LETHE: DEFENDING AGAINST UNAUTHORIZED IMAGE-TO-VIDEO GENERATION VIA VALUE DIVERSITY COLLAPSE

Anonymous authors
Paper under double-blind review

ABSTRACT

Image-to-video (I2V) diffusion models can animate a single reference image into high-quality, subject-consistent videos, posing threats of misuse as anyone can turn personal photos into fabricated videos without consent. A promising solution is to proactively add an adversarial perturbation to the image before release, such that videos generated from it would be unfaithful or visually corrupted. However, due to an insufficient understanding of what governs how subsequent frames read the reference image in modern I2V models, existing protections struggle to offer strong protection under common perturbation budgets. In this paper, we approach the problem by exploring this underlying reading mechanism. Through a convex geometric analysis, we theoretically identify the spread of the reference value vectors as a key factor governing this reading process, and empirically find that even a mild contraction of this spread can severely degrade the generated video, making it an efficient lever for protection. Guided by this insight, we propose LETHE, a novel protection method that directly disrupts this reading process. LETHE pairs two complementary objectives, where a value collapse objective drives the reference value vectors toward uniformity, while an attention attraction objective draws subsequent frames toward these collapsed tokens to amplify corruption. We further construct two human-centric I2V protection benchmarks and conduct extensive experiments with multiple modern I2V models, which demonstrate the state-of-the-art protection performance and versatility of LETHE.

1 INTRODUCTION

Image-to-video (I2V) diffusion models (Yang et al., 2025; Wan et al., 2025) have achieved remarkable progress in video synthesis. Given a reference image (e.g., a selfie) and a text prompt, these models can generate a high-quality video in which the subject faithfully retains the appearance in the reference image while performing the actions described in the prompt (Huang et al., 2024). However, this capability also enables malicious individuals to animate a personal photo shared online into videos without the owner's consent, such as videos that fabricate the person's words and actions. Since such videos depict the person in natural motion, they can be deceptively realistic and thus pose serious threats to personal privacy and reputation, which have drawn growing public concern and legislative attention (Gui et al., 2025; Song et al., 2026; Kamachee et al., 2025).

To mitigate this threat, one emerging solution is *protective perturbation* (Zhang et al., 2026), which turns the vulnerability of deep learning models to adversarial perturbations (Szegedy et al., 2013) into a means of protecting personal images. The key idea is to proactively add a small, carefully crafted perturbation to the photo before sharing it online, so that diffusion models taking the protected photo as input are misled into producing unfaithful or low-quality outputs (Salman et al., 2023; Liang et al., 2023; Xue et al., 2024; Gui et al., 2025; Song et al., 2026). In this way, forgeries could be blocked at the source, substantially reducing the potential harm from their spread.

While protective perturbations have achieved notable success in image editing and customization (Liang et al., 2023; Xue et al., 2024; Salman et al., 2023), protection against I2V generation is still at an early stage. Early efforts mainly targeted UNet-based models extended from Stable Diffusion, such as pose-driven animation models (Zhou et al., 2025). Only recently, a few studies

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

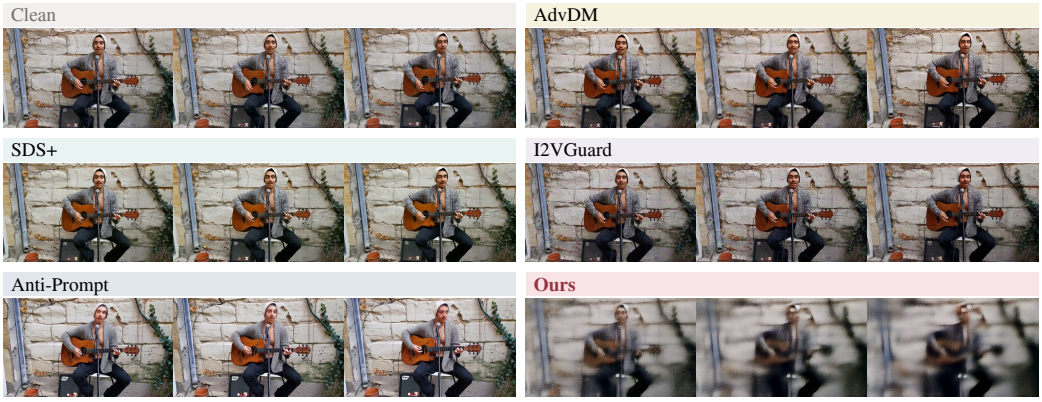


Figure 1: Frames sampled from videos generated from reference images protected by different methods. The perturbation budget is $16/255$ and the model is Wan2.1-I2V-14B (Wan et al., 2025).

have begun to consider DiT-based models (Long et al., 2026; Song et al., 2026), the mainstream architecture of state-of-the-art I2V generation (Shao et al., 2026). However, disrupting modern DiT-based video diffusion models appears considerably more challenging: as shown in Fig. 1, under the commonly used budget of $\epsilon = 16/255$, which typically suffices to effectively disrupt image generative models (Salman et al., 2023; Xue et al., 2024), the protection offered by existing methods is still not satisfactorily strong: they tend to introduce only local artifacts or mild visual changes into the generated videos, while the identity and appearance of the reference image remain somewhat recognizable. This raises concerns about the effectiveness of current protections and highlights the need for stronger protection tailored to modern attention-based I2V models.

To approach this problem, in this paper, we first seek to understand the key factor that governs how I2V models read information from the reference image. In modern I2V models, the reference image is encoded into a set of tokens, from which subsequent frames read primarily through attention operations (Liu et al., 2026). Characterizing this reading process from a convex geometric perspective, we theoretically identify the spread of the reference value vectors, measured by the radius of their minimum enclosing ball, as an upper bound on the diversity of the reference readouts. This analysis points to reference-value diversity as a key target towards disrupting the reading process. We further empirically find that even a mild contraction of this spread suffices to severely degrade the generated video, making it an efficient lever for protection. Motivated by this insight, we propose LETHE, which crafts protective perturbation by disrupting this reading process. LETHE pairs two complementary objectives, where the value collapse objective drives the reference value vectors toward uniformity, aiming to tighten the above bound, while another attention attraction objective draws the attention of subsequent frames toward these collapsed tokens to amplify disruption. In this way, LETHE constrains both the content and the source of the information read by subsequent frames, so that the protected photo can hardly be faithfully animated into a high-quality video. Lastly, since a standardized and publicly available benchmark tailored to human-centric I2V protection is still largely lacking, we construct two high-quality I2V protection benchmarks by re-annotating CelebV-Text (Yu et al., 2023) and Kinetics-700 (Carreira et al., 2019) with carefully curated reference frames and human-verified motion prompts. Extensive experiments on these benchmarks with multiple I2V models confirm the state-of-the-art effectiveness and versatility of LETHE.

In conclusion, our contributions are three-fold:

- We characterize the reference-reading pathway in attention-based I2V models from a convex geometric perspective and identify reference-value diversity as a key quantity that bounds the diversity of the reference readouts. Combining this analysis with direct value interventions, we establish reference-value diversity as a useful and efficient target for protective perturbations.
- We propose LETHE, a novel protection method that safeguards images against unauthorized I2V, by directly disrupting how these models read the reference image. With a pair of complementary objectives, LETHE jointly constrains the content and the source of the information read by subsequent frames, thereby effectively preventing them from reproducing the reference image.
- We construct two human-centric I2V protection benchmarks with carefully curated reference frames and human-verified motion prompts, and extensively evaluate LETHE on various I2V mod-

els. Results show that LETHE consistently outperforms existing methods, transfers to models not involved in the optimization, and remains effective against transformations and various attacks.

2 RELATED WORK

Image-to-Video Diffusion Models. Given a reference image and a prompt, I2V diffusion models aim to generate videos that preserve the appearance of the reference image while following the motion described by the prompt, and they have been widely adopted in content creation and animation (Wang et al., 2026). Early I2V models, such as SVD (Blattmann et al., 2023), are mostly built upon UNet architectures extended from text-to-image models. Constrained by their model capacity and spatiotemporal modeling ability, these models remain limited in generation fidelity and temporal consistency (Wang et al., 2026). More recent models, such as CogVideoX (Yang et al., 2025), Wan (Wan et al., 2025), and LTX-Video (HaCohen et al., 2024), shift to DiT architectures with full spatiotemporal attention, substantially improving subject fidelity, generation quality, and temporal coherence. These models typically encode the reference image into latent tokens aligned with the first frame, and subsequent frames primarily interact with these tokens through attention to acquire visual information. This reading pathway plays a key role in video generation, as evidenced by recent long video generation methods (Liu et al., 2026; Yi et al., 2025), which demonstrate that retaining the first-frame tokens as attention anchors can help improve long-range subject consistency. In this work, we show that this reading pathway also offers an efficient lever for protection.

Protective Perturbations against Unauthorized Visual Generation. While generative models have greatly facilitated visual content creation, they are also increasingly misused, e.g., for unauthorized style mimicry and identity forgery (Zhang et al., 2026). To prevent such misuse, protective perturbations add small adversarial noise to images before release, aiming to block unauthorized generation at the source. In the era of diffusion models, PhotoGuard (Salman et al., 2023) and AdvDM (Liang et al., 2023) are among the earliest to protect images from being edited and customized. These methods typically perturb the image's VAE latent or maximize the denoising loss, so that models like Stable Diffusion cannot faithfully process the protected images. To date, this research area remains highly active, with subsequent studies further improving efficiency (Xue et al., 2024) and enhancing protection in specific scenarios (Li et al., 2025; Jia et al., 2026).

With the rapid advancement of I2V models and the growing concerns over their misuse, recent studies have begun to extend protective perturbations to I2V generation (Gui et al., 2025; Song et al., 2026). Early efforts mainly target UNet-based models (Pang et al., 2024; Zhou et al., 2025). Since these models largely inherit their components from image diffusion models, these methods naturally build upon designs from image protection, e.g., disrupting the encoders. As DiT-based models become the mainstream, more recent studies further introduce objectives on intermediate features or the denoising process (Long et al., 2026; Song et al., 2026), demonstrating the promise on modern I2V models. However, designing such objectives is inherently challenging: in DiT-based models, the reference image's features propagate to subsequent frames through deeply stacked spatiotemporal attention, and it remains largely unclear which internal factor governs how much of the reference is reproduced. Without such an understanding, protection objectives have to be defined on proxies that are only indirectly related to this reading process, and part of the budget may be spent on features that contribute little to what subsequent frames read, limiting the defenses' effectiveness. In this paper, we aim to fill this gap by examining how I2V models read the reference image, which allows us to concentrate the budget more directly on the reading process and unlock more effective protection.

