Anti-Prompt: Image Protection against Text-Guided Image-to-Video Generation

Yeonghwan Song[1] Chanhui Lee[2] Jinsoo Park[2] Jeany Son[2⋆]

[1] GIST [2] POSTECH

https://yeonghwansong.github.io/Anti-prompt/

Fig. 1: Text-guided I2V generation and protection. CogVideoX [51] generates videos from the original image (top) and our protected image (bottom).

Abstract. Recent advances in Image-to-Video (I2V) generation allow a single image to be animated into a convincing video under text guidance, raising serious copyright and privacy risks. We propose Anti-Prompt, an image protection approach that injects imperceptible perturbations into an image, inducing visible inconsistencies and structural failures in text-guided I2V generation. Our method is motivated by a simple empirical observation: when text guidance is removed from modern I2V models, generation quality degrades markedly, not only in motion realism but also in subject preservation, structural coherence, and temporal consistency. Building on this insight, Anti-Prompt exploits the model's reliance on textual guidance by attenuating text-conditioned interactions during denoising while strengthening visual-only pathways. To further systematically evaluate protection effectiveness, we introduce a Video-LLM–assisted evaluation protocol that provides interpretable, frame-grounded analyses of generation artifacts and inconsistencies. Experiments on two representative I2V architectures demonstrate that our method achieves strong protection performance while improving efficiency and cross-model transferability.

Keywords: Safety and Privacy · Image Protection · I2V Generation

⋆ Corresponding author.

1 Introduction

Diffusion models have rapidly advanced image synthesis [9,12,18,23,31–34,36,40] and video generation [4,5,15–17,19,39,41,44,45,51], making it increasingly easy to animate a single photograph into a realistic short video. Modern Image-to-Video (I2V) models [4,7,46,50,55] can generate temporally coherent motion and high visual fidelity from one image, guided by a text prompt. These capabilities enable compelling content creation, but they also introduce a practical risk for publicly shared images [6,30,42,48]. Once an image is posted online, it becomes a reusable source that can be repeatedly animated under a wide range of prompts, without the owner's intent. Compared to static misuse, video generation can be more persuasive because it adds actions, expressions, and temporal narratives, thereby amplifying privacy and copyright risks [42]. This motivates proactive image protection [10,13,24,37,38] tailored to diffusion-based I2V dynamics.

In this work, we study I2V image protection. Given an image intended for sharing, the owner adds an imperceptible perturbation so that attempts to animate the image are less likely to produce a coherent and image-faithful video. Here, image-faithful refers to preserving the input image's subject identity and overall structure throughout the generated video. A successful protection, therefore, causes generated videos to exhibit noticeable inconsistencies or artifacts, making them unsuitable for convincing misuse. Figure 1 illustrates representative examples of such protected generations.

Our approach is motivated by a simple empirical observation. In several modern I2V models [15,45,51], removing the text prompt during generation causes severe degradation that extends beyond prompt-following. As illustrated in Figure 2, prompt-free generation often exhibits subject deformation, structural collapse, and temporal incoherence. These failures are broad in scope: they affect not only motion quality but also subject preservation and structural consistency. This behavior suggests that text guidance plays a broader role in stabilizing generation than typically assumed.

Based on this insight, we propose Anti-Prompt, an image protection method that weakens the influence of text during I2V denoising. Rather than targeting specific prompts, our method targets the architectural pathways through which text affects generation. Concretely, our objective suppresses text-conditioned interactions while reinforcing visual-only pathways, shifting the model's updates away from text guidance. We instantiate this idea for two common I2V architectures: full-attention models, where video tokens jointly attend to video, image, and text tokens, and cross-attention models, where text influences the latent through a dedicated residual branch. Our optimization operates on intermediate attention statistics rather than generated videos, eliminating the need for expensive reference video generation and significantly improving efficiency compared with prior approaches [13].

Finally, we introduce an evaluation protocol tailored to I2V protection. Standard video benchmarks (*e.g.*, VBench [20,21,56]) measure overall video quality but provide limited insight into how generation degrades on protected images. In protection settings, success is inherently asymmetric: even a single salient

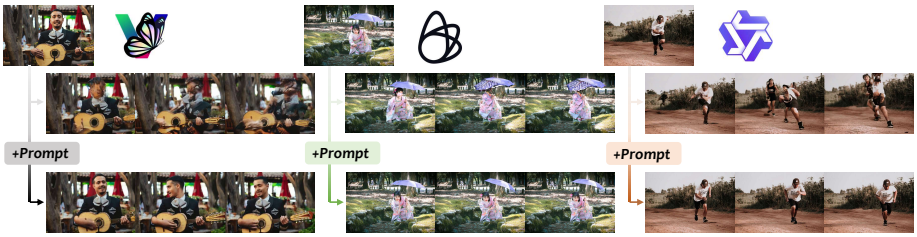

Fig. 2: Empirical observation: prompt removal degrades I2V generation broadly. Top: prompt omitted (empty text). Bottom: standard text-guided generation with all other settings fixed. Removing text guidance causes failures that extend beyond motion quality to subject preservation, structural coherence, and temporal consistency. This pattern is observed across CogVideoX [51], LTX-Video [15], and Wan [45].

error can render a generated video unsuitable for convincing misuse. To better capture such behaviors, we complement standard benchmarks with a Video-LLM–assisted evaluation protocol [2, 26, 29, 52, 53]. The protocol evaluates four interpretable dimensions, *Subject Preservation, Structural Consistency, Dynamic Consistency, and Artifact Suppression*, using structured outputs that cite frame-grounded observations before assigning discrete scores.

Our contributions are summarized as follows:

- We propose Anti-Prompt, an I2V image protection method that attenuates text-conditioned interactions during denoising while amplifying visual-only pathways, instantiated for both full-attention and cross-attention I2V architectures.
- We introduce a Video-LLM-assisted evaluation protocol that provides inspectable, frame-grounded diagnoses of protection-induced failures across four visual dimensions.
- Our method achieves effective disruption of plausible animation with improved efficiency and high imperceptibility across two representative I2V architectures under our evaluation suite.

