Anti-I2V: Safeguarding your photos from malicious image-to-video generation

Supplementary Material

Overview: We first introduce preliminaries on the CIELAB ($L^*a^*b^*$) color space in Sec. 8 and the Discrete Cosine Transform (DCT) in Sec. 9. Sec. 10 then outlines the experimental settings and parameters for all methods. Sec. 11 details the hyperparameter settings for the purification and transformation techniques described in Sec. 6.1. In addition, Secs. 12 to 17 provide ablation studies on the design of the perturbation optimization space, further transferability experiments, a component analysis of $\mathcal{L}_{Anti-I2V}$, and the method's effectiveness across diverse prompts. We additionally provide more details on the benchmark construction in Sec. 19 and qualitative examples in Sec. 20.

Note: Ablation experiments in the supplementary material are conducted using CogVideoX-5B [25] on a subset of 200 videos from CelebV-Text [65], where the first frame serves as the image condition and the provided caption serves as the prompt. For each image–prompt pair, we generate five samples, resulting in a total of **1,000 videos** for evaluation. Unless stated otherwise, all experiments follow this setup.

8. CIELAB ($L^*a^*b^*$) color space

The conversion from linear RGB (normalized to [0, 1]) to Lab is a two-stage process. First, a linear transformation to CIE XYZ space is performed using a predefined matrix $\mathbf{M}$: $\begin{bmatrix} X & Y & Z \end{bmatrix}^T = \mathbf{M} \begin{bmatrix} R & G & B \end{bmatrix}^T$. Subsequently, a non-linear transformation yields the Lab values.

$$\begin{aligned} &X_n = 0.95047, \quad Y_n = 1.0, \quad Z_n = 1.08883, \\ &f(t) = \begin{cases} t^{1/3} & \text{if } t > (6/29)^3 \\ \frac{1}{3}(29/6)^2 t + 4/29 & \text{otherwise} \end{cases}, \\ &L^* = 116 f(Y/Y_n) - 16, \\ &a^* = 500[f(X/X_n) - f(Y/Y_n)], \\ &b^* = 200[f(Y/Y_n) - f(Z/Z_n)]. \end{aligned} \tag{13}$$

9. Discrete Cosine Transform (DCT)

For an RGB image $x_0 \in \mathbb{R}^{3\times h\times w}$, its frequency representation X_0 is given by:

$$X_0(k,u,v) = c_u c_v \sum_{i=0}^{h-1} \sum_{j=0}^{w-1} x_0(k,i,j)\phi(i,u)\phi(j,v), \tag{14}$$

where k is the channel index, u and v are 2D coordinates in the frequency space, $c_u = \sqrt{1/h}$ if $u = 0$ or $c_u = \sqrt{2/h}$ otherwise, $\phi(i,u) = \cos\left(\frac{\pi(0.5+i)}{h}u\right)$, and c_v and $\phi(j,v)$ have similar formulas as c_u and $\phi(i,u)$, respectively. The inverse function, i.e., IDCT, to map from frequency domain to RGB domain is defined as:

$$x_0(k,i,j) = \sum_{u=0}^{h-1} \sum_{v=0}^{w-1} c_u c_v X_0(k,u,v)\phi(i,u)\phi(j,v). \tag{15}$$

10. Implementation Details

For all experiments, we fix the number of update iterations to $N = 200$ and use a perturbation budget of $\Delta_{RGB} = 16/255$. For our **Anti-I2V** method, we additionally set $\Delta_{Lab} = 16/255$. To improve efficiency and reduce memory usage, we use only the first four frames of each video as input to the VDMs. Following [58], we adopt the AdamW optimizer with a learning rate of $1\mathrm{e}{-2}$. Baseline methods are optimized using PGD [34] with a step size of $1/255$. All experiments are run on a single NVIDIA A100 GPU 40GB.

11. Robustness Settings

For JPEG Compression, we compress each image at the compression rate of 40%. For Gaussian blur, we set the kernel size to 7 and $\sigma = 1.5$. For Gaussian noise, we set the noise scale to 0.05. For DIffPure [40], we set the number of iterations to 100, with $\epsilon_{adv} = 0.07$. For GrIDpure [68], we also set the number of iterations to 100, with $\gamma = 0.1$. All experiments use the same objective function, $\mathcal{L}_{Anti-I2V}$.