3 MOTIVATING ANALYSIS

As discussed in Sec. 2, understanding how subsequent frames read from the reference image is central towards effective protection. In this section, we first formalize this reading process in modern I2V models and then analyze the factors influencing it, which motivates the design of our protection.

The Reading Pathway. Given a reference image x_{ref} and a text prompt y, I2V models such as Wan 2.1 and CogVideoX generate a video by starting from Gaussian noise z_T and iteratively denoising it. At each step, the current latent z_t is split into spatio-temporal patches and linearly projected into tokens, on which the network operates. The visual content of x_{ref} is encoded and carried by N tokens at specific positions (typically those of the first frame), which we refer to as reference tokens.

Inside the denoising network, the MLP and the normalization layers act on each token separately, and attention is the main operator that moves information between tokens (Yang et al., 2025; Wan et al., 2025). Concretely, within one head of one layer, each reference token i is projected into a key vector $k_i = W_K h_i$ and a value vector $v_i = W_V h_i$, where h_i is its hidden state. The key serves for matching and encodes what the token depicts, while the value carries the appearance features that are transferred. In turn, a subsequent-frame token is projected into a query $q_m = W_Q h_m$, which is matched against the keys, and the token receives the values in proportion to the match. For instance, when x_{ref} depicts a person, the query of the token generating the face in a subsequent frame matches the keys of the facial tokens, and it then receives the facial values, while the token generating the background likewise receives the background values. In this way, each position obtains the content it needs, and the video reproduces the visual features of x_{ref}. Formally, let o_m denote the content that the m-th subsequent-frame token reads from the reference tokens in this head (with α_{mi} renormalized within the N reference tokens), we have:

$$o_m = \sum_{i=1}^{N} \alpha_{mi}\, v_i, \qquad \alpha_{mi} = \frac{\exp\!\left(q_m^\top k_i/\sqrt{d}\right)}{\sum_{i'=1}^{N} \exp\!\left(q_m^\top k_{i'}/\sqrt{d}\right)}. \tag{1}$$

Theoretical Analysis from a Convex Hull Perspective. One can observe from Eq. (1) that the weights α_{mi} are non-negative and sum to one, so that o_m corresponds to a point in the convex hull (Ball et al., 1997) of the reference values. Specifically, since $\alpha_{mi} \geq 0$ and $\sum_i \alpha_{mi} = 1$, the weight vector $\alpha_m = (\alpha_{m1}, \ldots, \alpha_{mN})^\top$ lies in the probability simplex $\Delta^{N-1} := \{\alpha \in \mathbb{R}^N \mid \alpha_i \geq 0,\ \sum_i \alpha_i = 1\}$. The map $\alpha \mapsto \sum_i \alpha_i v_i$ is linear and sends this simplex to the convex hull of $\{v_i\}$, with the vertices mapped to the individual values. Therefore, we obtain

$$o_m \in \mathcal{H}(V) := \operatorname{conv}(v_1, \ldots, v_N) = \left\{ \sum_{i=1}^{N} \alpha_i v_i \;\middle|\; \alpha \in \Delta^{N-1} \right\}, \tag{2}$$

i.e., what a subsequent-frame token can read from the reference tokens is confined to the convex hull of the reference tokens' value vectors $\{v_i\}$. Since a convex hull is determined by its vertices, the attention weights can only select a point within it as the readout o_m and cannot read content beyond what $\{v_i\}$ span. In our previous example, the video can reproduce distinct visual content (e.g., the face and the background) because $\mathcal{H}(V)$ is large enough for different queries to reach points that express completely different visual features. Thus, the size of $\mathcal{H}(V)$ is a key quantity that governs the range of content subsequent frames, as formally analyzed in the following proposition:

Proposition 1 (Value Spread Bounds Readout Diversity)**.** *Let $\bar{o} = \frac{1}{M}\sum_{m=1}^{M} o_m$, and let $R(V) = \min_c \max_i \|v_i - c\|$ denote the radius of the minimum enclosing ball of $\{v_1, \ldots, v_N\}$. Suppose the M subsequent-frame tokens read from the reference tokens according to Eq. (1), i.e., $o_m = \sum_i \alpha_{mi} v_i$. Then for any M and any attention weights $\alpha_m \in \Delta^{N-1}$,*

$$\frac{1}{M}\sum_{m=1}^{M} \left\|o_m - \bar{o}\right\|^2 \leq R(V)^2. \tag{3}$$

Proposition 1 shows that the spread of the reference values bounds the variance of their readouts across subsequent-frame tokens. A smaller $R(V)$ imposes a lower ceiling on this diversity, regardless of how attention is distributed over the reference tokens. In the extreme case where all reference values coincide, every normalized reference readout in this head equals the same vector. This observation motivates reducing reference-value diversity to constrain the reading pathway. We provide the proof for Proposition 1 and offer further discussions in Appendix A.

Empirical Validation. Proposition 1 shows that, in a single attention operation, the diversity of the reference values bounds what subsequent-frame tokens can read from them. However, in reality, the generation process involves multiple layers, denoising steps, and residual connections, and how the quality and fidelity of the generated video depend on the diversity of the reference values remains unclear from the bound alone. To empirically validate its impact, we conduct an intervention experiment on real video generation models, in which we directly contract the values of the reference tokens and observe how the generated video changes. Specifically, we replace the value of each reference token with $v_i \leftarrow c + \gamma\,(v_i - c)$, where c is the center of the minimum enclosing ball and $\gamma \in [0, 1]$, so that $R(V)$ shrinks linearly with γ, with $\gamma = 1$ recovering the original model and $\gamma = 0$ corresponding to the extreme case where the hull degenerates to a point. The intervention is

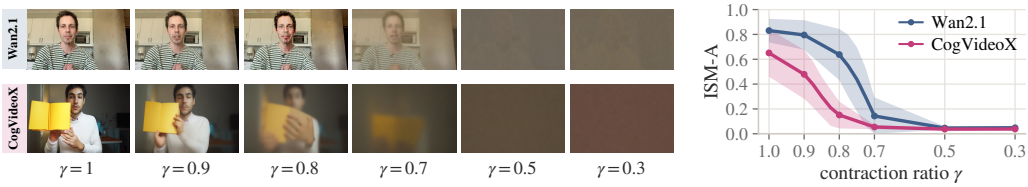


Figure 2: Effect of contracting the reference values toward their center by a ratio γ. **Left**: generated frames under different γ. **Right**: the ArcFace identity similarity (ISM-A) between generated frames and the reference, versus γ. A mild contraction already severely degrades the generated videos.

applied at every layer and the forward pass otherwise proceeds unchanged, with no other component (queries, keys, text condition, or noise seed) intervened.

As shown in Fig. 2, contracting $R(V)$ by only $20\%(\gamma = 0.8)$ already introduces visible artifacts on the face, and at 30% ($\gamma = 0.7$) the video largely collapses into degenerated color blocks with no recognizable subject. Moreover, even when the intervention is applied to only a subset of layers and denoising steps, a contraction of 30% also suffices to almost completely destroy the video. We also verify that random perturbations of the same magnitude do not cause comparable degradation. These results confirm that reducing the diversity of the reference values is an efficient lever for protection.

4 Methodology

Problem Formulation. We formulate the problem as follows. An image owner (the defender) wants to publish an image x_{ref} (e.g., a selfie), and does not want it to be used as the reference image of an I2V model to fabricate videos. To this end, the defender releases a protected image $x' = x_{\text{ref}} + \delta$. Following prior work on protective perturbations (Gui et al., 2025; Song et al., 2026), we adopt the commonly used ℓ_∞ budget with $\|\delta\|_\infty \leq \epsilon$. Then, an attacker may collect x' from the Internet and feed it, possibly after input transformations, into an I2V model with an arbitrary prompt y and random noise z_T. The defender's goal is to craft the protective perturbation δ such that the generated video can no longer faithfully reproduce x_{ref} with high quality.

Value Collapse Objective. Proposition 1 identifies the spread of the reference values as an upper bound on reference-readout diversity, while our intervention experiment shows that contracting these values can substantially disrupt generation. This motivates us to disrupt video generation by collapsing the values of the reference tokens. However, $R(V)$ is inconvenient to optimize directly: computing it requires solving a minimum enclosing ball problem at every forward pass, and its gradient is sparse, being nonzero only for the few values lying on the boundary. Moreover, the absolute scale of $R(V)$ varies across layers and heads, which makes it hard to be aggregated.

To this end, we propose a value collapse loss, which retains the target of $R(V)$ while avoiding the aforementioned difficulties. Let $V^{(\ell,h,t)}(x') = \{v_i\}_{i=1}^N$ denote the values of the reference tokens at layer ℓ, head h, and denoising step t of the surrogate model, $\bar{v}$ their mean, and $\mathcal{B}$ the set of layers and heads on which the objectives are imposed. We define the value collapse loss $\mathcal{L}_{\text{vc}}$ as follows:

$$\mathcal{L}_{\text{vc}}(\delta) \;=\; \mathbb{E}_t \sum_{(\ell,h)\in\mathcal{B}} \frac{\sum_{i=1}^N \left\| v_i^{(\ell,h,t)}(x_{\text{ref}}+\delta) - \bar{v}^{(\ell,h,t)} \right\|^2}{\sum_{i=1}^N \left\| v_i^{(\ell,h,t)}(x_{\text{ref}}+\delta) \right\|^2}. \tag{4}$$

Intuitively, since $\sum_i \|v_i\|^2 = \sum_i \|v_i - \bar{v}\|^2 + N\|\bar{v}\|^2$, each term in the sum is the fraction of the total value energy that is not shared by all reference tokens. Minimizing it pushes the energy onto the common component $\bar{v}$ and drives all reference tokens toward a single vector. Unlike $R(V)$, whose gradient involves only the few most distant values, this objective has a gradient involving all reference tokens. Moreover, for nonzero total value energy, each term is differentiable, bounded in $[0, 1]$, and invariant to a common nonzero scaling of the values, which allows direct aggregation across layers and heads. We further show in Appendix B that each summand in Eq. (4) directly bounds the normalized readout diversity, and that it vanishes if and only if the normalized $R(V)^2$ does, with both relations supported empirically in Appendix D.1.