2 Related Work

Image-to-Video Generative Models. Modern image-to-video (I2V) synthesis is largely built by extending text-to-image diffusion backbones with temporal modules to synthesize temporally coherent frames from a single image. Early works such as Tune-A-Video [49] and AnimateDiff [14] demonstrated that injecting lightweight temporal components enables compelling image animation. Stable Video Diffusion [4] further scaled latent video diffusion and has served as a widely used open-source baseline. Recent models improve fidelity and controllability through stronger spatio-temporal architectures, including Lumiere [3] and large-scale diffusion transformers such as CogVideoX [51]. In parallel, LTX-Video [15] proposes efficiency-oriented designs and Wan [45] continues to reduce

the barrier to high-quality I2V generation as a large-scale open source model. Community-driven efforts such as VideoCrafter2 [8] and Open-Sora [58] further broaden access to I2V capabilities. The increasing realism and accessibility of I2V generation have spurred interest in protecting images against video synthesis.

Image Protection against Generative Models. Protective perturbations aim to reduce the utility of images for downstream generation or editing by disrupting key components such as latent encoders, denoising dynamics, or attention. PhotoGuard [37] introduced diffusion-targeted image immunization to prevent unauthorized edits, and Distraction [28] improved practicality by focusing on attention-level objectives with reduced memory cost. DiffusionGuard [10] strengthens robustness under challenging masked-editing settings, while AdvPaint [22] explicitly targets attention interactions to defend against inpainting manipulation. Beyond editing, Glaze [38] and AdvDM [25] protect against artistic style imitation, and PRIME [24] extends perturbation-based protection to video content. Most closely related to our setting, I2VGuard [13] studies safeguards against diffusion-based I2V generation by degrading spatio-temporal consistency, but its optimization can require generating auxiliary videos and additional model passes. These observations motivate more efficient I2V protection mechanisms that directly target text-dependent interactions to induce generation failures.

3 Background

Diffusion Models. Diffusion models are trained to reverse a noise process that gradually corrupts data. A noisy latent z_t at the timestep t is expressed as:

$$z_t = \sqrt{\bar{\alpha}_t}\, z_0 + \sqrt{1-\bar{\alpha}_t}\, \epsilon, \quad \epsilon \sim \mathcal{N}(0, I), \tag{1}$$

where z_0 is the clean latent and $\bar{\alpha}_t$ is the noise schedule. A denoising network ϵ_θ (*e.g.*, U-Net [35], DiT [31]) is trained to predict ϵ from z_t at timestep t, under conditioning input c (*e.g.*, text embeddings). In diffusion-based I2V systems, such conditioning is typically injected through attention mechanisms that control how signals from different modalities are combined [34, 43].

Conditioning Mechanisms in Diffusion Models. We focus on two representative conditioning designs commonly used in diffusion-based I2V models: *Full-Attention* and *Cross-Attention*. To describe both architectures with a unified perspective, we distinguish *text-dependent* interactions that explicitly access text tokens, from *visual-only* interactions that operate on image and video tokens only.

Full-Attention. In *Full-Attention* architectures (*e.g.*, CogVideoX [51]), video tokens jointly attend to video, image, and text tokens under shared softmax normalization. Under a shared softmax, modalities compete for attention mass [43],

which directly determines their relative contribution to the latent update.

$$[\mathcal{A}_{z,z}, \mathcal{A}_{z,i}, \mathcal{A}_{z,\text{txt}}] = \text{softmax}([A_{z,z}, A_{z,i}, A_{z,\text{txt}}]) \tag{2}$$

$$z^{\ell}_{FA} = \mathcal{A}_{z,z} V^{\ell}_{z} + \mathcal{A}_{z,i} V^{\ell}_{i} + \mathcal{A}_{z,\text{txt}} V^{\ell}_{\text{txt}}, \tag{3}$$

where ℓ is the layer index. Here, $A_{z,\cdot}$ denotes the pre-softmax attention logits grouped by modality, $\mathcal{A}_{z,\cdot}$ the corresponding post-softmax weights, and $V^{\ell}_{\cdot}$ the value projections. z^{ℓ}_{FA} denotes the output of the full-attention block. Under our criterion, the term $\mathcal{A}_{z,\text{txt}} V^{\ell}_{\text{txt}}$ is *text-dependent*, while $\mathcal{A}_{z,z} V^{\ell}_{z} + \mathcal{A}_{z,i} V^{\ell}_{i}$ constitutes *visual-only* contributions.

Cross-Attention. In *Cross-Attention* architectures (*e.g.*, LTX-Video [15]), text enters through a dedicated cross-attention residual, while self-attention operates on image and video tokens:

$$z^{\ell}_{SA} = SA^{\ell}([z^{\ell}_{i}, z^{\ell}]) \tag{4}$$

$$z^{\ell}_{CA} = CA^{\ell}([z^{\ell}_{i}, z^{\ell}] + z^{\ell}_{SA}, c_{\text{txt}}) \tag{5}$$

$$[z^{\ell+1}_{i}, z^{\ell+1}] = [z^{\ell}_{i}, z^{\ell}] + z^{\ell}_{SA} + z^{\ell}_{CA}, \tag{6}$$

where z^{ℓ} and z^{ℓ}_{i} denote the video latent and the image-condition latent in the ℓ-th layer, c_{txt} is the text condition, and SA and CA denote the self-attention and cross-attention modules, respectively. In this design, text affects the video latent only through the cross-attention residual z^{ℓ}_{CA}. Thus, z^{ℓ}_{CA} is *text-dependent*, whereas z^{ℓ}_{SA} and the image-conditioned pathway are *visual-only*.

4 Method

In this section, we describe our I2V image protection approach. Our method consists of (1) the proposed Anti-Prompt objective that attenuates text influence during denoising, (2) an encoder attack that further distorts image conditioning, and (3) the overall objective to optimize the perturbation.

4.1 Anti-Prompt Objective

The goal of I2V image protection is to induce salient failures in text-guided I2V generation from a protected image. As shown in Section 1, we observe that removing text prompts from I2V models produces broad generation failures not only in motion quality but also in subject preservation and structural coherence (Figure 2). Anti-Prompt exploits this sensitivity: it optimizes an adversarial perturbation that attenuates text-conditioned interactions during denoising, reproducing, and intensifying the failure patterns seen under prompt removal.

