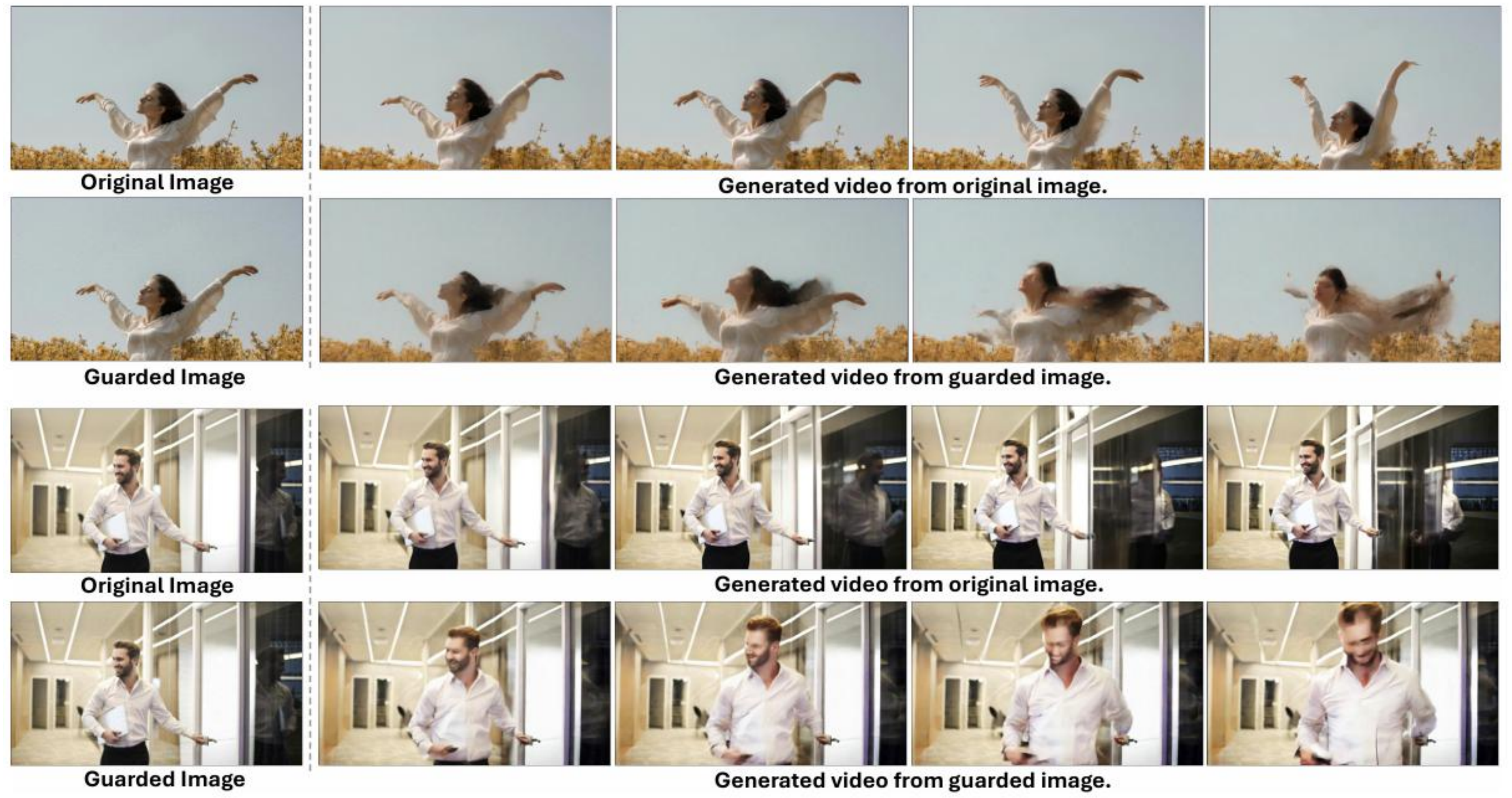
I2VGuard: Safeguarding Images against Misuse in Diffusion-based Image-to-Video Models

Dongnan Gui[1,*], Xun Guo[2], Wengang Zhou[1], Yan Lu[2]
[1]University of Science and Technology of China [2]Microsoft Research Asia