Attention Attraction Objective. In our previous analysis, the attention weights in Eq. (1) are normalized within the reference tokens. However, in practice, a subsequent-frame token distributes

its attention over the entire sequence, including other video tokens and embedding tokens of the reference image (Wang et al., 2026). Thus, how much the collapsed reference values contribute to its output depends on the attention allocated to them, which is not directly controlled by $\mathcal{L}_{\text{vc}}$.

To further amplify and stabilize the collapse's effect, we introduce another objective that attracts the attention of subsequent-frame tokens to the reference tokens. Our key insight is that the attention mass is governed by the logits between the subsequent-frame queries and the keys of the reference tokens, which could be influenced through δ. Thus, we propose to raise the average attention logit from the subsequent-frame queries to the reference keys. Let $\bar{q}^{(\ell,h,t)} = \frac{1}{M}\sum_m q_m$ denote the mean query of the subsequent-frame tokens and $\bar{k}^{(\ell,h,t)}(x_{\text{ref}}+\delta) = \frac{1}{N}\sum_i k_i$ the mean key of the reference tokens at layer ℓ, head h, and denoising step t. The attention attraction loss $\mathcal{L}_{\text{aa}}$ is defined as:

$$\mathcal{L}_{\text{aa}}(\delta) \;=\; -\,\mathbb{E}_t \sum_{(\ell,h)\in\mathcal{B}} \frac{\bar{q}^{(\ell,h,t)\top}\,\bar{k}^{(\ell,h,t)}(x_{\text{ref}}+\delta)}{\sqrt{d}}. \tag{5}$$

Intuitively, $\bar{q}^\top \bar{k}/\sqrt{d}$ is the average attention logit from the subsequent-frame queries to the reference keys, and raising it tends to increase the share of attention received by the reference tokens (Kim et al., 2023). We treat $\bar{q}$ as a constant, since the queries are given by the attacker's generation process and the perturbation should act through the keys alone, which also allows $\mathcal{L}_{\text{aa}}$ to be efficiently computed. Moreover, since the keys are usually QK-normalized (Wan et al., 2025; Yang et al., 2025), the average logit is bounded, and $\mathcal{L}_{\text{aa}}$ cannot be reduced by merely inflating the key norm.

Overall Objective. To get the final protected image, we obtain the perturbation δ by jointly optimizing the two objectives:

$$\min_{\|\delta\|_\infty \le \epsilon} \mathcal{L}_{\text{Lethe}}(\delta) \;=\; \mathcal{L}_{\text{vc}}(\delta) + \alpha\,\mathcal{L}_{\text{aa}}(\delta), \tag{6}$$

where ϵ is the perturbation budget and $\alpha = 1$ balances the two terms. Intuitively, $\mathcal{L}_{\text{vc}}$ and $\mathcal{L}_{\text{aa}}$ form a complementary pair that jointly shapes the reference-frame tokens. Specifically, $\mathcal{L}_{\text{vc}}$ suppresses deviations of the reference values from their mean, making them provide subsequent frames with highly homogeneous values. Complementarily, $\mathcal{L}_{\text{aa}}$ steers the reference keys, so that the readout of subsequent frames is largely drawn from these impoverished tokens. Together, LETHE leaves subsequent frames with little distinguishable information from the reference image and limited capacity to compensate through other tokens, leading to notable and persistent degradation in the generated video. We provide the implementation details and the pseudocode of LETHE in Appendix C.2.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

Models. We mainly conduct our experiments on 2 SOTA native image-to-video generative models, i.e., Wan2.1-I2V-14B (Wan et al., 2025) and CogVideoX-5B-I2V (Yang et al., 2025), which represent the mainstream DiT-based paradigm with distinct training objectives and attention designs. To evaluate transferability, we also evaluate on LTX-Video (HaCohen et al., 2024), SVD (Blattmann et al., 2023), LightX2V (LightX2V Team, 2026), and AccVideo (Zhang et al., 2025), covering models with completely different architectures and training data, as well as popular community models.

Baselines. We compare LETHE with 5 representative baselines that protect images against diffusion models, including PhotoGuard (Salman et al., 2023), AdvDM (Liang et al., 2023), SDS+ (Xue et al., 2024), I2VGuard (Gui et al., 2025), and AntiPrompt (Song et al., 2026), where AntiPrompt is a very recent work released shortly before our submission and represents the state of the art that specifically targets disrupting I2V generation. All baselines are reproduced following their original papers and official code with recommended hyperparameters. To ensure fairness, all methods share the same perturbation budget ($\epsilon = 16/255$ by default) and inference configurations.

Datasets. Existing I2V benchmarks (Huang et al., 2024) mainly target general-purpose content with few samples of human faces and bodies, while human-centric datasets (Yu et al., 2023; Carreira et al., 2019) carry annotations designed for other tasks (e.g., attribute descriptions or action labels) that cannot directly serve as I2V prompts. Therefore, we construct and annotate two evaluation datasets built upon CelebV-Text (Yu et al., 2023) and Kinetics-700 (Carreira et al., 2019), which

Table 1: **Main Results.** White-box protection performance on Wan2.1 and CogVideoX on the CelebV-Text and Kinetics-700 datasets. All methods use an ℓ_∞ budget of $\epsilon = 16/255$.

		CelebV-Text				Kinetics-700			
Model	Method	ISM-A ↓	ISM-F ↓	QAF ↓	QAV ↓	ISM-A ↓	ISM-F ↓	QAF ↓	QAV ↓
Wan2.1	Clean	0.8173	0.8725	0.9123	0.9462	0.7584	0.8242	0.7974	0.8381
	AdvDM	0.6082	0.7470	0.6199	0.6866	0.5621	0.6914	0.5623	0.6176
	SDS+	0.6758	0.8018	0.5894	0.6469	0.6119	0.7377	0.5449	0.5969
	PhotoGuard	0.7344	0.8224	0.6368	0.6967	0.6945	0.7747	0.5824	0.6404
	I2VGuard	0.5818	0.7263	0.6285	0.7045	0.5666	0.6952	0.5843	0.6497
	AntiPrompt	0.5237	0.7122	0.8447	0.8985	0.4956	0.6663	0.7341	0.8011
	Ours	**0.4280**	**0.5308**	**0.3970**	**0.4976**	**0.4125**	**0.4957**	**0.3852**	**0.4745**
CogVideoX	Clean	0.6611	0.7785	0.8561	0.9123	0.5742	0.7105	0.7623	0.8198
	AdvDM	0.4356	0.6339	0.5315	0.5750	0.3854	0.5762	0.5191	0.5514
	SDS+	0.4093	0.6104	0.5050	0.5549	0.3729	0.5617	0.4879	0.5273
	PhotoGuard	0.4440	0.6236	0.5538	0.6532	0.4302	0.5912	0.5619	0.6519
	I2VGuard	0.4920	0.6831	0.7035	0.7751	0.4535	0.6215	0.5763	0.6527
	AntiPrompt	0.3424	0.5598	0.7131	0.7769	0.3522	0.5621	0.7046	0.7650
	Ours	**0.2968**	**0.4895**	**0.4235**	**0.4946**	**0.2523**	**0.4299**	**0.4223**	**0.4828**

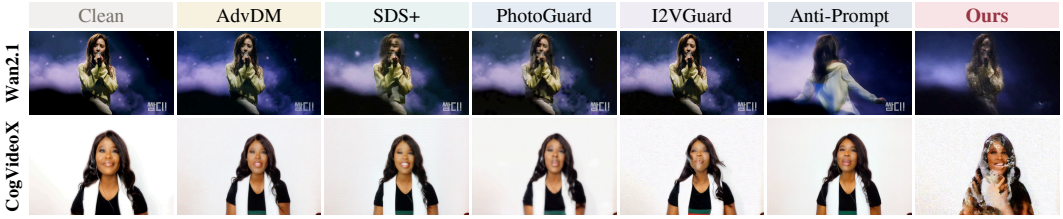


Figure 3: **Qualitative comparison** between LETHE and baselines on Wan2.1 and CogVideoX.

correspond to close-up facial portraits and human actions in natural scenes, respectively. For each dataset, we select 350 high-quality videos with distinct identities, extract one frame with a clearly visible face from each video as the conditioning image, and annotate each image with several I2V prompts describing the appearance and the motion of the subject (see selection criteria, annotation procedure, and manual verification in Appendix C.1). We will release the re-annotated datasets to facilitate future research on human-centric I2V protection.

Metrics. We evaluate protection performance from both identity preservation and generation quality perspectives. For identity preservation, we report the Identity Score Matching (ISM, ↓) (Van Le et al., 2023; Huang et al., 2026), i.e., the cosine similarity between the face embeddings of the generated frames and that of the clean conditioning image, averaged over frames. To reduce the dependence on a single recognizer, we report ISM under two face recognition models, ArcFace (Deng et al., 2019) and FaceNet (Schroff et al., 2015), denoted as ISM-A and ISM-F, respectively. For generation quality, we adopt Q-Align (Wu et al., 2024) to score the generated videos at both the frame level (QAF, ↓), which reflects the visual quality of individual frames, and the video level (QAV, ↓), which further accounts for video-level quality (e.g., temporal consistency and motion naturalness). Lower ISM, QAF, and QAV indicate that the generated video preserves less identity and has lower visual quality, and thus a stronger protection. Additionally, following Song et al. (2026), we adopt an MLLM-based evaluation protocol as a complement to the above metrics (see Appendix C.3).

5.2 Overall Performance

Main Results. We first evaluate all methods on Wan2.1-I2V-14B (Wan et al., 2025) and CogVideoX-5B-I2V (Yang et al., 2025) in the white-box setting. As shown in Tab. 1, LETHE achieves the best performance across all 16 settings spanning 2 models, 2 datasets, and 4 metrics, outperforming all evaluated baselines by a notable margin. For instance, on CelebV-Text with Wan2.1, LETHE reduces ISM-A from 0.817 on clean images to 0.428 and QAF from 0.912 to 0.397, well below the second-best baselines (0.524 and 0.589, respectively). The qualitative results in Fig. 3 further show that LETHE degrades both the identity and the visual structure of the generated videos more completely, where it erases the subject on Wan2.1 and smears it into unrecognizable textures on CogVideoX.