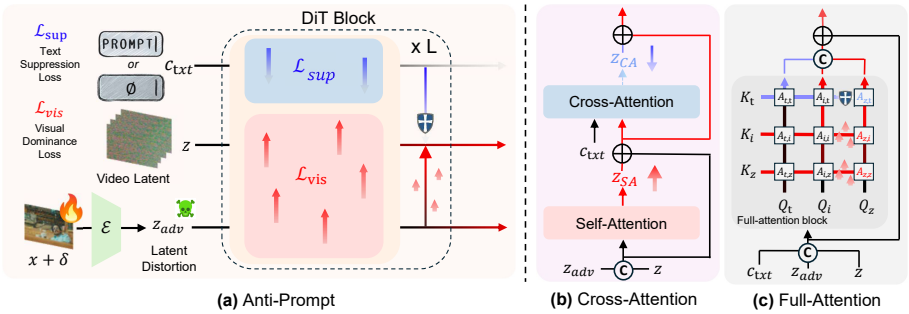

Fig. 3: Overview of Anti-Prompt. (a) We optimize an imperceptible perturbation to suppress text-dependent pathways and promote visual-only dominance. Latent distortion corresponds to the encoder attack. (b) **Cross-Attention:** text enters via a cross-attention residual. We suppress this text-dependent residual and strengthen the self-attention (visual-only) residual. (c) **Full-Attention:** video tokens attend jointly to video, image, and text tokens under a shared softmax. We reduce video-to-text interactions and reinforce visual interactions to promote visual dominance.

Text Suppression. Text suppression denotes attenuating text-dependent interactions during denoising. Since text influence in diffusion models is mediated through attention interactions, these interactions provide a natural target for disruption. We suppress text influence by penalizing direct text-dependent interactions at each layer:

$$\mathcal{L}_{\text{sup}} = \sum_{\ell=1}^{L} \tau^{\ell}. \tag{7}$$

Here, τ^{ℓ} is a scalar that measures the strength of text-dependent pathways at layer ℓ. For *Full-Attention*, we set $\tau^{\ell} = \mathbb{E}[A^{\ell}_{z,\text{txt}}]$, *i.e.*, the average pre-softmax attention logits from video to text tokens, where $\mathbb{E}[\cdot]$ denotes averaging over all attention heads and the corresponding token pairs. We suppress the text-related logits $A^{\ell}_{z,\text{txt}}$ to attenuate text-conditioned pathways. For *Cross-Attention*, we set $\tau^{\ell} = \mathbb{E}[\|z^{\ell}_{CA}\|_2^2]$, *i.e.*, the energy of the cross-attention residual that injects text features into the video latent, thereby shrinking text-dependent residual updates.

Visual Dominance. Text suppression directly weakens text-conditioned pathways, but its effect depends on the specific text tokens encountered during optimization and may transfer less reliably to unseen prompts. Visual dominance complements this by amplifying visual-only interactions regardless of prompt content, reducing the relative contribution of text guidance. By increasing the magnitude of activations that exclude text, we reduce textual influence through softmax competition in *Full-Attention* and residual dominance in *Cross-*

Attention:

$$\mathcal{L}_{\text{vis}} = -\sum_{\ell=1}^{L} \nu^{\ell}. \tag{8}$$

Here, ν^{ℓ} is a scalar that measures the strength of visual-only pathways at layer ℓ. For *Full-Attention*, we set $\nu^{\ell} = \mathbb{E}[A^{\ell}_{z,z}] + \mathbb{E}[A^{\ell}_{z,i}]$, *i.e.*, the average pre-softmax attention logits from video to video and image tokens. Since these modalities compete under the shared softmax, increasing ν^{ℓ} makes the visual terms more competitive and consequently decreases the relative contribution of text guidance. For *Cross-Attention*, we set $\nu^{\ell} = \mathbb{E}[\|z^{\ell}_{SA}\|_2^2]$, *i.e.*, the energy of the self-attention residual, encouraging it to dominate residual updates and reducing the relative contribution of cross-attention text features.

These complementary strategies provide an intuitive realization of our protection objective. Although the formulation differs for *Full-Attention* and *Cross-Attention*, both are designed to attenuate text-guided control and drive denoising toward failures that persist across unseen prompts.

4.2 Encoder Attack

Following [13], we adopt an encoder attack [37] to further weaken image conditioning. Empirically, while prompt omission often yields unstable outputs, some generations can still retain partial coherence depending on the pipeline and scene. To further reduce such residual image-conditioned coherence, we drive the encoder representation toward an uninformative target. Specifically, we perturb the input so that $\mathcal{E}(x+\delta)$ approaches $\mathcal{E}(x_{\text{tar}})$ (*e.g.*, a black image), suppressing usable visual cues for generation:

$$\mathcal{L}_{\text{enc}} = \|\mathcal{E}(x+\delta) - \mathcal{E}(x_{\text{tar}})\|_2^2, \tag{9}$$

where x is the input image, δ the perturbation, $\mathcal{E}$ the VAE encoder, and x_{tar} the uninformative target.

4.3 Overall Loss

To create the final protected image, we obtain the optimal imperceptible perturbation δ by jointly optimizing all three loss components. The final objective is formulated as:

$$\begin{aligned} \min_{\delta} \mathcal{L} &= \mathcal{L}_{\text{sup}} + \lambda_1 \mathcal{L}_{\text{vis}} + \lambda_2 \mathcal{L}_{\text{enc}}, \\ \text{s.t.} \quad & \|\delta\|_{\infty} \leq \epsilon, \end{aligned} \tag{10}$$

where λ_1 and λ_2 are balancing hyperparameters, and the constraint ensures imperceptibility. By jointly optimizing these objectives, our method learns a perturbation that reduces text-conditioned influence and makes text-guided generation more likely to exhibit visible failures. An overview of the proposed Anti-Prompt framework and its detailed optimization procedure is illustrated in Figure 3.

5 Evaluation Protocol for Text-Guided I2V Protection

Evaluating I2V image protection is fundamentally different from evaluating video generation quality. Prior I2V protection work [13] evaluates performance using video generation benchmarks such as VBench I2V [20,21] designed to assess how well a generated video satisfies multiple perceptual criteria. In contrast, the objective of protection is to break image-faithful animation of the protected input. Accordingly, protection effectiveness can often be established by identifying a single clear failure (*e.g.*, subject drift, structural breakdown, temporal instability, and prominent artifacts) rather than by exhaustively scoring fine-grained quality.

Motivated by this distinction, we introduce a failure-finding evaluation protocol that characterizes how generated videos depart from an image-faithful animation. Our protocol scores videos along four interpretable visual dimensions that capture common I2V failure patterns. Each dimension is rated on a discrete 1–5 scale, where higher scores indicate more stable, coherent generation and lower scores indicate more severe failures. Thus, lower scores correspond to stronger protection.