Introduction

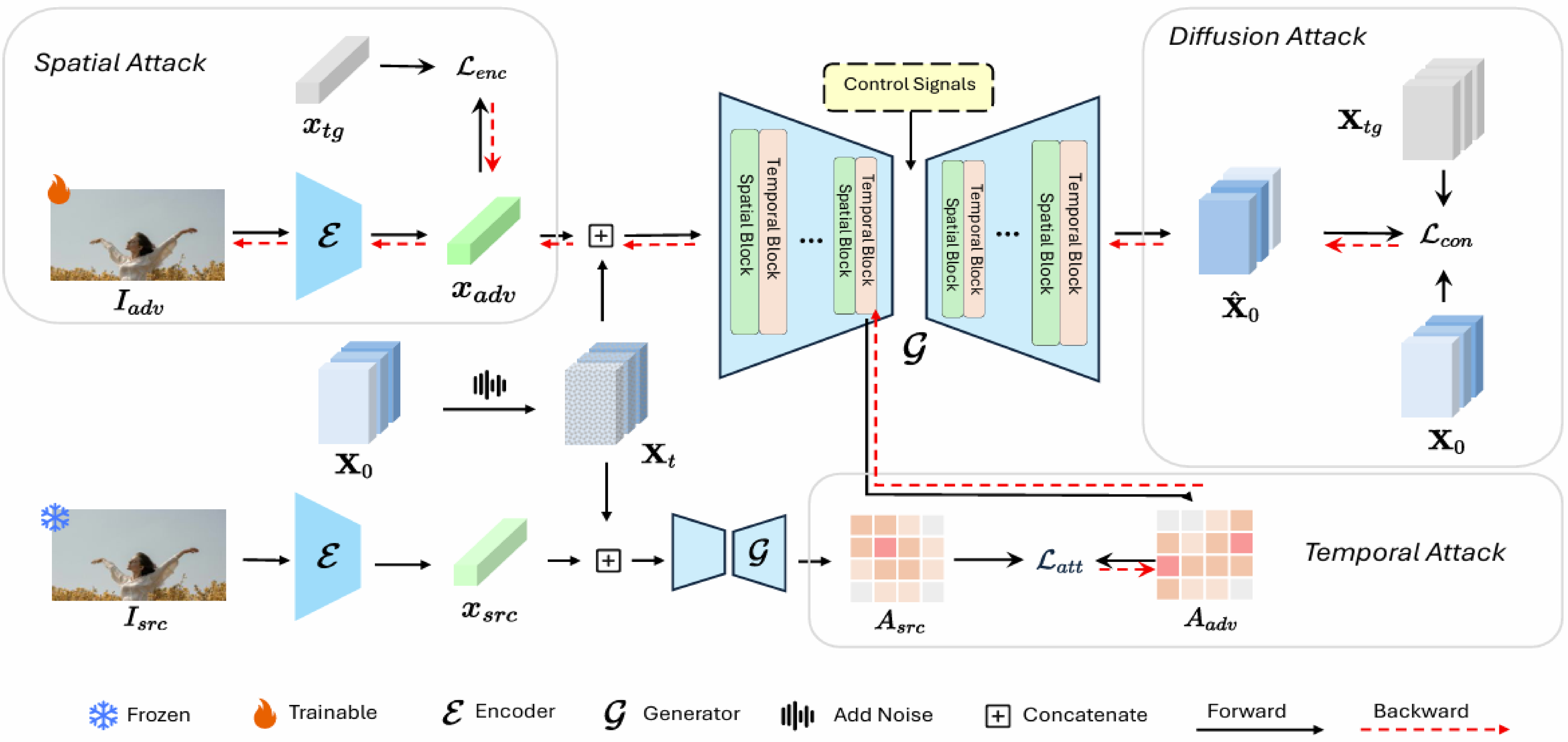
- **Motivation:** Image-to-video (I2V) diffusion models enable high-quality video generation but raise serious risks, including privacy breaches and copyright misuse. Existing defenses focus on protecting images from malicious image editing, not video generation.
- **Tasks:** We propose the first adversarial attack to protect images from misuse in diffusion-based I2V models, degrading the quality of generated videos with minimal, imperceptible perturbations.
- **Method Overview:** Our method targets three components of I2V models:
  - Spatial Attack: Alters image latents via the VAE encoder.
  - Temporal Attack: Disrupts motion by corrupting temporal attention.
  - Diffusion Attack: Uses contrastive loss to degrade denoising outputs.

This ensures robust protection across models and conditions (e.g., pose, text), and the protection results are shown below.



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 I2V generation.