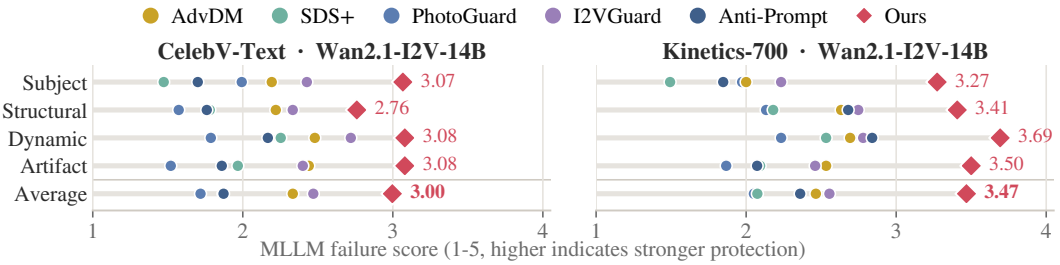


Figure 4: **MLLM-based evaluation** on Wan2.1. See full results in Tab. 3 (Appendix D.1).

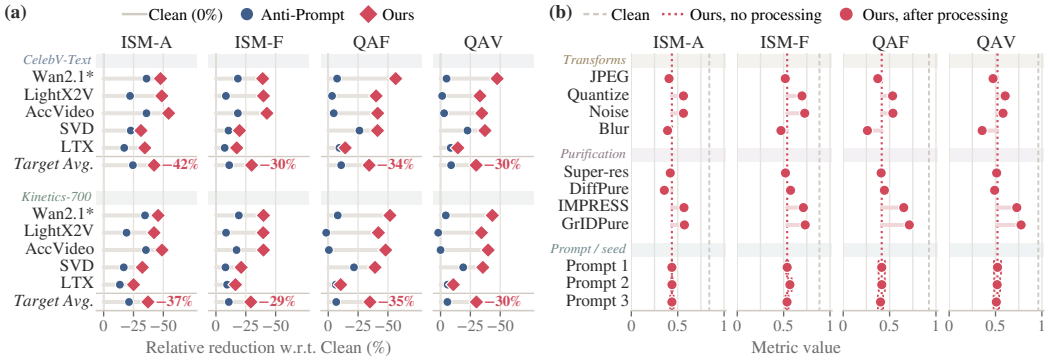


Figure 5: **Transferability and robustness results.** (a) Relative reduction of each metric w.r.t. clean images when the perturbations crafted on Wan2.1 are applied to four other I2V models. (b) Metric values on Wan2.1 (CelebV-Text) under different input transformations and generation conditions.

These results demonstrate that our value collapse and attention attraction strategies effectively block subsequent frames from reading the identity and appearance of the reference image, so that protected images can no longer be faithfully animated into high-quality videos by unauthorized users.

MLLM-based Evaluation. For a more comprehensive evaluation, we adopt the failure-oriented MLLM-based protocol of AntiPrompt (Song et al., 2026), where Qwen3-VL-8B-Instruct (Bai et al., 2025a) serves as the judge. The judge is given the generated video, the corresponding prompt, and the original input image, and scores four aspects: *Subject Preservation* (whether the subject remains the recognizable person the same as the one in the input image), *Structural Consistency* (whether shapes and proportions are stable), *Dynamic Consistency* (whether the motion is temporally coherent), and *Artifact Suppression* (whether rendering artifacts are present). Scores range from 1 to 5, and we invert them so that higher values indicate stronger protection (see Appendix C.3 for more details). As shown in Fig. 4, LETHE achieves the highest score over the four aspects on both datasets, and the results on CogVideoX in Tab. 3 (Appendix) show the same trend. This further confirms that LETHE provides stronger and more comprehensive protection against unauthorized I2V generation.

5.3 TRANSFERABILITY & ROBUSTNESS

Transferability. In practice, an attacker may not use exactly the same model as the surrogate. To account for this, we further evaluate whether the perturbations crafted on Wan2.1 transfer to models that are not directly involved in the optimization, and compare LETHE with AntiPrompt. Fig. 5(a) reports the relative reduction of each metric with respect to the clean images. On LightX2V and AccVideo, which are built on the same DiT architecture as Wan2.1, LETHE retains most of its effectiveness (e.g., lowering all metrics by $\sim 40\%$). On SVD and LTX-Video, which differ substantially from Wan2.1 in architecture, LETHE still yields clear degradation (e.g., $\sim 30\%$ ISM reduction for both models) and consistently outperforms AntiPrompt on all four metrics. We *hypothesize* that this is because tokens carrying reference information are also aggregated through softmax attention in all these models, so the reduced diversity of the reference representation induced by LETHE also generalizes beyond the surrogate. We leave a more thorough investigation to future work.

Robustness. Before being fed into the I2V model, a protected image may undergo image transformations (e.g., enhancement by platform), deliberate adversarial purification by the attacker, and

Table 2: **Ablation on the two objectives of LETHE** on Wan2.1. Each value is the mean of three runs, and standard deviations are reported in Appendix D.1.

Variant	ISM-A ↓	ISM-F ↓	QAF ↓	QAV ↓
CelebV-Text				
Clean	0.8173	0.8725	0.9123	0.9462
$\mathcal{L}_{\text{vc}}$ only	0.5899	0.6827	0.5224	0.6009
$\mathcal{L}_{\text{vc}} + \mathcal{L}_{\text{aa}}$ (LETHE)	**0.4280**	**0.5308**	**0.3970**	**0.4976**
Kinetics-700				
Clean	0.7584	0.8242	0.7974	0.8381
$\mathcal{L}_{\text{vc}}$ only	0.5084	0.6127	0.4600	0.5324
$\mathcal{L}_{\text{vc}} + \mathcal{L}_{\text{aa}}$ (LETHE)	**0.4125**	**0.4957**	**0.3852**	**0.4745**

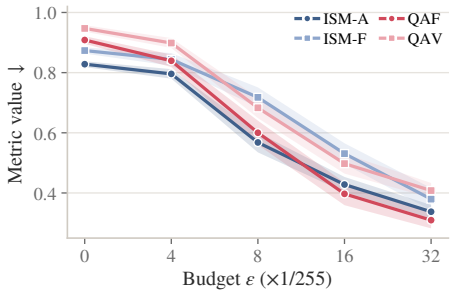


Figure 6: **Effect of the budget** ϵ on Wan2.1 and CelebV-Text. Shaded bands denote one standard deviation over three random seeds.

generation with prompts and seeds unknown to the defender. To test LETHE's robustness under these conditions, following Vu et al. (2026), we evaluate three groups of settings on Wan2.1: (I) 4 common input transformations, i.e., JPEG compression ($q = 70$), 3-bit color quantization, Gaussian noise ($\sigma = 0.2$), and Gaussian blur (kernel size 7); (II) 4 representative purifiers, i.e., Super-Res (Mustafa et al., 2019), DiffPure (Nie et al., 2022), IMPRESS (Cao et al., 2023), and GrIDPure (Zhao et al., 2024), each with its recommended configuration; and (III) 3 different prompts, each generated with 3 random seeds. As shown in Fig. 5(b), LETHE remains highly effective under these variations, and even the purifiers specifically designed to remove protective perturbations (IMPRESS and GrIDPure) recover only part of the gap to clean images, e.g., ISM-A rises from 0.43 to $\sim$ 0.56, which is still far below the clean ones ($\sim$ 0.80). Across prompts and seeds, the metrics also vary by no more than $\sim$ 0.03, suggesting that LETHE's effectiveness does not rely on specific generation conditions.

5.4 ABLATION STUDIES

Effect of each objective. We first ablate the contribution of the two objectives. As shown in Tab. 2, using $\mathcal{L}_{\text{vc}}$ alone already provides strong protection, e.g., reducing ISM-A from 0.81 to 0.43 and QAF from 0.912 to 0.522 on CelebV-Text, respectively. Adding $\mathcal{L}_{\text{aa}}$ further amplifies the protection by a notable margin, as reflected by consistently lower values on all four metrics across both datasets, together with smaller standard deviations across runs (Appendix D.1). This is consistent with our analysis in Sec. 4 that $\mathcal{L}_{\text{vc}}$ primarily collapses the convex hull, whereas $\mathcal{L}_{\text{aa}}$ keeps the attention on these tokens, which amplifies the collapse and makes the protection both stronger and more stable.

Effect of the budget. Fig. 6 shows how each metric varies with the perturbation budget ϵ. Consistent with the common observation in this field (Liang et al., 2023; Long et al., 2026), a larger budget yields stronger protection, and all four metrics decrease monotonically with ϵ. Notably, with only half of the perturbation budget ($\epsilon = 8/255$), LETHE already achieves protection comparable to several strong baselines at $\epsilon = 16/255$, e.g., an ISM-A of 0.58 and a QAV of 0.69. This suggests that LETHE spends the perturbation budget more efficiently, as both of its objectives directly target the pathway through which subsequent frames read the reference image.

6 CONCLUSION

This paper studies how to protect personal images against unauthorized I2V generation. We first show that subsequent frames read visual content from the reference frame primarily with attention, and demonstrate through a convex hull-based view that the diversity of these readouts is bounded by the spread of the reference values. Our intervention experiment further confirms that contracting this spread effectively disrupts generation. Building on this analysis, we propose LETHE, which consists of a value collapse loss that suppresses the diversity of the reference values and an attention attraction loss that draws the attention of subsequent frames to the collapsed reference tokens, so that the protected images can no longer be faithfully animated. Extensive experiments on multiple I2V models and two re-annotated human-centric datasets demonstrate LETHE's strong performance in various settings. We hope LETHE, together with our analyses and the dataset can deepen the community's understanding and inspire more effective image protection methods in the future.

AI USE STATEMENT

In this work, we used generative AI tools to annotate our evaluation datasets: for each conditioning image, Qwen2.5-VL-72B-Instruct (Bai et al., 2025b) generated several candidate I2V prompts, from which the authors manually selected and verified the final prompts (Appendix C.1). Additionally, we used Claude Opus 4.8, 5.5 (Anthropic, 2026b) and Claude Fable 5.1 (Anthropic, 2026a) to polish the grammar and word choice of author-written text. Separately, following Song et al. (2026), Qwen3-VL-8B-Instruct (Bai et al., 2025a) serves as an evaluation judge, as described in Appendix C.3. The authors have carefully reviewed all AI-assisted content and take full responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work is motivated by the growing concern that personal images can be used without authorization to generate videos of the depicted person through image-to-video models. By analyzing how these models read visual content from the reference image, we hope to provide insights into what governs the effectiveness of protective perturbations in this setting, and we believe LETHE represents a valuable step toward more effective protection of personal images. All models used in this work are publicly released and used in accordance with their licenses. Our evaluation sets are built upon CelebV-Text and Kinetics-700, which we use for non-commercial research in accordance with their terms of use. Our re-annotation only adds textual descriptions and does not collect any new personal data. To comply with the original terms, we will release only the textual annotations together with the video identifiers and frame indices, so that users obtain the original data from the official sources. The qualitative examples in this paper are limited to those necessary to illustrate the protection effect. We have read and ensured that our work adheres to the ICLR Code of Ethics.