5.1 Evaluation Dimensions

We score the following dimensions because protection can disrupt I2V generation through distinct mechanisms, and mixing them obscures diagnosis. Each dimension is scored on a discrete 1–5 scale, where *5* indicates that the dimension is largely preserved and *1* indicates a severe failure. Accordingly, *lower scores indicate stronger protection.*

Subject Preservation. We assess whether the main subject from the first frame remains present and recognizable across the video. This captures identity drift, disappearance, and replacement, which are often the most salient failures even when the overall rendering still looks plausible.

Structural Consistency. We examine whether the subject's geometry and boundaries remain stable across frames. This highlights warping, tearing, and part-level inconsistencies that can appear even when the subject is still identifiable.

Dynamic Consistency. We check whether motion and state transitions remain temporally coherent and physically plausible. This targets flicker, abrupt jumps, and implausible dynamics that may occur without large appearance changes.

Artifact Suppression. We assess how well visible artifacts are suppressed (*e.g.*, checkerboard patterns, banding, flickering textures, unnatural noise, or rendering glitches). Higher scores indicate minimal visible artifacts, whereas lower scores indicate severe and dominant artifacts.

Table 1: White-box quantitative results on VBench metrics [20, 21]. Lower indicates stronger protection efficacy, since our metrics are formulated to quantify deviations from an image-faithful animation.

Method	I2V Subject	I2V Background	Subject Consistency	Background Consistency	Aesthetic Quality	Imaging Quality	Temporal Flickering	Motion Smoothness
CogVideoX [51]								
Clean	97.15	98.87	95.06	97.38	61.48	70.23	96.99	98.15
I2VGuard	94.58	97.14	93.32	96.17	58.61	65.30	96.47	97.61
Ours	**91.20**	**94.28**	**87.54**	**93.16**	**54.72**	**63.71**	**94.47**	**95.66**
LTX-Video [15]								
Clean	97.70	98.15	95.09	97.87	60.87	68.22	99.04	99.36
I2VGuard	94.36	95.56	90.25	95.00	56.28	62.63	98.23	98.90
Ours	**93.64**	**95.07**	**89.36**	**94.99**	**54.99**	**61.44**	**97.95**	**98.67**

Table 2: VBench [20, 21] results of generated videos with unseen prompt.

Method	I2V Subject	I2V Background	Subject Consistency	Background Consistency	Aesthetic Quality	Imaging Quality	Temporal Flickering	Motion Smoothness
CogVideoX [51]								
Clean	96.87	97.65	94.88	97.40	60.84	70.21	97.41	98.31
I2VGuard	93.51	93.79	92.45	95.19	57.37	63.58	96.34	97.43
Ours	**88.90**	**90.60**	**84.70**	**91.76**	**53.14**	**61.80**	**94.10**	**95.36**
LTX-Video [15]								
Clean	95.85	97.30	90.12	95.50	56.96	64.37	98.53	99.11
I2VGuard	91.83	92.09	83.46	90.28	50.69	56.36	97.35	98.40
Ours	**91.14**	**91.38**	**82.47**	**90.14**	**49.92**	**55.85**	**96.93**	**98.06**

5.2 Video-LLM scoring with Inspectable Evidence

For each generated video, we uniformly sample 4 frames per second, including the first and last. Given the sampled frames, a Video-LLM produces (1) a small set of concrete observations grounded to frame indices, and then (2) a 1–5 score for each dimension. Requiring observations before scoring makes the output more inspectable and reduces arbitrary ratings [27, 57].

To evaluate videos at scale under this protocol, we employ Video-LLMs [2, 26, 29, 52, 53] as evaluators that inspect sampled frames from each generated video and assign dimension-wise scores based on observable visual evidence. The prompt requests a structured output containing frame-grounded observations and per-dimension scores. We release the prompt template and parsing scripts to facilitate reproducibility. We report both the per-dimension scores and their average as a compact summary statistic.

Video-LLM–based evaluation is inherently imperfect and may exhibit model-specific biases or stochastic variability. We therefore complement it with standard VBench metrics and validate the reliability through a human study in Section 6. As shown in Table 6, the human judgments align well with our protocol-induced rankings, supporting its use as a scalable failure-finding evaluation. Additional evaluator repeatability and cross-evaluator consistency analyses are provided in the supplementary material.

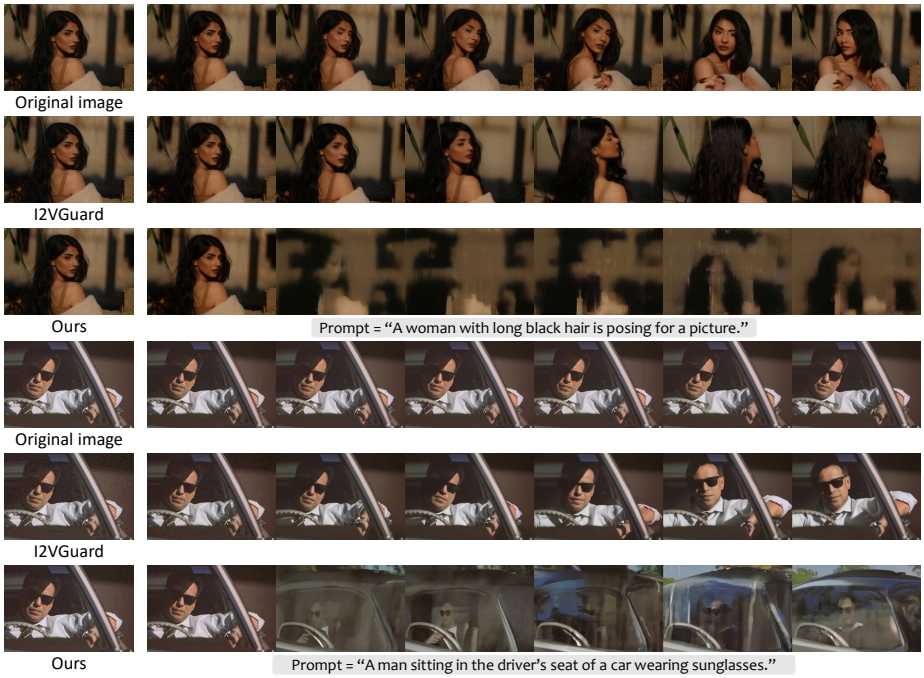

Fig. 4: Qualitative results on CogVideoX [51]. We compare generations from the original image, I2VGuard [13], and Ours.