12. Analysis of Perturbation Update Space

To evaluate the Dual-Space Perturbation design (Sec. 4.3), we compare perturbations applied in RGB space, $L^*a^*b^*$ space, frequency space, and their combinations. For a fair comparison, all adversarial attacks use only the vanilla denoising loss as the optimization objective.

Table 4. **Quantitative results** of perturbation optimization spaces.

Method	ISM↓	Q-A (F)↓	Q-A (V)↓	DINO-SIM↓
RGB	**0.582**	0.624	0.692	0.788
$L^*a^*b^*$	0.672	0.707	0.765	0.833
Frequency	0.633	0.576	0.641	0.802
RGB + $L^*a^*b^*$	0.613	0.525	0.618	0.784
RGB + Frequency	0.587	0.567	0.645	0.796
$L^*a^*b^*$ + Frequency **(DSP)**	**0.582**	**0.521**	**0.610**	**0.781**
RGB + DSP	0.654	0.566	0.639	0.805

As shown in Tab. 4, under identical objectives and settings, our **Dual-Space Perturbation (DSP)** yields stronger cloaking effects than traditional RGB-space perturbations. While

Table 5. **Analysis of perturbation budget for $L^*a^*b^*$ space.**

Budget	ISM↓	C-FIQA ↓	Q-A (F)↓	Q-A (V)↓	DINO↓
$8/255$	0.469	0.456	0.485	0.567	0.775
$16/255$	**0.462**	**0.448**	**0.481**	**0.562**	**0.760**
$32/255$	0.473	0.468	0.491	0.588	0.775

perturbing only in the $L^*a^*b^*$ or frequency domains underperforms RGB, combining these domains substantially improves protection, as reflected by the drops in Q-Align (V) and Q-Align (F). Adding RGB to DSP, however, degrades performance because the fixed perturbation budget must be split across more spaces, reducing the impact of each. Although RGB performs comparably on ISM, it falls short in other overall quality and aesthetics metrics. DSP is preferred for its stronger robustness to purification, as elaborated in Sec. 6.1

13. Perturbation budget for $L^*a^*b^*$ color space

Tab. 5 shows the performance of our method under different perturbation budgets in the $L^*a^*b^*$ color space. With all other parameters and loss components fixed and using the same objective function, $\mathcal{L}_{Anti-I2V}$, we evaluate three levels of Δ_{Lab}: $8/255$, $16/255$, and $32/255$. Increasing the perturbation budget does not always enhance protection. The best balance between identity concealment and video quality is achieved at $\Delta_{Lab} = 16/255$, which achieves the lowest DINO-SIM while maintaining the lowest ISM and Q-Align scores. This suggests that Δ_{Lab} can be flexibly selected based on the desired protection and objective.

14. Transferability

Following the protocol in Sec. 6.2, we additionally evaluate transferability on the recent Wan2.2-TI2V-5B [55], as shown in Tab. 6. Consistent with the other transfer settings, Anti-I2V clearly surpasses all baselines on Wan2.2, further demonstrating strong generalization to up-to-date models.

Table 6. **Quantitative comparison of transferbility** from CogVideoX-5B to Wan2.2-TI2V-5B.

	CogVideoX-5B - Wan2.2-TI2V-5B				
Method	ISM ↓	C-FIQA ↓	Q-A(F) ↓	Q-A(V) ↓	DINO ↓
Clean	0.672	0.517	0.841	0.899	0.815
SDS+	0.538	0.478	0.512	**0.622**	**0.741**
SDS-	0.608	0.504	0.573	0.667	0.766
AdvDM	0.544	0.456	0.528	0.624	<u>0.743</u>
MIST	0.635	0.476	0.574	0.678	0.790
VGMShield	0.628	0.499	0.530	0.624	0.778
Anti-I2V	**0.439**	**0.450**	**0.502**	<u>0.623</u>	<u>0.743</u>

Table 7. **Ablation Study of Loss Components**: Comparison of different loss components on performance metrics. (U) denotes untargeted attack, while (T) denotes targeted attack.