REFERENCES

Anthropic. System card: Claude fable 5.1 & claude mythos 5.1. https://www-cdn.anthropic.com/0339e6a7c5c7b87f5c07798616dc32c215d14235/Claude%20Fable%205.1%20&%20Claude%20Mythos%205.1%20System%20Card.pdf, September 2026a. Accessed: 2026-09-24.

Anthropic. System card: Claude opus 5.5. https://www-cdn.anthropic.com/fc1b44717c85dc068bc6ba5024219938094694bd/Claude%20Opus%205.5%20System%20Card.pdf, September 2026b. Accessed: 2026-09-24.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025b. URL https://arxiv.org/abs/2502.13923.

Keith Ball et al. An elementary introduction to modern convex geometry. *Flavors of geometry*, 31 (1-58):26, 1997.

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.

Bochuan Cao, Changjiang Li, Ting Wang, Jinyuan Jia, Bo Li, and Jinghui Chen. Impress: Evaluating the resilience of imperceptible perturbations against unauthorized data usage in diffusion-based generative ai. *Advances in Neural Information Processing Systems*, 36:10657–10677, 2023.

Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. A short note on the kinetics-700 human action dataset. *arXiv preprint arXiv:1907.06987*, 2019.

Huanran Chen, Yichi Zhang, Yinpeng Dong, Xiao Yang, Hang Su, and Jun Zhu. Rethinking model ensemble in transfer-based adversarial attacks. In *International Conference on Learning Representations*, volume 2024, pp. 19199–19225, 2024.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4690–4699, 2019.

Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9185–9193. IEEE, 2018.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64 (12):86–92, 2021.

Jindong Gu, Xiaojun Jia, Pau de Jorge, Wenqain Yu, Xinwei Liu, Avery Ma, Yuan Xun, Anjun Hu, Ashkan Khakzar, Zhijiang Li, et al. A survey on transferability of adversarial examples across deep neural networks. *arXiv preprint arXiv:2310.17626*, 2023.

Dongnan Gui, Xun Guo, Wengang Zhou, and Yan Lu. I2vguard: Safeguarding images against misuse in diffusion-based image-to-video models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12595–12604. IEEE, 2025.

Jia Guo, Jiankang Deng, Alexandros Lattas, and Stefanos Zafeiriou. Sample and computation redistribution for efficient face detection. *arXiv preprint arXiv:2105.04714*, 2021.

Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, et al. Ltx-video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024.

Ziming Hong, Tianyu Huang, Runnan Chen, Shanshan Ye, Mingming Gong, Bo Han, and Tongliang Liu. Adlift: Lifting adversarial perturbations to safeguard 3d gaussian splatting assets against instruction-driven editing. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=RL565ePGHR.

Yuxin Huang, Ziming Hong, Mingming Gong, Wanyu Wang, Jing Zhang, and Tongliang Liu. Delving into the temporal challenges of unified video protection against image-to-video and fine-tuning-based customization. *arXiv preprint arXiv:2607.13336*, 2026.

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21807–21818. IEEE, 2024.

Jun Jia, Hongyi Miao, Yingjie Zhou, Wangqiu Zhou, Jianbo Zhang, Linhan Cao, Dandan Zhu, Hua Yang, Xiongkuo Min, Wei Sun, et al. Adapter shield: A unified framework with built-in authentication for preventing unauthorized zero-shot image-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 30120–30129, 2026.

Max Kamachee, Stephen Casper, Michelle L Ding, Rui-Jie Yew, Anka Reuel, Stella Biderman, and Dylan Hadfield-Menell. Video deepfake abuse: How company choices predictably shape misuse patterns. *arXiv preprint arXiv:2512.11815*, 2025.

Hyunjik Kim, George Papamakarios, and Andriy Mnih. The lipschitz constant of self-attention. In *International Conference on Machine Learning*, pp. 5562–5571. PMLR, 2021.

Yunji Kim, Jiyoung Lee, Jin-Hwa Kim, Jung-Woo Ha, and Jun-Yan Zhu. Dense text-to-image generation with attention modulation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7667–7677. IEEE, 2023.

Yanjie Li, Wenxuan Zhang, Xinqi Lyu, Yihao Liu, and Bin Xiao. Styleguard: Preventing text-to-image-model-based style mimicry attacks by style perturbations. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

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. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 20763–20786. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/liang23g.html.

LightX2V Team. LightX2V: AI video generation and multimodal content production platform. https://x2v.light-ai.top/, 2026. Accessed: 2026-09-22.

Kunhao Liu, Wenbo Hu, Jiale Xu, Ying Shan, and Shijian Lu. Rolling forcing: Autoregressive long video diffusion in real time. In *International Conference on Learning Representations*, volume 2026, pp. 91177–91196, 2026.

Yixin Liu, Chenrui Fan, Yutong Dai, Xun Chen, Pan Zhou, and Lichao Sun. Metacloak: Preventing unauthorized subject-driven text-to-image diffusion-based synthesis via meta-learning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24219–24228. IEEE, 2024.

Zeqian Long, Ozgur Kara, Haotian Xue, Yongxin Chen, and James M Rehg. Immune2v: Image immunization against dual-stream image-to-video generation. *arXiv preprint arXiv:2604.10837*, 2026.

Aamir Mustafa, Salman H Khan, Munawar Hayat, Jianbing Shen, and Ling Shao. Image super-resolution as a defense against adversarial attacks. *IEEE Transactions on Image Processing*, 29: 1711–1724, 2019.

Kamal Nasrollahi and Thomas B Moeslund. Face quality assessment system in video sequences. In *European Workshop on Biometrics and Identity Management*, pp. 10–18. Springer, 2008.

Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Animashree Anandkumar. Diffusion models for adversarial purification. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 16805–16827. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/nie22a.html.

Yan Pang, Baicheng Chen, Yang Zhang, and Tianhao Wang. Vgmshield: Mitigating misuse of video generative models. *arXiv preprint arXiv:2402.13126*, 2024.

Hadi Salman, Alaa Khaddaj, Guillaume Leclerc, Andrew Ilyas, and Aleksander Madry. Raising the cost of malicious ai-powered image editing. *arXiv preprint arXiv:2302.06588*, 2023.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 815–823, 2015.

Shitong Shao, Lichen Bai, Pengfei Wan, James Kwok, and Zeke Xie. Efficient video diffusion models: Advancements and challenges. *arXiv preprint arXiv:2604.15911*, 2026.

Yeonghwan Song, Chanhui Lee, Jinsoo Park, and Jeany Son. Anti-prompt: Image protection against text-guided image-to-video generation. In *European Conference on Computer Vision (ECCV)*, 2026.

Zekun Sun, Zijian Liu, Shouling Ji, Chenhao Lin, and Na Ruan. Pretender: Universal active defense against diffusion finetuning attacks. In *34th USENIX Security Symposium (USENIX Security 25)*, pp. 1017–1036, 2025.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Thanh Van Le, Hao Phung, Thuan Hoang Nguyen, Quan Dao, Ngoc N Tran, and Anh Tran. Anti-dreambooth: Protecting users from personalized text-to-image synthesis. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2116–2127. IEEE, 2023.

Duc Vu, Anh Nguyen, Chi Tran, and Anh Tran. Anti-i2v: Safeguarding your photos from malicious image-to-video generation. *arXiv preprint arXiv:2603.24570*, 2026.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Xianlong Wang, Wenbo Pan, Shijia Zhou, Ke Li, Yuqi Wang, Zeyu Ye, Hangtao Zhang, Leo Yu Zhang, and Xiaohua Jia. Image-to-video diffusion: From foundations to open frontiers. *arXiv preprint arXiv:2605.17248*, 2026.

Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, Qiong Yan, Xiongkuo Min, Guangtao Zhai, and Weisi Lin. Q-align: Teaching LMMs for visual scoring via discrete text-defined levels. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=PHjkVjR78A.

Haotian Xue, Chumeng Liang, Xiaoyu Wu, and Yongxin Chen. Toward effective protection against diffusion-based mimicry through score distillation. In *International Conference on Learning Representations*, volume 2024, pp. 25734–25758, 2024.

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. In *International Conference on Learning Representations*, 2025.

Jung Yi, Wooseok Jang, Paul Hyunbin Cho, Jisu Nam, Heeji Yoon, and Seungryong Kim. Deep forcing: Training-free long video generation with deep sink and participative compression. *arXiv preprint arXiv:2512.05081*, 2025.

Jianhui Yu, Hao Zhu, Liming Jiang, Chen Change Loy, Weidong Cai, and Wayne Wu. Celebv-text: A large-scale facial text-video dataset. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14805–14814. IEEE, 2023.

Haiyu Zhang, Xinyuan Chen, Yaohui Wang, Xihui Liu, Yunhong Wang, and Yu Qiao. Accvideo: Accelerating video diffusion model with synthetic dataset. *arXiv preprint arXiv:2503.19462*, 2025.

Jiaming Zhang, Boyang Chen, Zherui Li, Fuyao Zhang, Xinyu Yan, Hong Xi Tae, Wenwen He, Xuan Wang, Siqi Guo, Junhao Dong, et al. Adversarial attacks for good: A survey of proactive protection across the visual content lifecycle. *arXiv preprint arXiv:2608.04314*, 2026.

Yechao Zhang, Shengshan Hu, Leo Yu Zhang, Junyu Shi, Minghui Li, Xiaogeng Liu, Wei Wan, and Hai Jin. Why does little robustness help? a further step towards understanding adversarial transferability. In *2024 IEEE Symposium on Security and Privacy (SP)*, pp. 3365–3384. IEEE, 2024.