6 Experiments

Implementation Details. Following I2VGuard [13], we set the PGD hyperparameters to $K = 50$, $\epsilon = 8/255$, and $\alpha = 1/255$ for all experiments. We unroll $T = 10$ denoising steps during optimization. Unless specified otherwise, we set $\lambda_1 = 1$ and $\lambda_2 = 1$ for CogVideoX [51], and use $\lambda_1 = 0.1$ and $\lambda_2 = 0.01$ for LTX-Video [15] to account for objective scale differences across architectures. For video generation, we use 720×480 resolution with 49 frames at 8 FPS for CogVideoX, and 704×480 with 121 frames at 25 FPS for LTX-Video.

Experimental Setup. We employ representative open-source image-to-video generators in our text-guided I2V setting. Specifically, we use CogVideoX-5B [51] (full-attention) and LTX-Video-2B [15] (cross-attention) as target models, as they cover the two common conditioning designs and remain feasible to run at scale under practical memory budgets.

Datasets and Metrics. We evaluate on 355 images from the VBench Image-to-Video dataset [21], using the ground-truth text prompts paired with each image for consistent evaluation. We report VBench [20,21] metrics and our proposed failure-finding scores computed with an open-source Video-LLM evaluator

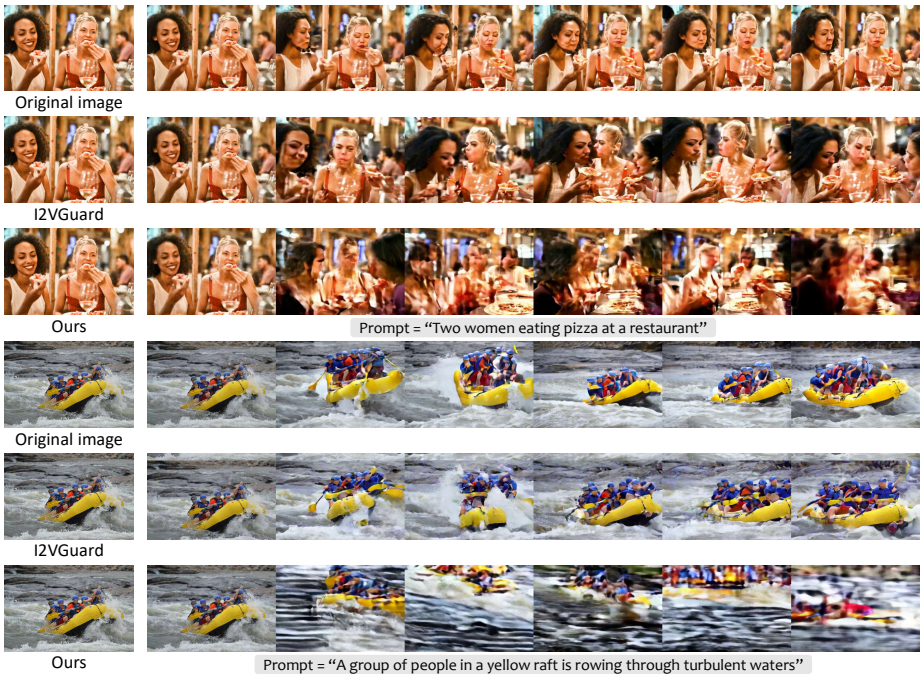

Fig. 5: Qualitative results on LTX-Video [15]. We compare generations from the original image, I2VGuard [13], and Ours.

(Qwen3-VL-8B [2]). We follow the official VBench implementation for metric computation; our failure-oriented evaluation protocol is described in Section 5.

Baselines. We compare our method with I2VGuard [13], which is designed to prevent unauthorized video generation. The original I2VGuard paper reports results on a non-public dataset, so exact numeric matching is not possible. We reproduce I2VGuard using the hyperparameters and design choices described in the paper, and evaluate all methods on the public VBench I2V benchmark under identical generation settings and perturbation budgets. In our setup, the reproduced I2VGuard achieves effectiveness broadly consistent with the reported results. For fair comparison, both methods use the encoder attack by default and share the same uninformative target (a black image).

Qualitative Results We compare generations from the original image and protected images optimized by I2VGuard [13] and Ours. Figures 4 and 5 show that our method more reliably induces salient failures and temporally unstable dynamics than the baseline across both CogVideoX and LTX-Video. Overall, videos generated from our protected images become incoherent and inconsistent across frames, while the input perturbations remain imperceptible. These failures are consistent with the degradation patterns observed under prompt re-

Table 3: Black-box transfer evaluation on VBench metrics [20, 21].

Method	I2V Subject	I2V Background	Subject Consistency	Background Consistency	Aesthetic Quality	Imaging Quality	Temporal Flickering	Motion Smoothness
Optimization: LTX-Video [15], Generation: CogVideoX [51]								
Clean	97.15	98.87	95.06	97.38	61.48	70.23	96.99	98.15
I2VGuard	97.49	98.62	95.41	97.31	60.77	70.43	96.98	98.24
Ours	**95.28**	**96.72**	**94.34**	**96.57**	**59.56**	**68.42**	**96.56**	**97.83**
Optimization: CogVideoX [51], Generation: LTX-Video [15]								
Clean	97.70	98.15	95.09	97.87	60.87	68.22	99.04	99.36
I2VGuard	**94.52**	**95.33**	90.98	95.33	55.78	**61.42**	98.29	98.90
Ours	94.61	95.60	**89.87**	**94.98**	**55.40**	62.63	**98.14**	**98.83**

Table 4: White-box quantitative results under our I2V protection evaluation protocol. Lower scores indicate more severe failures. Higher scores indicate more faithful animation of the input image. Scores are reported as *a/b* for *seen/unseen* prompts.

Method	Subject Preservation	Structural Consistency	Dynamic Consistency	Artifact Suppression	Average
CogVideoX [51]					
Clean	4.86/4.72	4.55/4.22	4.19/3.91	4.79/4.66	4.58/4.38
I2VGuard	4.72/4.71	4.18/3.95	3.81/3.66	2.97/3.13	3.92/3.86
Ours	**4.37**/**4.26**	**3.24**/**2.82**	**2.91**/**2.61**	**2.49**/**2.28**	**3.25**/**2.99**
LTX-Video [15]					
Clean	4.64/3.78	4.01/2.83	3.70/2.85	4.17/3.08	4.13/3.13
I2VGuard	4.24/3.37	3.18/2.21	2.99/**2.17**	2.70/1.95	3.28/2.46
Ours	**4.06**/**3.21**	**3.02**/**2.20**	**2.85**/**2.17**	**2.58**/**1.92**	**3.12**/**2.38**

moval (Figure 2), further supporting our motivation of disrupting text-guided stabilization.