Loss Type	ISM ↓	Q-A (F) ↓	Q-A (V) ↓	DINO ↓
[A1]: Denoising Loss	0.582	0.521	0.610	0.781
[A2]: **[A1]** + IRC	0.535	0.518	0.607	0.776
[A3]: **[A1]** + IRA-VAE (U)	0.514	0.507	0.583	0.774
[A4]: **[A1]** + IRA-VAE (T)	0.507	0.503	0.580	0.771
[A5]: **[A1]** + IRA-Denoiser (U)	0.507	0.486	0.572	0.775
[A6]: **[A1]** + IRA-Denoiser (T)	0.493	0.484	0.575	0.772
[A7]: **[A4]** + **[A6]**	0.484	0.483	0.567	0.768
[A8]: **[A2]** + **[A4]** + **[A6]**	0.476	0.481	0.567	0.764
[A9]: **[A8]** + Auxiliary Loss	**0.462**	**0.481**	**0.562**	**0.760**

15. Loss Components

Tab. 7 highlights the effectiveness of applying IRC and IRA compared to the vanilla denoising loss. Specifically, both IRC and IRA, when individually combined with the vanilla denoising loss, lead to a decline across all metrics, particularly in identity-related features. This indicates that these components disrupt the information flow within the denoising modules, causing the reference image features to diverge. Furthermore, the results suggest that high-level semantic features (e.g., human identity) are significantly affected. Combining both losses further reduces ISM and DINO-SIM, implying that IRC and IRA complement each other.

16. IRC Layer Selection

Tab. 8 highlights the effectiveness of applying IRC under different layer configurations. We compare the optimal layer selection identified in Sec. 4.4 with a simplified variant that applies the IRC loss only on the last three layers. We further investigate more configurations by applying IRC to the last one, two, and four layers. The results show virtually identical performance, indicating that the simplified setting of IRC loss provides comparable protection without any need of complicated layer selection.

Table 8. **Ablation on IRC Layer Selection**. We compare applying IRC to different subsets of layers. **Full** applies IRC to all layers after the 27$^{\text{th}}$ layer. **Last-k** applies IRC only to the final k layers. **Bold** indicates the best performance, and <u>underline</u> indicates the second best.

Setting	ISM ↓	Q-A (F) ↓	Q-A (V) ↓	DINO ↓
Full (27+)	**0.458**	**0.479**	**0.560**	0.762
Last-3	0.462	<u>0.481</u>	<u>0.562</u>	**0.760**
Last-1	<u>0.460</u>	0.486	0.565	0.762
Last-2	<u>0.460</u>	0.487	0.568	0.767
Last-4	0.470	0.489	0.570	0.764

17. Evaluation on Different Prompts

We evaluate whether our method can generate effective attacks independent of the provided prompts. For each image ID in our evaluation subset, we use [10] to generate three distinct text prompts. Captions are obtained using the query: "Return me three different text prompts for video generation based on this image. The prompts should focus on the human subject, their appearance, and their actions." For each image–prompt pair, we generate five samples, resulting in a total of 3000 videos for evaluation. We use CogVideoX [25] and DynamiCrafter [61] as representative models for DiT-based and UNet-based architectures, respectively. Tab. 9 demonstrates that despite the difference in provided prompts, our method significantly successfully degrades both identity features and video quality.Our method consistently achieves strong protection across different prompts, as evidenced by lower ISM, CLIB-FIQA, Q-Align, and DINO-SIM scores. Moreover, our method maintains its effects on DynamiCrafter [61], similar to experiments in Sec. 5. This suggests that our approach is robust to prompt variations, effectively disrupting video generation regardless of the textual descriptions used.

Table 9. **Quantitative comparisons of protections with different set of prompts.** ↓ indicates that a lower value of the metric signifies poorer video quality and thus better protection.

Method	Metric ↓	Clean	Anti-I2V
CogVideoX [25]	ISM	0.646	**0.407**
	C-FIQA	0.519	**0.462**
	Q-A (Frame)	0.771	**0.474**
	Q-A (Video)	0.825	**0.553**
	DINO	0.869	**0.776**
DynamiCrafter [61]	ISM	0.558	**0.208**
	C-FIQA	0.521	**0.367**
	Q-A (Frame)	0.883	**0.084**
	Q-A (Video)	0.912	**0.104**
	DINO	0.875	**0.234**

18. Perturbation Visibility

We conduct experiments to evaluate the imperceptibility of each method, using SSIM and PSNR as our primary metrics. Notably, perturbations in the Lab color space prioritize perceptually meaningful changes, which do not fully align with how PSNR and SSIM measure image similarity. These metrics emphasize pixel-wise differences and structural consistency in the RGB space, making them less reflective of human visual perception. As a result, while perturbations in the Lab space subtly alter color information in a way that is less noticeable to humans, they can still lead to lower PSNR and SSIM scores due to significant pixel-level differences. Nevertheless, as shown in Tab. 10, our method remains competitive across all metrics despite its lower SSIM and PSNR values.