Longjie Zhao, Ziming Hong, Jiaxin Huang, Runnan Chen, Mingming Gong, and Tongliang Liu. Intellectual property protection for 3d gaussian splatting assets: A survey. *arXiv preprint arXiv:2602.03878*, 2026.

Zhengyue Zhao, Jinhao Duan, Kaidi Xu, Chenan Wang, Rui Zhang, Zidong Du, Qi Guo, and Xing Hu. Can protective perturbation safeguard personal data from being exploited by stable diffusion? In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24398–24407. IEEE, 2024.

Jiachen Zhou, Mingsi Wang, Tianlin Li, Guozhu Meng, and Kai Chen. Dormant: Defending against pose-driven human image animation. In *34th USENIX Security Symposium (USENIX Security 25)*, pp. 5209–5228, 2025.

A Proof of Proposition 1

Let $A = (\alpha_{mi}) \in \mathbb{R}^{M\times N}$ collect the attention weights of the M subsequent tokens over the reference tokens, where each row α_m lies in Δ^{N-1} (Eq. (1)). We denote the LHS of Eq. (3) by

$$D(A,V) = \frac{1}{M}\sum_{m=1}^{M}\|o_m - \bar{o}\|^2, \qquad o_m = \sum_{i=1}^{N}\alpha_{mi}v_i, \qquad \bar{o} = \frac{1}{M}\sum_{m=1}^{M} o_m, \tag{7}$$

and refer to it as the readout diversity throughout this appendix for better readability.

Proof. Let $c^\star$ and $R = R(V)$ denote the center and the radius of the minimum enclosing ball of V, which exist and are unique since $c \mapsto \max_i \|v_i - c\|^2$ is continuous, coercive, and strictly convex. For any $c \in \mathbb{R}^d$, expanding the square around $\bar{o}$ and noting that $\sum_m (o_m - \bar{o}) = 0$ gives

$$\frac{1}{M}\sum_{m=1}^{M}\|o_m - c\|^2 = D(A,V) + \|\bar{o} - c\|^2. \tag{8}$$

Upper bound. Every v_i lies in the ball $B(c^\star, R)$. Since the ball is convex, it contains the convex hull $\mathcal{H}(V)$, and hence every readout $o_m \in \mathcal{H}(V)$ satisfies $\|o_m - c^\star\| \le R$. Taking $c = c^\star$ in Eq. (8) and dropping the nonnegative term $\|\bar{o} - c^\star\|^2$, we obtain

$$D(A,V) \le \frac{1}{M}\sum_{m=1}^{M}\|o_m - c^\star\|^2 \le R(V)^2. \tag{9}$$

Eq. 9 only uses $\alpha_m \in \Delta^{N-1}$, so the bound holds for every M and every attention configuration.

Sharpness Analysis. We show that the supremum of $D(A,V)$ over M and the attention weights equals $R(V)^2$. Assume $R > 0$ and let $\mathcal{S} = \{i : \|v_i - c^\star\| = R\}$ be the support points. A standard property of the minimum enclosing ball is that $c^\star \in \mathrm{conv}\{v_i\}_{i\in\mathcal{S}}$, since otherwise $c^\star$ could be moved toward all support points to reduce the radius, contradicting minimality. Hence, there exist weights $\lambda_i \ge 0$ on $\mathcal{S}$ with $\sum_i \lambda_i = 1$ and $\sum_i \lambda_i v_i = c^\star$. Let each subsequent token place its entire attention on a single support point, with the fraction of tokens assigned to point i approaching λ_i as $M \to \infty$. Then every readout lies on the boundary of the ball, and $\bar{o} \to c^\star$. By Eq. (8), $D(A,V) = R^2 - \|\bar{o} - c^\star\|^2 \to R^2$. Since softmax weights can approximate one-hot weights arbitrarily well, the same supremum holds for softmax attention. □

Remarks. Proposition 1 shows that the readout diversity cannot exceed $R(V)^2$, and that this constant cannot be improved in general. One might note that the equality is attained only under specific configurations, so that the readout diversity might possibly lie well below $R(V)^2$ in reality. However, as we empirically observed in Appendix D.1, reducing the value spread through our objective is accompanied by a substantial reduction in the readout diversity across nearly all DiT blocks. This suggests that the configurations under which the bound becomes loose are not observed in practice, and that $R(V)$ provides an informative guide to what subsequent frames can read.

B Relating the Value Collapse Objective to Readout Diversity

Fix a layer, a head, and a denoising step, and let $\bar{v} = \frac{1}{N}\sum_i v_i$, $\sigma_V^2 = \frac{1}{N}\sum_i \|v_i - \bar{v}\|^2$, and $E_V = \frac{1}{N}\sum_i \|v_i\|^2 = \sigma_V^2 + \|\bar{v}\|^2$. For $E_V > 0$, each summand of Eq. (4) equals

$$C(V) = \frac{\sigma_V^2}{E_V} = \frac{\sigma_V^2}{\sigma_V^2 + \|\bar{v}\|^2}, \tag{10}$$

i.e., the fraction of the value energy not shared by all reference tokens.

Boundedness of the value energy. In both Wan2.1 and CogVideoX, the hidden states entering W_V are produced by a LayerNorm followed by an affine modulation computed from the timestep embedding, which does not depend on the input frame. Hence $\|v_i\| \le C_t$ for a constant C_t determined by the model weights and t, and $\|\bar{v}\|^2 \le E_V \le C_t^2$. It follows that $C(V) \ge \sigma_V^2/(\sigma_V^2 + C_t^2)$, so

that driving the objective toward zero requires reducing the token-specific variation σ_V^2, rather than enlarging the common component $\bar{v}$.

Relation to the readout diversity. Let $p_i = \frac{1}{M}\sum_m \alpha_{mi}$ denote the average weight received by reference token i, and $\kappa(A) = N \max_i p_i$, which equals one when the average weights are uniform.

Proposition 2. *For any attention matrix A with rows in Δ^{N-1} and $E_V > 0$,*

$$D(A,V) \le \sum_{i=1}^{N} p_i \|v_i - \bar{v}\|^2 \le \kappa(A)\,\sigma_V^2, \qquad \textit{and} \qquad \frac{D(A,V)}{E_V} \le \min\Big\{\frac{R(V)^2}{E_V},\ \kappa(A)\,C(V)\Big\}. \tag{11}$$

Proof. For each m, we have $\sum_i \alpha_{mi}(v_i - o_m) = 0$. Expanding $\|v_i - \bar{o}\|^2$ around o_m, weighting by α_{mi}, and averaging over m gives

$$\sum_{i=1}^{N} p_i \|v_i - \bar{o}\|^2 = D(A,V) + \frac{1}{M}\sum_{m=1}^{M}\sum_{i=1}^{N} \alpha_{mi}\|v_i - o_m\|^2. \tag{12}$$

Dropping the last term yields $D(A,V) \le \sum_i p_i \|v_i - \bar{o}\|^2$. Since $\bar{o} = \sum_i p_i v_i$ minimizes $c \mapsto \sum_i p_i \|v_i - c\|^2$, replacing $\bar{o}$ by $\bar{v}$ gives the first inequality, and bounding each p_i by $\max_i p_i$ gives the second. Dividing by E_V and combining with Proposition 1 completes the proof. □

Relation to the minimum enclosing ball. Since $\bar{v}$ minimizes the sum of squared distances and is also an admissible center for $R(V)$, we have

$$\sigma_V^2 \le \frac{1}{N}\sum_i \|v_i - c^\star\|^2 \le R(V)^2 \le \max_i \|v_i - \bar{v}\|^2 \le N\sigma_V^2, \quad \text{i.e.,} \quad \frac{R(V)^2}{N\,E_V} \le C(V) \le \frac{R(V)^2}{E_V}. \tag{13}$$

Hence, $C(V)$ and the normalized squared radius vanish together, and combined with the energy bound, $R(V)^2 \le NC_t^2\,C(V)$. The factor N arises from bounding the maximum deviation by the sum, which allows the gradient of the objective to involve all reference tokens rather than only the most distant ones. One might note that this factor could make the two quantities weakly coupled in principle. However, as analyzed in Appendix D.1, minimizing $\mathcal{L}_{\text{vc}}$ reduces the normalized $R(V)^2$ by a margin comparable to that of the objective itself across blocks, suggesting that the two quantities are closely coupled in practice.

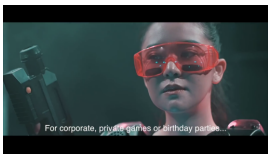
Relation to the value intervention. For the contraction $v_i^{(\gamma)} = c + \gamma(v_i - c)$ used in Sec. 3, linearity of the readout gives $o_m^{(\gamma)} = c + \gamma(o_m - c)$, and hence $D(A, V^{(\gamma)}) = \gamma^2 D(A,V)$, $R(V^{(\gamma)}) = \gamma R(V)$, and $\sigma^2_{V^{(\gamma)}} = \gamma^2\sigma_V^2$ for any center c and any attention weights at the intervened operation. The intervention thus contracts the readout diversity by the same factor as the value spread, which is consistent with the substantial degradation observed in Sec. 3.

Implications for the design. Taken together, these results motivate and guide the design of LETHE. The value collapse objective reduces $C(V)$, which bounds the normalized readout diversity (Proposition 2) and drives the normalized radius toward zero (Eq. (13)). The analysis above concerns the component read from the reference tokens. In the full attention, this component is combined with that read from the other tokens, and the attention attraction objective encourages a larger share of the former, so that the collapsed component accounts for a larger part of what subsequent frames read. This is consistent with the ablation in Sec. 5.4, where the two objectives are complementary.

C IMPLEMENTATION DETAILS

C.1 BENCHMARK CONSTRUCTION DETAILS

Why New Benchmarks? As discussed in Sec. 5, the primary targets of I2V misuse are personal photos. Therefore, evaluating protection against this threat requires reference images that depict real persons, paired with prompts that drive the depicted person to perform plausible actions. However, existing datasets do not fully meet this need. General-purpose I2V benchmarks, such as VBench (Huang et al., 2024), are designed to evaluate generation quality across diverse content



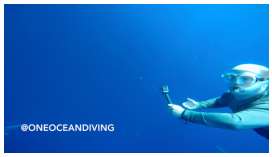
(a) Glasses



(b) Headgear or similar items



(c) False detection



(d) Near-profile or back view



Description: A gray-haired man holds a smoking pipe in a workshop filled with tools and papers.

Primary prompt: The man gently exhales smoke from his pipe, keeping his face towards the camera.