6.1 Quantitative Evaluation

We first evaluate on the standard VBench benchmark [20, 21]. Throughout this section, we use a unified convention: lower scores indicate more severe failures (stronger protection), since our metrics are formulated to quantify deviations from an image-faithful animation. As shown in Table 1, our approach improves protection effectiveness over I2VGuard across various I2V metrics. We further evaluate using our failure-finding protocol with an open-source Video-LLM evaluator (Qwen3-VL-8B [2] in our implementation). As shown in Table 4, our method achieves consistently lower scores across the four dimensions, indicating stronger disruption of image-faithful generation.

Robustness against Unseen Prompts. In real-world scenarios, the defender has no access to the attacker's prompt, making generalization to unseen prompts critical for dependable protection. Accordingly, we evaluate robustness under unseen generation prompts by using GPT-4 [1] to synthesize diverse, plausible video scenarios conditioned on each input image. We note that the generated unseen prompts are typically more informative and dynamic than the original

Table 5: Black-box transfer evaluation on our evaluation protocol.

Method	Subject Preservation	Structural Consistency	Dynamic Consistency	Artifact Suppression	Average
Optimization: LTX-Video [15], Generation: CogVideoX [51]					
Clean	4.86	4.55	4.19	4.79	4.58
I2VGuard	**4.84**	4.44	4.09	4.50	4.47
Ours	**4.84**	**4.31**	**3.92**	**3.51**	**4.15**
Optimization: CogVideoX [51], Generation: LTX-Video [15]					
Clean	4.64	4.01	3.70	4.17	4.13
I2VGuard	4.24	3.23	3.14	2.97	3.40
Ours	**4.07**	**3.11**	**2.95**	**2.86**	**3.25**

Table 6: Human study and protocol alignment. The study uses a blinded triplet comparison over Clean, I2VGuard, and Ours. A higher worst rate indicates stronger protection-induced degradation.

Method	Cons. Rank ↓	Plaus. Rank ↓	Best (%) ↑	Worst (%) ↑
Clean	1.50	1.57	59.5	10.0
I2VGuard	1.77	1.74	33.5	10.9
Ours	2.72	2.69	7.0	79.1
Protocol–human ranking agreement				
Criterion	Spearman ρ		Kendall τ	
Overall ranking	0.84		0.71	
Consistency	0.81		0.68	
Plausibility	0.79		0.66	

VBench [20] prompts, making generation itself more challenging. As a result, absolute scores tend to degrade for all methods, and this effect is particularly pronounced for LTX-Video [15]. Nevertheless, our method maintains a clear advantage over I2VGuard [13] in this harder regime, indicating that the learned perturbations transfer robustly even when the prompt distribution shifts (Table 2 and 4). We use the same unseen prompts for all methods and clean generations. The unseen prompts are provided in the supplementary material.

Black-box Transferability. We further conduct cross-architecture black-box experiments to assess protection transferability. As shown in Table 3, Anti-Prompt is generally favorable to I2VGuard [13] in black-box transfer. When optimized on LTX-Video [15], it outperforms I2VGuard across all reported VBench metrics, while the CogVideoX [51]-to-LTX setting shows a few exceptions on isolated dimensions but still favors Anti-Prompt overall. In Table 5, Anti-Prompt exhibits a clearer and more consistent advantage under our evaluation protocol in Sec. 5. Taken together, these results indicate effective black-box transfer for Anti-Prompt, with its advantage appearing more consistently under the proposed failure-finding evaluation.

Table 7: Imperceptibility comparison.

Model	Method	DISTS ↓	LPIPS ↓	PSNR ↑	SSIM ↑
CogVideoX	I2VGuard	0.138	0.246	32.86	0.838
	Ours	**0.111**	**0.196**	**35.10**	**0.906**
LTX-Video	I2VGuard	**0.130**	0.237	32.35	0.816
	Ours	0.138	**0.228**	**34.20**	**0.892**

Table 8: Ablation of Anti-Prompt components.

Method	I2V Subject	I2V Background	Subject Consistency	Background Consistency	Aesthetic Quality	Imaging Quality	Temporal Flickering	Motion Smoothness	Video-LLM [2] Average↓
$\mathcal{L}_{\text{enc}}$	94.47	97.71	91.96	96.76	57.83	64.22	96.23	97.12	4.11
$\mathcal{L}_{\text{enc}} + \mathcal{L}_{\text{sup}}$	94.77	97.22	92.94	96.22	58.33	66.55	96.57	97.53	3.92
$\mathcal{L}_{\text{enc}} + \mathcal{L}_{\text{vis}}$	92.92	95.91	90.53	94.90	56.38	66.95	95.32	96.66	3.59
Ours	91.20	94.28	87.54	93.16	54.72	63.71	94.47	95.66	3.25

Human Study and Protocol Alignment. To verify that the generated failures are perceptually meaningful, we conduct a blinded human study on generated video triplets. Participants compare the results from Clean, I2VGuard, and Ours for the same image and prompt, and select the best and worst videos based on consistency and plausibility. To reduce the cognitive burden of comparing video triplets, we ask participants to judge two higher-level criteria rather than all four protocol dimensions: consistency, which mainly reflects subject preservation and structural consistency, and plausibility, which mainly reflects dynamic consistency and artifact suppression. As shown in Table 6, generations from our protected images are most frequently selected as the worst, indicating stronger human-perceived degradation. The protocol-induced rankings also agree well with aggregated human preferences across overall, consistency, and plausibility criteria.

6.2 Analysis on Imperceptibility

Beyond protection effectiveness, practical protective perturbation should be visually imperceptible. We measure DISTS [11], LPIPS [54], PSNR, and SSIM [47], comparing our method with I2VGuard [13]. As shown in Table 7, our perturbations achieve consistently lower LPIPS and higher PSNR/SSIM across both models, indicating improved perceptual fidelity. DISTS is improved on CogVideoX and is slightly higher on LTX-Video.