Method	SSIM↑	LPIPS↓	PSNR↑
SDS (+)	0.84	0.206	32.5
SDS (-)	**0.86**	0.192	**33.7**
AdvDM	0.84	0.205	32.6
MIST	0.82	0.271	31.4
VGMShield	0.84	**0.191**	33.2
Ours	0.80	0.200	32.2

Table 10. **Ablation Study of Perturbation Visibility**: Similarity metrics between perturbed images and their original images of different methods.

19. Benchmark Construction

To obtain reference videos for the inputs, we use [10] to generate captions describing the adversarial examples. Specifically, we query the model with the prompt: "Return an extremely detailed prompt ONLY describing the person in this image (including appearance, emotion, and action)." For CelebV-Text [65], we first crawl videos with unique identities, using Qwen [47] to obtain person-centric descriptions, from which we synthesize the first-frame images using FLUX [29]. We then pair each image with the corresponding video prompt described above, crop the image to satisfy the model input requirements, and generate five samples for each image–prompt pair.

20. More qualitative results

We provide more qualitative examples to compare our method Anti-I2V against other baselines with different diffusion models. Please refer to Figs. 4 to 7.

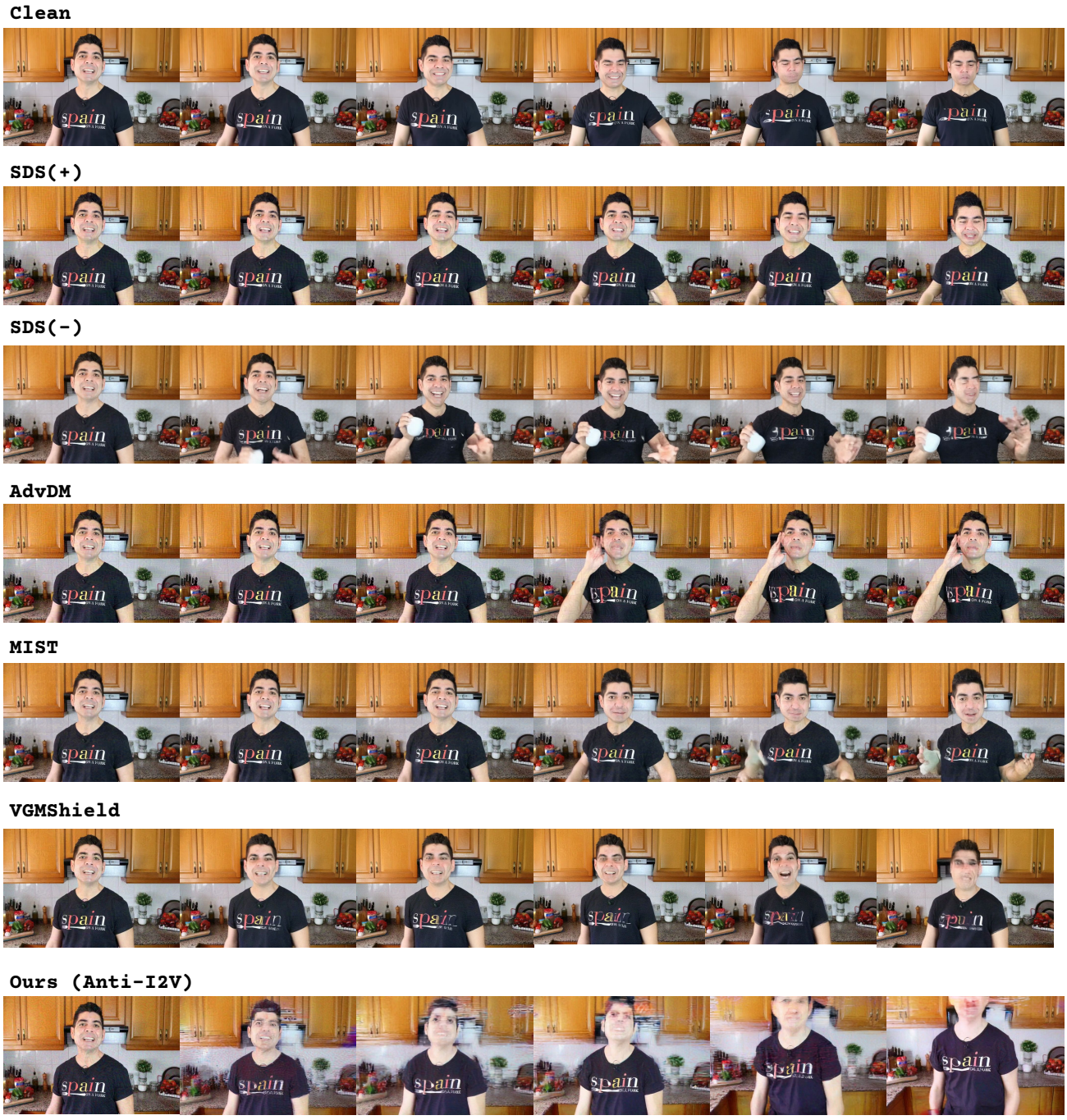

Figure 4. Qualitative comparison of adversarial attack methods against against CogVideoX [25]. The first column shows the reference frame. The remaining columns present the generated outputs from models.

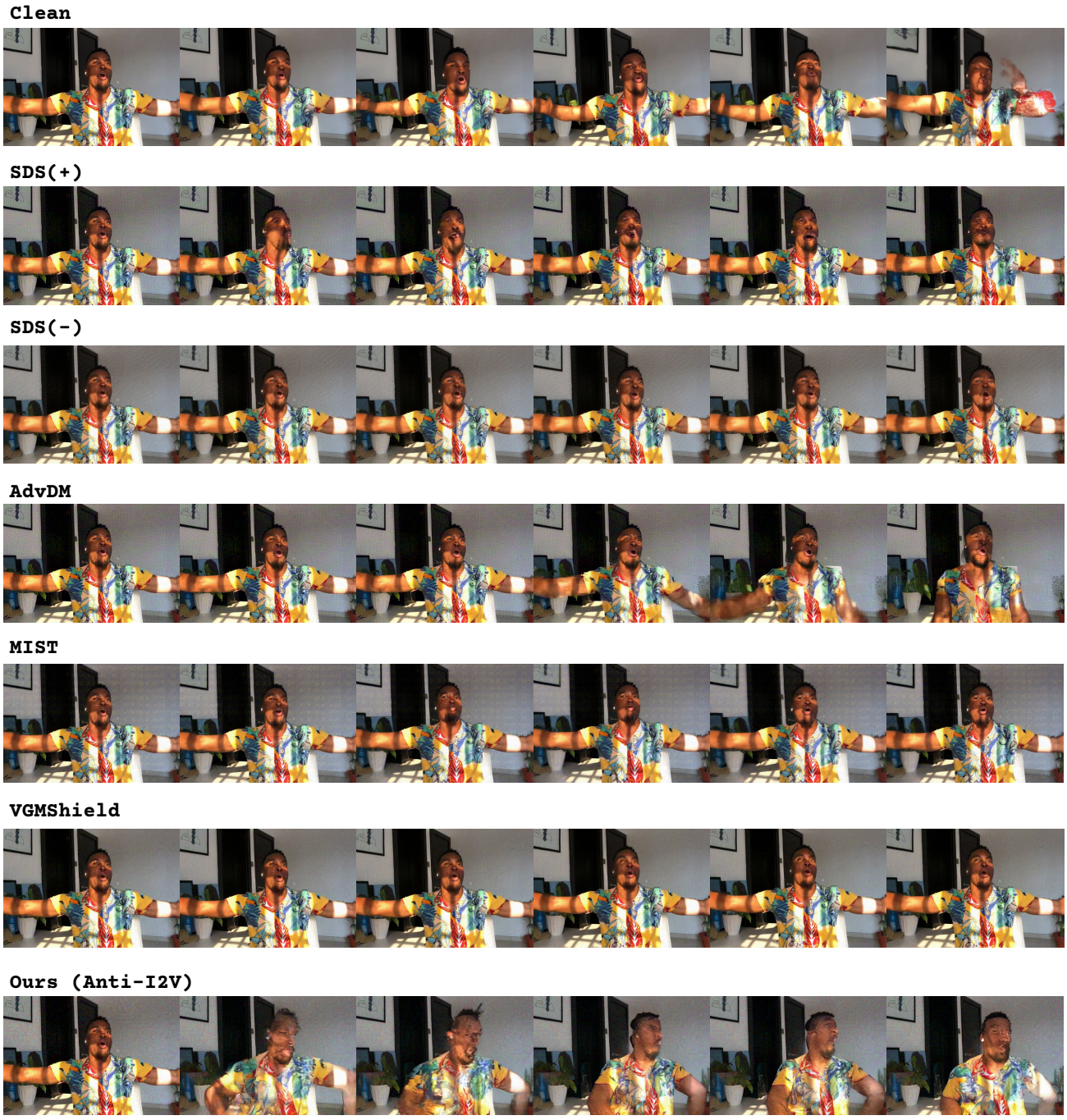

Figure 5. Qualitative comparison of adversarial attack methods against against CogVideoX [25]. The first column shows the reference frame. The remaining columns present the generated outputs from models.