Auxiliary prompt 1: The gray-haired man slightly adjusts his grip on the pipe, maintaining eye contact.

Auxiliary prompt 2: The man raises his left eyebrow subtly, while holding the pipe steady in hand.

Figure 7: **Examples of candidate frames. Left:** representative frames rejected during manual review, where the face is (a) occluded by glasses, (b) covered by headgear or similar items, (c) a false detection, or (d) seen from a near-profile or back view. **Right:** an accepted frame with its final annotation, consisting of a scene description, a primary prompt, and two auxiliary prompts.

and not human-centric. Human-centric video datasets, on the other hand, are not designed for I2V generation. For instance, the texts in CelebV-Text (Yu et al., 2023) are generated from templates to describe the attributes presented in the entire source clip, such as appearance, rather than to drive the person in a given frame to perform plausible motions. Similarly, Kinetics-700 (Carreira et al., 2019) only provides action class labels. Consequently, existing studies on I2V protection are evaluated on disparate data, such as 355 images from VBench-I2V (Song et al., 2026) or 50 first frames from DAVIS (Long et al., 2026). The benchmarks most closely related to ours are constructed by Vu et al. (2026) based on CelebV-Text and UCF101, yet to the best of our knowledge, they are not publicly available. To fill this gap, we construct two human-centric benchmarks by re-annotating CelebV-Text (Yu et al., 2023) and Kinetics-700 (Carreira et al., 2019).

Source Video Selection. For CelebV-Text, we randomly sample 1,000 clips from the dataset, download them, and trim them according to the official timestamps. The trimming only cuts the temporal range without re-encoding, which avoids introducing additional compression artifacts. For Kinetics-700, we randomly select 400 of the 700 action classes to cover diverse human actions. Since Kinetics videos often contain multiple persons or focus on the scene, we first apply YOLOv8 for person detection and retain only videos with a single person. We then rank the remaining videos by video resolution and keep at most the top 10 videos per class.

Reference Frame Selection. We uniformly sample frames from each video according to its duration, i.e., 10 frames per video for CelebV-Text and 8 frames per video for Kinetics-700. Each frame is processed with the `buffalo_l` model of InsightFace (Deng et al., 2019), where the SCRFD (Guo et al., 2021) detection module provides face bounding boxes and detection confidences, and the 68-point 3D landmark module provides head pose angles (yaw and pitch). A frame becomes a candidate only if it satisfies all of the following four conditions with empirically chosen thresholds: (i) the frame is in landscape orientation with an aspect ratio close to 16:9, which is consistent with the default output resolution of mainstream I2V models and avoids additional cropping, (ii) the face occupies at least 1% of the frame area, (iii) the face is nearly frontal, with both $|\text{yaw}|$ and $|\text{pitch}|$ no larger than 20°, and (iv) the face region is sufficiently sharp, with a Laplacian variance computed by OpenCV of at least 50. A video is retained if at least one of its frames satisfies all conditions, and is discarded otherwise. After this step, 388 videos remain for CelebV-Text and 655 videos remain for Kinetics-700. Similar to Nasrollahi & Moeslund (2008), we then score each candidate frame by a weighted combination of four criteria, namely detection confidence, face area, sharpness, and frontalness. We keep the highest-scoring frame of each video and save it as a PNG image at its original resolution. Finally, to ensure that each identity appears only once and that the samples are independent, we further remove duplicate identities based on InsightFace identity embeddings, keeping only one frame when the similarity between two frames exceeds 0.8. In practice, no pair exceeded this threshold, so no frames were discarded.

Motion Prompt Annotation. The original annotations of these datasets are not directly applicable as I2V prompts, since they either describe the content of the source clip or consist of class labels only, rather than instructing the person in the selected frame to perform a plausible action. We thus

annotate motion prompts for each reference frame in two stages. First, we use Qwen2.5-VL-72B-Instruct (Bai et al., 2025b) to generate ten diverse candidate action descriptions for each frame. The instruction requires each description to contain about 15 words, to be directly usable as a generation prompt, to be consistent with the depicted scene, and to avoid motions so large that the person or the face leaves the frame. Second, two of the authors review all images and annotations with a web annotation interface, and every decision is made jointly by both reviewers. The reviewers first decide whether to retain each image. An image is discarded if the person is incomplete (e.g., cropped), large occlusion appears in the frame, the face is occluded by sunglasses, headgear, or similar items, the person is close to a profile or back view, the face is a false detection (e.g., a face printed on a US dollar), or none of the ten candidates suits the depicted scene (e.g., asking a person sitting indoors to drive a car). For each retained image, the reviewers select the best candidate as the primary prompt, together with two auxiliary prompts that are distinct from each other and from the primary prompt while remaining plausible in the scene. Candidates that conflict with the pose or environment, or that involve objects absent from the frame (e.g., glasses, microphones, or pets), are discarded in this step. Additionally, we prompt the model to generate an appearance description, which contains about 20 words and only describes the visible characteristics of the person and the surrounding environment, without involving any action. Finally, the reviewers correct inaccurate content in the appearance description and remove unnecessary background details added by the model.

After manual review, we keep the 350 highest-scoring images for each benchmark, which matches the scale of existing evaluations (Song et al., 2026). The ranking criterion is the same criterion used in reference frame selection as described earlier. Each image is paired with one primary prompt, two auxiliary prompts, and one appearance description. The primary prompts are used in the main experiments, the auxiliary prompts are used to evaluate robustness to prompt variation (Fig. 5), and the appearance descriptions are provided as supplementary annotations.

License and Release. Neither source dataset hosts videos directly. Instead, both distribute YouTube video IDs with timestamps, and the copyright of the videos belongs to their original uploaders. According to its terms of use, CelebV-Text is available for non-commercial research purposes only. The Kinetics annotations are released under the CC BY 4.0 license, which permits use and redistribution with attribution. We strictly comply with these terms. Our use of both datasets is limited to non-commercial academic research, and we do not redistribute any original videos, video clips, or extracted reference frames. Instead, we only release our own annotations, including the source video IDs and clip identifiers, the frame indices of the reference frames, the primary and auxiliary prompts (all provided via indices), and the appearance descriptions. We also provide scripts to reproduce the reference frames, so that users can recover them after obtaining the source videos under the terms of the original datasets. The annotations built upon CelebV-Text and Kinetics-700 will be released under the CC BY-NC 4.0 and CC BY 4.0 licenses, respectively, with proper attribution to the source datasets. We refrain from releasing any generated videos, and limit the qualitative examples in this paper to a small number of frames for illustration. Upon acceptance, we will release the annotations, scripts, and a dataset card following Gebru et al. (2021) in a public repository, together with a contact for handling removal requests from individuals appearing in the data. We hope these two standardized, publicly available benchmarks can further promote research in this immature area.

C.2 More Implementation Details

All experiments are conducted on NVIDIA H100 NVL (94GB) GPUs, where the models run in `bf16` precision and the losses are computed in `fp32` precision. We keep the surrogate models' parameters frozen and optimizing only δ. Both models use their default number of frames and default samplers with 25 sampling steps. At each iteration, the layers in $\mathcal{B}$ and the timesteps along the sampling trajectory are randomly sampled to estimate the expectations in the losses. We solve the overall objective with MI-FGSM (Dong et al., 2018) under a budget of $\epsilon = 16/255$ for 100 iterations, using its default step size and momentum. To improve robustness against common image processing, we adopt Expectation over Transformation (EOT) with transformation settings following Sun et al. (2025). Protecting one image takes about 7–8 minutes on both models, and the pseudocode is provided in Alg. 1. For a fair comparison, all baselines are optimized with the same surrogate model, same perturbation budget, input resolution, and prompts as LETHE, as well as identical generation and evaluation configurations. QAF and QAV are computed by Q-Align in image and video modes, respectively, and the raw scores are linearly mapped from $[1, 5]$ to $[0, 1]$ for better readability. Unless otherwise stated, all results are averaged over all videos with 3 random seeds.

Algorithm 1 Crafting the protective perturbation with LETHE

Input: reference image x_{ref}, surrogate model f, budget ϵ, weight α, iterations T_{iter}, step size η, momentum μ, transformation set $\mathcal{T}$
Output: protected image $x' = x_{\text{ref}} + \delta$

$\delta \sim \mathcal{U}(-\epsilon/4, \epsilon/4)$, $g \leftarrow 0$
for $k = 1, \ldots, T_{\text{iter}}$ **do**
  Sample layers $\mathcal{B}_k$, denoising step t, noise z_T, and transformation $\tau \sim \mathcal{T}$
  $\tilde{x} \leftarrow \tau(x_{\text{ref}} + \delta)$
  Run one forward pass of f at step t with $\tilde{x}$, z_T, and hooks on $\mathcal{B}_k$
  Collect reference values $\{v_i\}$, reference keys $\{k_i\}$, and subsequent-frame queries $\{q_m\}$ for each $(\ell, h) \in \mathcal{B}_k$
  $\mathcal{L} \leftarrow \mathcal{L}_{\text{vc}} + \alpha\, \mathcal{L}_{\text{aa}}$ ▷ Eq. (4), Eq. (5); $\bar{q}$ detached
  $g \leftarrow \mu\, g + \nabla_\delta \mathcal{L} \,/\, \|\nabla_\delta \mathcal{L}\|_1$
  $\delta \leftarrow \Pi_{\|\cdot\|_\infty \le \epsilon}\big(\delta - \eta\, \text{sign}(g)\big)$
  $\delta \leftarrow \text{clip}(x_{\text{ref}} + \delta, 0, 1) - x_{\text{ref}}$
end for
return $x_{\text{ref}} + \delta$

Following Liu et al. (2024), we further confirm that the improvements of LETHE over all baselines are statistically significant under the Wilcoxon signed-rank test ($p \le 0.01$). We will release our code and detailed configurations upon acceptance.

