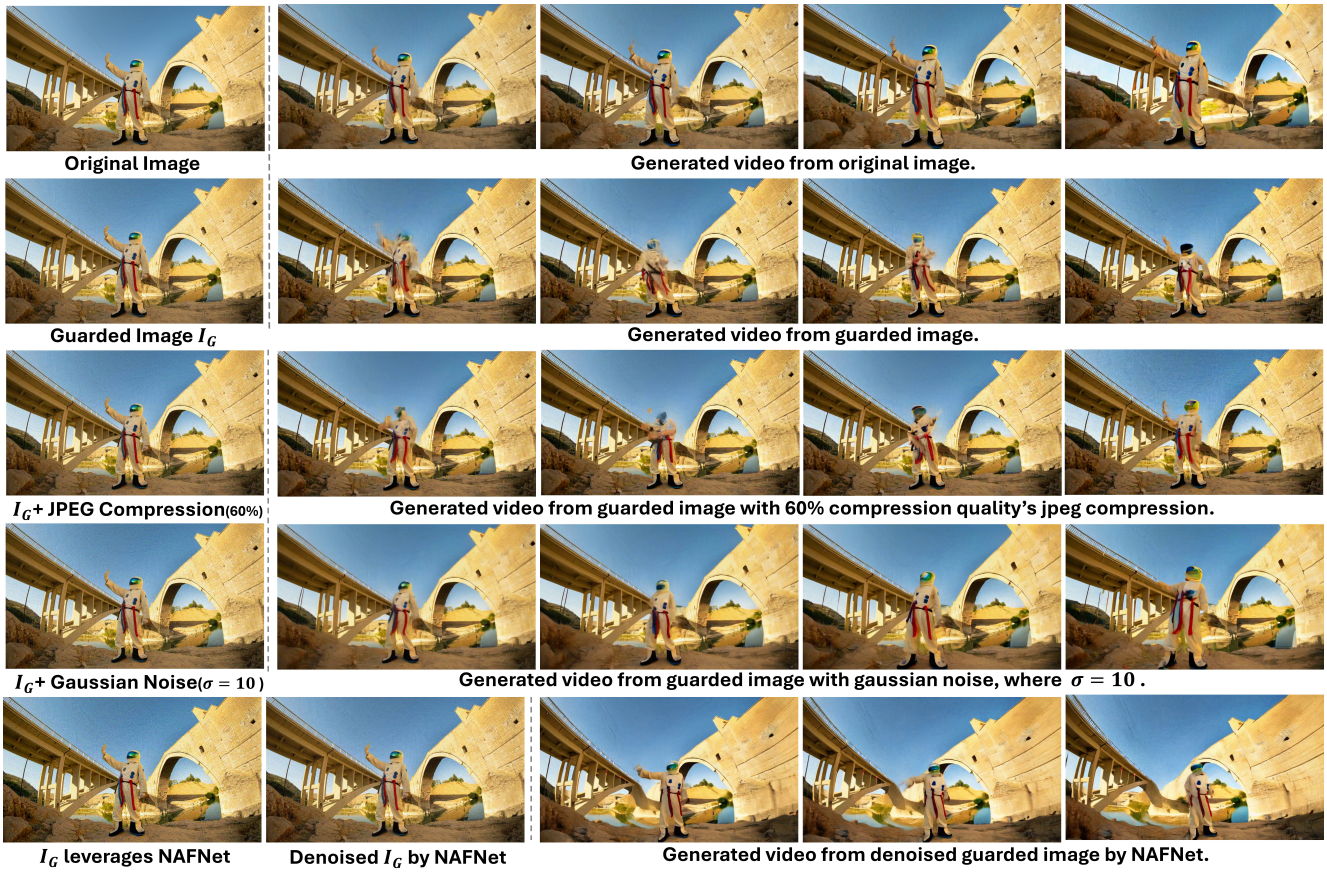

Figure A. Robustness of our I2VGuard by assessing its qualitative performance under various pre-processing techniques that could potentially affect the quality of the safeguarded image. Specifically, we consider three pre-processing methods: JPEG compression with a 60% quality ratio, the addition of Gaussian noise with $\sigma = 10$, and denoising module using NAFNet. All results are using the same seed.

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 A. Analysis of video generation effects of SVD from original images, images with random noise, images guarded by our method and ablation study on the different attack methods. 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.

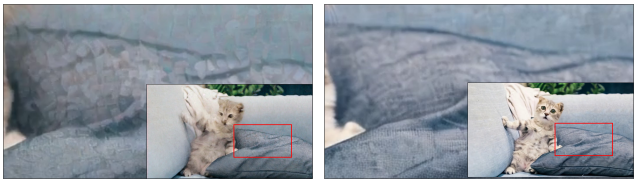

Figure B. **Detailed visualization**. Comparison between generated results from the guarded(left) and original images(right). It shows that some unreasonable textures are occurred in the generated frames from guarded image.