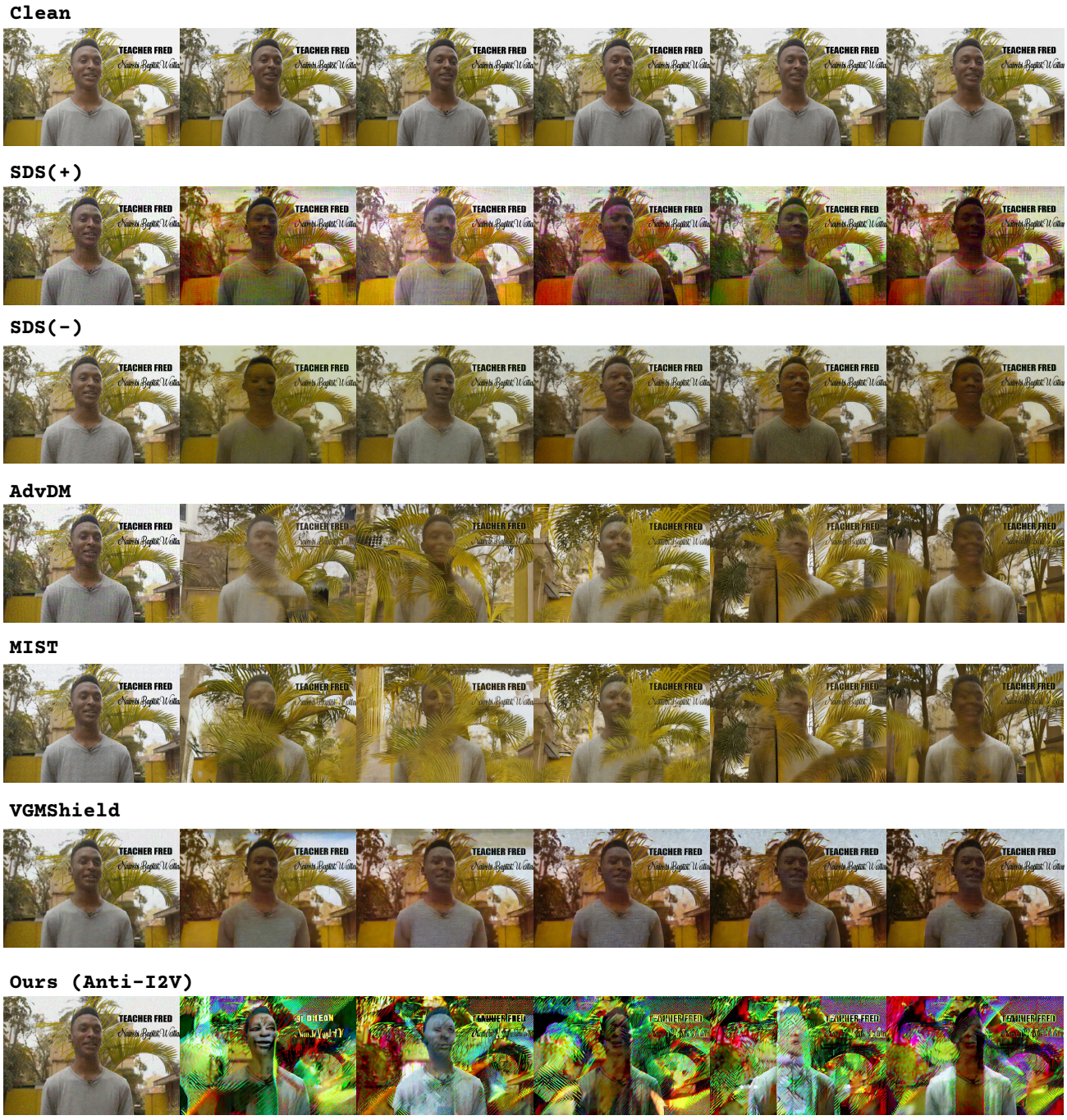

Figure 6. Qualitative comparison of adversarial attack methods against against DynamiCrafter [61]. The first column shows the reference frame. The remaining columns present the generated outputs from models.