Method



Method Overview: Training starts with a trainable copy I_{adv} of the original image I_{src}. We first perform inference to generate the original video V_0 and obtain latent frames $\boldsymbol{X}_0$. Noisy latent frames $\boldsymbol{X}_t$ and latent images x_{adv}, x_{src} are processed by the denoising module to reconstruct the original frames. The encoded x_{adv} is used for a spatial encoder attack, while the predicted frames $\hat{\boldsymbol{X}}_0$ are used in a diffusion-based contrastive loss. Within the denoising module, we extract the temporal attention map and modify A_{adv} to diverge from A_{src}, enabling the temporal attack.

Algorithm 1 Adversarial Attack on Image-to-Video Generation

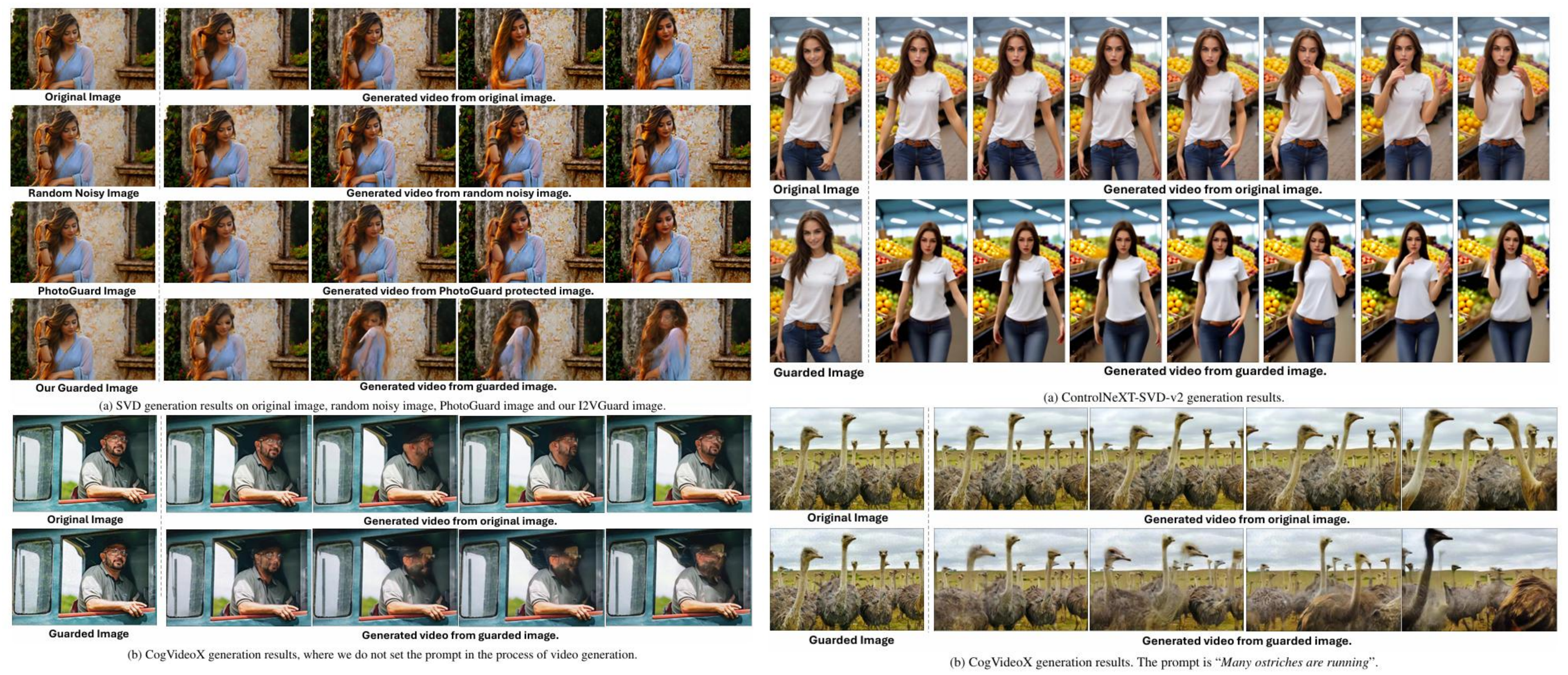
Input: Source Image I_{src} and Trainable Copy I_{adv}, Target Image I_{tg}, Target Video V_{tg}, Image-to-Video Pipeline $\mathcal{P}$, Generator $\mathcal{G}$, Encoder $\mathcal{E}$, Diffusion Scheduler $\mathcal{S}$ (Optional:Condition c)
Hyperparameters $\tau_1, \tau_2, \alpha, \beta, \gamma, \lambda$
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 $x_{adv} = \mathcal{E}(I_{adv})$ — Spatial Attack
  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 — Diffusion Attack
$$\mathcal{L}_{con} = ||\hat{\mathbf{X}}_{0,adv} - \mathbf{X}_{tg}||^2 + \max\left(0, \tau_1 - ||\hat{\mathbf{X}}_{0,adv} - \mathbf{X}_0||^2\right)$$
  Compute temporal attention loss — Temporal Attack
$$\mathcal{L}_{att} = \tau_2 - ||A_{adv} - A_{src}||^2$$
  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 $I_{adv} \leftarrow I_{adv} - \lambda \cdot \nabla_{I_{adv}} \mathcal{L}$
end for