C. Hyperparameters Setup

In this section, we present the key hyperparameters used in the experiments, as summarized in Tab. B.

Hyperparameters	value
λ	0.01
α	10.0
β	1.0
γ	1.0
τ_1	2.0
τ_2	10.0
ϵ	0.03
Epochs	50
Width	SVD: 1024 CVX: 720
Height	SVD: 576 CVX: 480
Num Frames	SVD: 25 CVX: 49
FPS	7

Table B. Hyperparameters used in the experiments are presented. It is worth mentioning that these hyperparameters should be case-specific for optimal effects. Here, ϵ represents the maximum modification between I_{adv} and I_{src}. CogVideoX is abbreviated as CVX.

D. Ablation Study

We conduct an ablation study, as illustrated in Tab. A, to examine the impact of omitting each of the three attack methods.

We firstly 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.

As for the ablation, 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.

E. Qualitative Comparison

Compared to prior image-to-image generation approaches, challenge of our method lies in disrupting temporal consistency through perturbations applied to single images. To illustrate this, we present a qualitative comparison between spatial and temporal attacks in Fig. C. While spatial attacks degrade individual frames and reduce smoothness, they do not directly break temporal consistency. In contrast, temporal attacks specifically target frame-to-frame consistency, leading to unnatural motion shifts even when quantitative metrics remain relatively high.

Particularly in this case, spatial attacks result in abnormal textures within individual whole frames, like the textures on and around the dog, whereas temporal attacks cause noticeable motion inconsistencies, like the blur of the dog's head, highlighting the distinct impact of each approach.

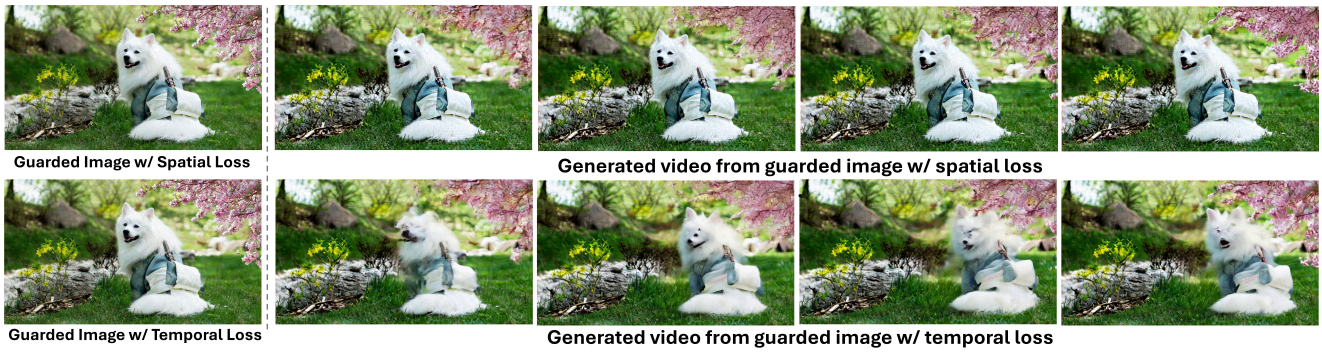

Figure C. Comparison of effects of spatial loss and temporal loss.

F. Failure Cases Analysis

In this section, we present a failure case study illustrated in Fig. D.

In the first example, the original video depicts smooth and coherent motion as a woman drops a cup, maintaining spatial quality throughout. However, in the guarded version, the motion remains largely static, failing to replicate or disrupt the original dynamics. Additionally, spatial quality deteriorates, with noticeable artifacts and abnormal textures, particularly around the woman's hand and the cup, highlighting challenges in generating coherent spatial features under the guarded method. In the second example, the original video shows smooth and consistent background motion, with pedestrians gradually walking into the scene. The guarded version partially disrupts this motion, causing pedestrians to appear sporadically with abrupt, inconsistent movements. Abnormal textures also emerge in the foreground, further degrading spatial quality and introducing artifacts.

In summary, failure cases primarily occur when the original video lacks significant motion, making it challenging for our method to introduce meaningful disruption. Alternatively, our method may occasionally damage the motion to the extent that the scene becomes static. However, the spatial quality attack is generally effective in degrading the visual fidelity of the generated content.

G. More Visualization Results

We present additional visualization results in Fig. E. These examples demonstrate that, with our guarded method ap-

Figure D. Failure case study. The generation results are with the same seed.