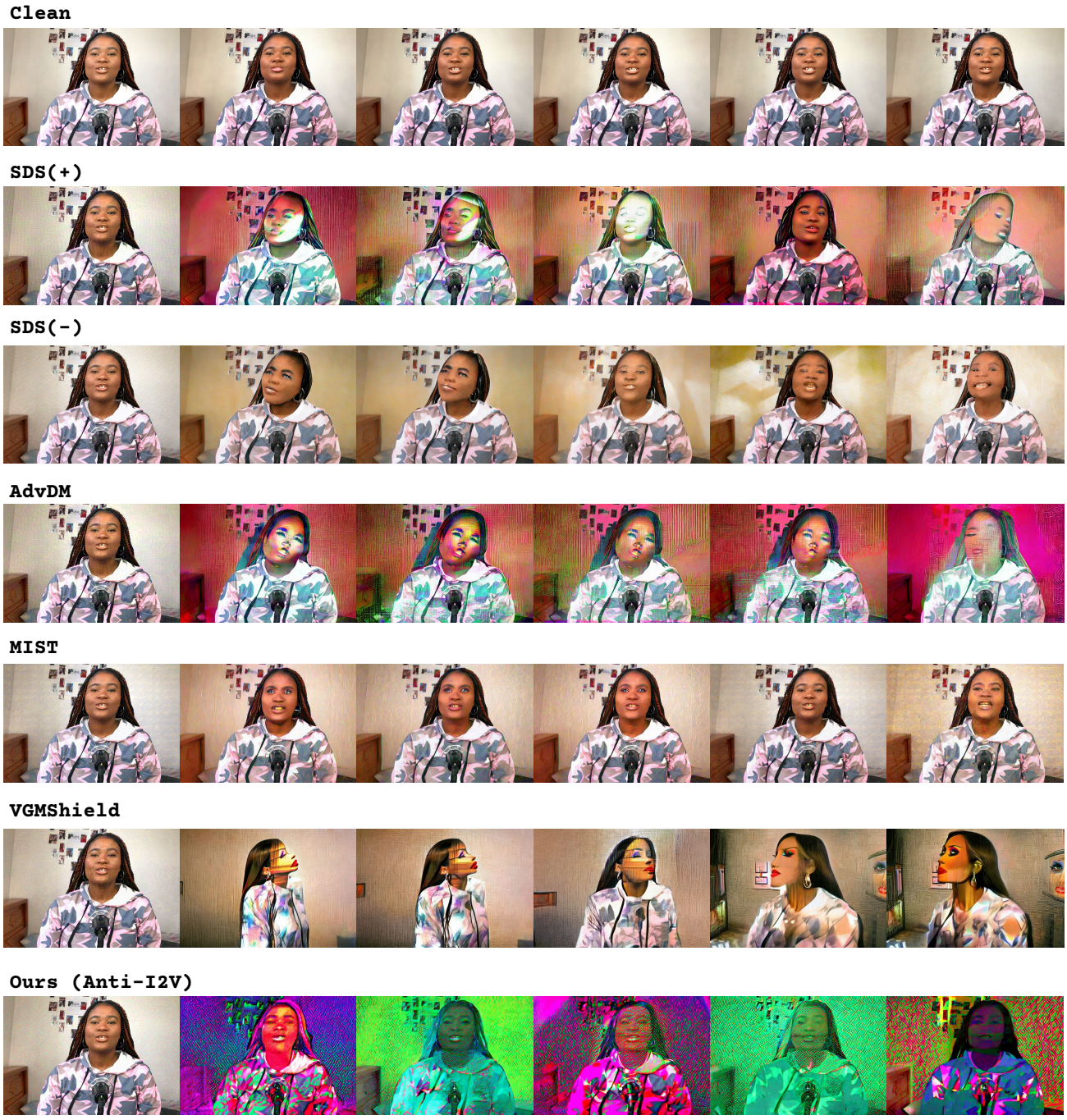

Figure 7. Qualitative comparison of adversarial attack methods against against DynamiCrafter [61]. The first column shows the reference frame. The remaining columns present the generated outputs from models.