Attention Visualizations of temporal self-attention map between generated frames from original image (left) and guarded image (right).

📌 Several Points:

Objective: Protect images from misuse in video generation by attacking latent diffusion pipelines.
Three-Pronged Attack: Spatial Attack: Breaks down encoder representation fidelity; Temporal Attack: Distorts attention over time to ruin motion; Diffusion Attack: Leverages contrastive loss to shift generation trajectory.
Optimization Strategy: Combined multi-term loss minimizes visibility while maximizing disruption.
Plug-and-Play: Model-agnostic design works across various diffusion frameworks.

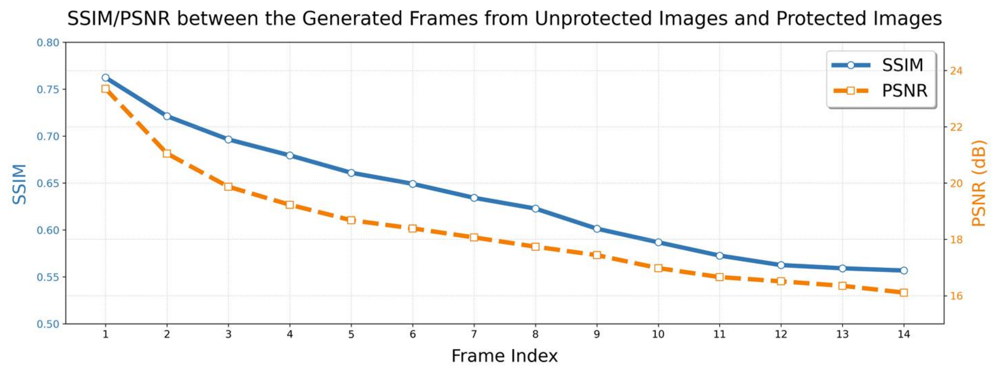
Experimental Results



Left: Qualitative results of adversarial attacks on I2V models SVD and CogVideoX. We also include generation results of SVD with random noise and PhotoGuard perturbations for comparison. **Right:** Qualitative results of adversarial attacks on conditional I2V models ControlNeXt and CogVideoX. All generation results are using the same seed.

Video Source	Model	Subject Consistency(%,↓)	Motion Smoothness(%,↓)	Aesthetic Quality(%,↓)	Image Quality(%,↓)
Original Image	SVD	95.86±2.62	97.90±1.43	56.76±4.75	67.28±6.18
Guarded Image		**91.57**±3.95	**97.18**±1.21	**53.42**±4.93	**64.38**±8.23
Original Image	CogVideoX	97.02±1.96	99.19±0.27	59.94±5.53	67.60±6.49
Guarded Image		**93.50**±3.58	**97.97**±0.32	**53.95**±5.62	**65.24**±9.85

Analysis of video generation results of SVD and CogVideoX from original images and images guarded by our method.



Experimental Analysis:
- In Qualitative Analysis: our method effectively disrupts both spatial content and temporal consistency in generated videos.
- In Quantitative Analysis, our method disrupts both temporal consistency, motion smoothness and spatial quality, leading to a propagated deviation from the original generation.

Conclusion

We introduce **I2VGuard**, a novel adversarial defense that applies imperceptible image perturbations to protect against misuse by diffusion-based I2V models.
Our method includes three targeted attack modules:
- Spatial Attack: disrupts visual fidelity
- Temporal Attack: breaks temporal consistency
- Diffusion Attack: ensures robustness across models

Tested on cutting-edge models like CogVideoX and SVD, I2VGuard proves highly effective in safeguarding image content.