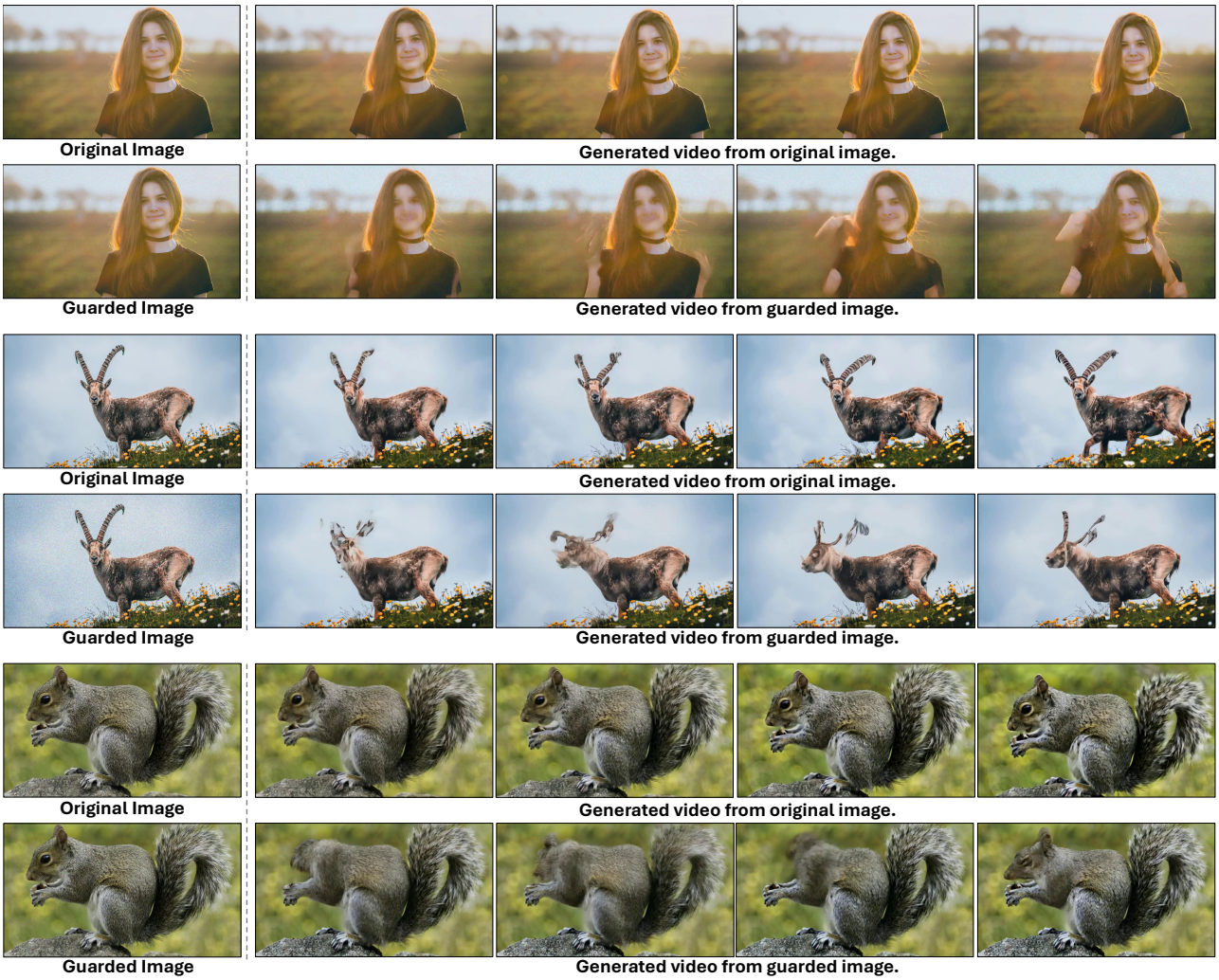

Figure E. More visualizations about our methods I2VGuard. The generation results are with the same seed.

plied to the input image, the video generation model fails to produce motion-consistent and spatially high-quality videos. For instance, in the first example, the video generated from the original image is consistent and visually appealing, showing a woman smiling naturally at the camera. In contrast, the guarded version fails to produce reasonable motion, with arbitrary arm movements that disrupt both temporal and spatial aesthetics. We also present two additional examples involving animals. In the second example, a deer with long antlers is depicted. While the original video successfully captures both the temporal consistency and spatial details of the deer, the guarded version struggles to reconstruct the head and antlers. Lastly, the third example focuses on a squirrel. In the original video, the motion remains smooth and consistent. In contrast, the guarded version introduces temporal disruptions, with abrupt frame-to-frame changes, and fails to maintain the spatial accuracy of the squirrel's head, resulting in visible artifacts and distorted features.

In summary, our method successfully introduces both spatial and temporal disruptions to the generated videos generated from guarded images. The visualization results demonstrate the robustness of our I2VGuard in compromising both spatial and temporal quality of generated videos.