6.3 Ablation Studies

Impact of Each Proposed Component. We conduct ablation experiments to evaluate the contribution of our two objectives, $\mathcal{L}_{\text{sup}}$ and $\mathcal{L}_{\text{vis}}$, as shown in Table 8. In this analysis, the encoder attack ($\mathcal{L}_{\text{enc}}$) is used as a baseline component. We evaluate the protection efficacy when adding $\mathcal{L}_{\text{sup}}$ only, $\mathcal{L}_{\text{vis}}$ only, and both combined. All ablation experiments are conducted in the white-box setting on CogVideoX [51].

Table 9: Efficiency comparison.

Model	Method	Video Gen. (PFLOPs)	δ Optim. (PFLOPs)	Total FLOPs (PFLOPs)	Peak Memory (GB)
CogVideoX-5B	I2VGuard	16.96	37.29	54.25	51.56
	Ours	**0**	**26.82**	**26.82**	**48.98**
LTX-Video-2B	I2VGuard	1.896	3.517	5.413	35.18
	Ours	**0**	**2.609**	**2.609**	**28.86**

Adding $\mathcal{L}_{\text{sup}}$ alone shows mixed effects on VBench metrics, where some scores slightly increase, but consistently improve the Video-LLM failure score. This suggests that text suppression induces failures that are visible to a Video-LLM evaluator inspecting individual frames but are not fully captured by the automated VBench pipeline. Adding $\mathcal{L}_{\text{vis}}$ alone produces consistent improvements across both evaluation frameworks. Their combination achieves the lowest scores on both VBench and the Video-LLM protocol, confirming that text suppression and visual dominance are complementary.

Efficiency Analysis. As shown in Table 9, our method significantly reduces runtime and memory consumption compared to I2VGuard [13]. This is because our losses operate on intermediate attention statistics within a single denoising pass, whereas I2VGuard requires generating a reference video as an additional forward pass. This reduced memory footprint makes the optimization feasible under practical GPU budgets and leaves headroom to scale to larger I2V models when more compute becomes available.

7 Conclusion

We introduced Anti-Prompt, an image protection strategy that aims to prevent unauthorized text-guided image-to-video generation from publicly shared images. Anti-Prompt combines text suppression to attenuate text-dependent interactions with a visual dominance loss that increases robustness to unseen prompts. We further propose a Video-LLM-based failure-finding evaluation protocol that exposes protection-induced breakdowns along four visual dimensions. Extensive experiments on VBench and our protocol across cross-attention and full-attention I2V models demonstrate strong protection with high imperceptibility and improved computational efficiency. Moreover, our method demonstrates generally stronger protection efficacy in unseen-prompt and model-shift evaluation settings.

Acknowledgement

This work was supported by the AI Graduate School Program at POSTECH (RS-2019-II191906 (5%)), the NRF grants (RS-2025-24535146 (10%), RS-2026-25491789 (40%)), the IITP grants (RS-2022-II220926 (25%)) funded by MSIT, and the KRIT grant (No. 21-107-E00-009-02 (20%)) funded by DAPA, Korea.

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
3. Bar-Tal, O., Chefer, H., Tov, O., Herrmann, C., Paiss, R., Zada, S., Ephrat, A., Hur, J., Liu, G., Raj, A., et al.: Lumiere: A space-time diffusion model for video generation. In: SIGGRAPH Asia (2024)
4. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al.: Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127 (2023)
5. Blattmann, A., Rombach, R., Ling, H., Dockhorn, T., Kim, S.W., Fidler, S., Kreis, K.: Align your latents: High-resolution video synthesis with latent diffusion models. In: CVPR (2023)
6. Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M.S., Bohg, J., Bosselut, A., Brunskill, E., et al.: On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258 (2021)
7. Chen, H., Xia, M., He, Y., Zhang, Y., Cun, X., Yang, S., Xing, J., Liu, Y., Chen, Q., Wang, X., et al.: Videocrafter1: Open diffusion models for high-quality video generation. arXiv preprint arXiv:2310.19512 (2023)
8. Chen, H., Zhang, Y., Cun, X., Xia, M., Wang, X., Weng, C., Shan, Y.: Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In: CVPR (2024)
9. Chen, J., Yu, J., Ge, C., Yao, L., Xie, E., Wu, Y., Wang, Z., Kwok, J., Luo, P., Lu, H., et al.: Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. arXiv preprint arXiv:2310.00426 (2023)
10. Choi, J.S., Lee, K., Jeong, J., Xie, S., Shin, J., Lee, K.: Diffusionguard: A robust defense against malicious diffusion-based image editing. In: ICML (2024)
11. Ding, K., Ma, K., Wang, S., Simoncelli, E.P.: Image quality assessment: Unifying structure and texture similarity. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(5), 2567–2581 (2020)
12. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: ICML (2024)
13. Gui, D., Guo, X., Zhou, W., Lu, Y.: I2vguard: Safeguarding images against misuse in diffusion-based image-to-video models. In: CVPR (2025)
14. Guo, Y., Yang, C., Rao, A., Liang, Z., Wang, Y., Qiao, Y., Agrawala, M., Lin, D., Dai, B.: Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. In: ICLR (2024)
15. HaCohen, Y., Chiprut, N., Brazowski, B., Shalem, D., Moshe, D., Richardson, E., Levin, E., Shiran, G., Zabari, N., Gordon, O., et al.: Ltx-video: Realtime video latent diffusion. arXiv preprint arXiv:2501.00103 (2024)
16. He, Y., Yang, T., Zhang, Y., Shan, Y., Chen, Q.: Latent video diffusion models for high-fidelity long video generation. arXiv preprint arXiv:2211.13221 (2022)
17. Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D.P., Poole, B., Norouzi, M., Fleet, D.J., et al.: Imagen video: High definition video generation with diffusion models. arXiv preprint arXiv:2210.02303 (2022)

18. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: NeurIPS (2020)
19. Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., Fleet, D.J.: Video diffusion models. In: NeurIPS (2022)
20. Huang, Z., He, Y., Yu, J., Zhang, F., Si, C., Jiang, Y., Zhang, Y., Wu, T., Jin, Q., Chanpaisit, N., et al.: Vbench: Comprehensive benchmark suite for video generative models. In: CVPR (2024)
21. Huang, Z., Zhang, F., Xu, X., He, Y., Yu, J., Dong, Z., Ma, Q., Chanpaisit, N., Si, C., Jiang, Y., et al.: Vbench++: Comprehensive and versatile benchmark suite for video generative models. arXiv preprint arXiv:2411.13503 (2024)
22. Jeon, J., Kim, W.J., Ha, S., Son, S., Yoon, S.e.: Advpaint: Protecting images from inpainting manipulation via adversarial attention disruption. In: ICLR (2025)
23. Labs, B.F., Batifol, S., Blattmann, A., Boesel, F., Consul, S., Diagne, C., Dockhorn, T., English, J., English, Z., Esser, P., et al.: Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. arXiv preprint arXiv:2506.15742 (2025)
24. Li, G., Yang, S., Zhang, J., Zhang, T.: Prime: Protect your videos from malicious editing. arXiv preprint arXiv:2402.01239 (2024)
25. Liang, C., Wu, X., Hua, Y., Zhang, J., Xue, Y., Song, T., Xue, Z., Ma, R., Guan, H.: Adversarial example does good: preventing painting imitation from diffusion models via adversarial examples. In: ICML (2023)
26. Lin, B., Ye, Y., Zhu, B., Cui, J., Ning, M., Jin, P., Yuan, L.: Video-llava: Learning united visual representation by alignment before projection. In: EMNLP (2024)
27. Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., Zhu, C.: G-eval: Nlg evaluation using gpt-4 with better human alignment. In: EMNLP (2023)
28. Lo, L., Yeo, C.Y., Shuai, H.H., Cheng, W.H.: Distraction is all you need: Memory-efficient image immunization against diffusion-based image editing. In: CVPR (2024)
29. Maaz, M., Rasheed, H., Khan, S., Khan, F.: Video-chatgpt: Towards detailed video understanding via large vision and language models. In: ACL (2024)
30. Mirsky, Y., Lee, W.: The creation and detection of deepfakes: A survey. ACM computing surveys (CSUR) **54**(1), 1–41 (2021)
31. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: ICCV (2023)
32. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023)
33. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 (2022)
34. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022)
35. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI (2015)
36. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. NeurIPS (2022)
37. Salman, H., Khaddaj, A., Leclerc, G., Ilyas, A., Madry, A.: Raising the cost of malicious ai-powered image editing. In: ICML (2023)
38. Shan, S., Cryan, J., Wenger, E., Zheng, H., Hanocka, R., Zhao, B.Y.: Glaze: Protecting artists from style mimicry by {Text-to-Image} models. In: USENIX Security (2023)

39. Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., Hu, Q., Yang, H., Ashual, O., Gafni, O., et al.: Make-a-video: Text-to-video generation without text-video data. arXiv preprint arXiv:2209.14792 (2022)
40. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
41. Teng, H., Jia, H., Sun, L., Li, L., Li, M., Tang, M., Han, S., Zhang, T., Zhang, W., Luo, W., et al.: Magi-1: Autoregressive video generation at scale. arXiv preprint arXiv:2505.13211 (2025)
42. Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., Ortega-Garcia, J.: Deepfakes and beyond: A survey of face manipulation and fake detection. Information fusion **64**, 131–148 (2020)
43. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. NeurIPS (2017)
44. Villegas, R., Babaeizadeh, M., Kindermans, P.J., Moraldo, H., Zhang, H., Saffar, M.T., Castro, S., Kunze, J., Erhan, D.: Phenaki: Variable length video generation from open domain textual description. arXiv preprint arXiv:2210.02399 (2022)
45. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., Zeng, J., Wang, J., Zhang, J., Zhou, J., Wang, J., Chen, J., Zhu, K., Zhao, K., Yan, K., Huang, L., Feng, M., Zhang, N., Li, P., Wu, P., Chu, R., Feng, R., Zhang, S., Sun, S., Fang, T., Wang, T., Gui, T., Weng, T., Shen, T., Lin, W., Wang, W., Wang, W., Zhou, W., Wang, W., Shen, W., Yu, W., Shi, X., Huang, X., Xu, X., Kou, Y., Lv, Y., Li, Y., Liu, Y., Wang, Y., Zhang, Y., Huang, Y., Li, Y., Wu, Y., Liu, Y., Pan, Y., Zheng, Y., Hong, Y., Shi, Y., Feng, Y., Jiang, Z., Han, Z., Wu, Z.F., Liu, Z.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
46. Wang, X., Yuan, H., Zhang, S., Chen, D., Wang, J., Zhang, Y., Shen, Y., Zhao, D., Zhou, J.: Videocomposer: Compositional video synthesis with motion controllability. NeurIPS (2023)
47. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)
48. Weidinger, L., Rauh, M., Marchal, N., Manzini, A., Hendricks, L.A., Mateos-Garcia, J., Bergman, S., Kay, J., Griffin, C., Bariach, B., et al.: Sociotechnical safety evaluation of generative ai systems. arXiv preprint arXiv:2310.11986 (2023)
49. Wu, J.Z., Ge, Y., Wang, X., Lei, S.W., Gu, Y., Shi, Y., Hsu, W., Shan, Y., Qie, X., Shou, M.Z.: Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In: ICCV (2023)
50. Xing, J., Xia, M., Zhang, Y., Chen, H., Yu, W., Liu, H., Liu, G., Wang, X., Shan, Y., Wong, T.T.: Dynamicrafter: Animating open-domain images with video diffusion priors. In: ECCV (2024)
51. Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., et al.: Cogvideox: Text-to-video diffusion models with an expert transformer. In: ICLR (2025)
52. Zhang, B., Li, K., Cheng, Z., Hu, Z., Yuan, Y., Chen, G., Leng, S., Jiang, Y., Zhang, H., Li, X., et al.: Videollama 3: Frontier multimodal foundation models for image and video understanding. arXiv preprint arXiv:2501.13106 (2025)
53. Zhang, H., Li, X., Bing, L.: Video-llama: An instruction-tuned audio-visual language model for video understanding. In: EMNLP (2023)
54. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018)

55. Zhang, S., Wang, J., Zhang, Y., Zhao, K., Yuan, H., Qin, Z., Wang, X., Zhao, D., Zhou, J.: I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models. arXiv preprint arXiv:2311.04145 (2023)
56. Zheng, D., Huang, Z., Liu, H., Zou, K., He, Y., Zhang, F., Gu, L., Zhang, Y., He, J., Zheng, W.S., et al.: Vbench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. arXiv preprint arXiv:2503.21755 (2025)
57. Zheng, L., Chiang, W.L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., et al.: Judging llm-as-a-judge with mt-bench and chatbot arena. NeurIPS (2023)
58. Zheng, Z., Peng, X., Yang, T., Shen, C., Li, S., Liu, H., Zhou, Y., Li, T., You, Y.: Open-sora: Democratizing efficient video production for all. arXiv preprint arXiv:2412.20404 (2024)