C.3 DETAILS ON MLLM-BASED EVALUATION PROTOCOL

We adopt the failure-oriented evaluation protocol of AntiPrompt (Song et al., 2026), including its judge model (Qwen3-VL-8B-Instruct (Bai et al., 2025a)), frame sampling, prompt template, and output format, and refer readers to the original paper for details. The protocol scores four interpretable dimensions that capture common I2V failures, and the judge's reliability has been calibrated and validated through a human study and consistency analyses in the original paper. To make the scores more faithfully reflect the protection effectiveness in our setting, we make the following minor adaptations, and all methods are evaluated under the same adapted protocol. First, the original protocol uses the first generated frame as the reference for Subject Preservation. However, in Wan2.1 and CogVideoX, the first frame is also generated by the model and can thus be affected by the input perturbation, making the judgement occasionally less faithful. Thus, we use the original input image as the reference instead. Second, in the original protocol, each dimension is scored from 1 to 5, where lower scores indicate more severe failures and thus stronger protection. For consistency with other metrics, we report six minus the original score, so that higher scores indicate stronger protection. Notably, LETHE achieves the best average score under both the original and the adapted protocols.

D ADDITIONAL RESULTS & DISCUSSIONS

D.1 ADDITIONAL RESULTS

Additional Results. Tab. 4 gives the mean and standard deviation over three runs for the ablation in Tab. 2. Besides, Tab. 3 gives the full numerical results for the Fig. 4. Finally, we provide more qualitative examples of LETHE and baselines in Fig. 10 and 11.

Empirical Validation of LETHE's Mechanism. To examine how LETHE operates in practice, we measure, on Wan2.1, the value collapse loss $\mathcal{L}_{\text{vc}}$, the normalized $R(V)^2$, and the normalized readout diversity of subsequent frames at each DiT block, as well as the attention mass that subsequent frames allocate to the reference tokens, for both clean and protected images, with all quantities averaged over denoising steps (Fig. 8). As shown in Fig. 8(a)–(c), the normalized $R(V)^2$ decreases by a margin comparable to that of the loss across blocks, and the two follow nearly identical trends. The readout diversity of subsequent frames decreases even more substantially, by over 50% in nearly all blocks. This suggests that the perturbation reduces the diversity of the reference representation from the input onward, and that this effect propagates through the network. Since the readout diversity scales with the square of the contraction factor in the intervention of Sec. 3, the observed reduction

Table 3: **MLLM-based evaluation.** Judge scores on Wan2.1-I2V-14B and CogVideoX-5B-I2V (1–5, higher means stronger protection). Darker cells indicate higher scores, scaled per column across both models. The best protection in each column of each block is in bold.

		CelebV-Text					Kinetics-700				
Model	Method	Subj.	Struct.	Dyn.	Artif.	*Avg.*	Subj.	Struct.	Dyn.	Artif.	*Avg.*
Wan2.1-I2V	AdvDM	2.19	2.22	2.48	2.44	2.33	2.00	2.63	2.69	2.53	2.46
	SDS+	1.47	1.78	2.25	1.97	1.87	1.49	2.18	2.53	2.09	2.07
	PhotoGuard	1.99	1.57	1.79	1.52	1.72	1.97	2.13	2.23	1.87	2.05
	I2VGuard	2.43	2.33	2.72	2.40	2.47	2.23	2.75	2.78	2.46	2.55
	AntiPrompt	1.70	1.76	2.17	1.86	1.87	1.85	2.68	2.84	2.07	2.36
	Ours	**3.07**	**2.76**	**3.08**	**3.08**	**3.00**	**3.27**	**3.41**	**3.69**	**3.50**	**3.47**
CogVideoX	AdvDM	1.60	1.90	2.89	2.08	2.12	2.05	2.76	3.28	2.07	2.54
	SDS+	1.89	2.03	2.69	2.17	2.20	1.75	2.51	2.92	2.26	2.36
	PhotoGuard	1.65	1.74	2.54	1.75	1.92	1.71	2.37	2.81	2.06	2.24
	I2VGuard	1.98	2.07	2.62	2.11	2.19	1.99	2.70	3.35	2.40	2.61
	AntiPrompt	2.07	2.12	2.77	2.01	2.24	2.30	2.69	3.31	2.12	2.61
	Ours	**2.31**	**2.55**	**3.29**	**2.67**	**2.71**	**2.54**	**2.93**	**3.54**	**2.78**	**2.95**

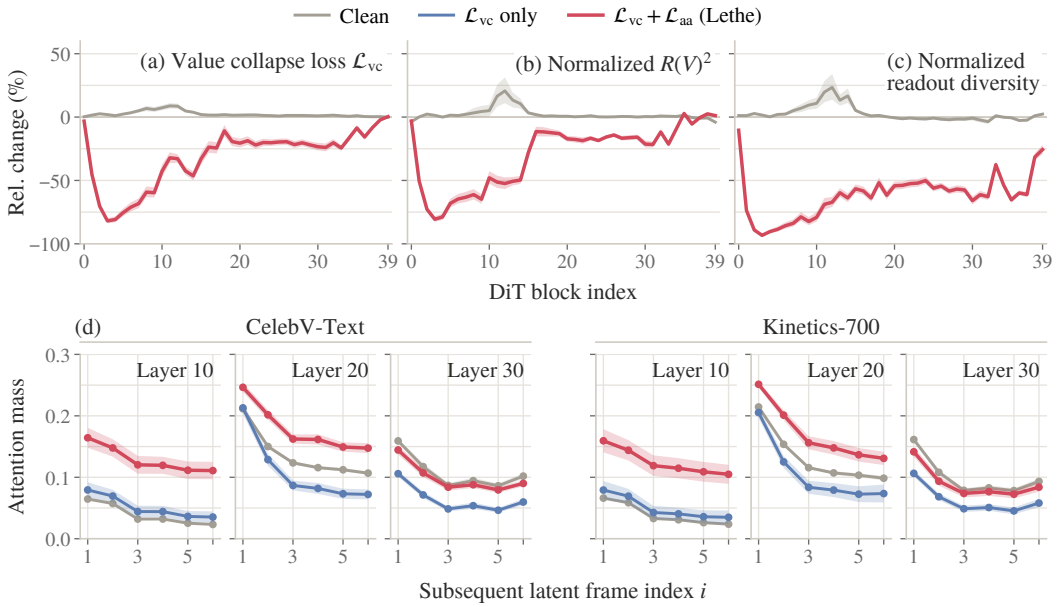


Figure 8: **Mechanism Validation of LETHE.** (a)–(c) LETHE lowers the value collapse loss, and with it the normalized $R(V)^2$ and the readout diversity it bounds, across DiT blocks. (d) LETHE draws more attention from subsequent latent frames to the reference frame at different blocks.

roughly reaches the level of a 25% contraction ($\gamma = 0.75$) and exceeds it in many blocks. This indicates that the disruption mechanism revealed by the intervention can indeed be triggered by a small perturbation ($\epsilon = 16/255$) in the input space. Moreover, as shown in Fig. 8(d), incorporating $\mathcal{L}_{aa}$ increases the attention mass that subsequent frames allocate to the reference tokens, especially at early layers ($\sim 40\%$), compared with the $\mathcal{L}_{vc}$-only variant, which demonstrates the effectiveness of this objective in directing subsequent frames to read from the reference. Interestingly, with $\mathcal{L}_{vc}$ alone, subsequent frames allocate less attention to the reference tokens than with clean images. This may be because, as the reference values collapse and no longer provide sufficient information, the model turns to reading more from other tokens. This trend is consistent across both datasets and agrees with the complementary effect of the two objectives observed in the ablation study. Overall, these observations suggest that the relations in Appendix B are considerably tighter in practice than in the worst case, and that LETHE indeed operates through the reading pathway identified in Sec. 3.

Table 4: Effect of each objective of LETHE (mean ± std over three runs).

Dataset	Variant	ISM-A ↓	ISM-F ↓	QAF ↓	QAV ↓
CelebV-Text	Clean	0.8173 ± 0.0120	0.8725 ± 0.0038	0.9123 ± 0.0057	0.9462 ± 0.0021
	$\mathcal{L}_{\text{vc}}$ only	0.5899 ± 0.0156	0.6827 ± 0.0335	0.5224 ± 0.1624	0.6009 ± 0.1603
	$\mathcal{L}_{\text{vc}} + \mathcal{L}_{\text{aa}}$ (LETHE)	0.4280 ± 0.0051	0.5308 ± 0.0138	0.3970 ± 0.0389	0.4976 ± 0.0428
Kinetics-700	Clean	0.7584 ± 0.0158	0.8242 ± 0.0185	0.7974 ± 0.0239	0.8381 ± 0.0205
	$\mathcal{L}_{\text{vc}}$ only	0.5084 ± 0.0284	0.6127 ± 0.0100	0.4600 ± 0.0801	0.5324 ± 0.0890
	$\mathcal{L}_{\text{vc}} + \mathcal{L}_{\text{aa}}$ (LETHE)	0.4125 ± 0.0168	0.4957 ± 0.0218	0.3852 ± 0.0176	0.4745 ± 0.0184

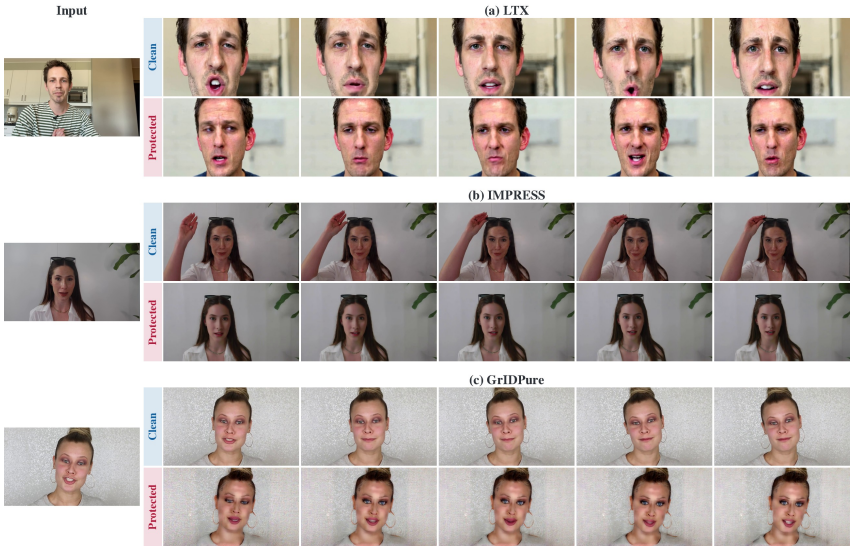


Figure 9: **Challenging cases analysis.** (a) The protected image is rendered by LTX-Video as an unseen black-box model. (b, c) the protected image is first passed through a strong generative purifier, IMPRESS and GrIDPure respectively, and then re-rendered by Wan2.1-I2V. **Left**: reference image. **Right**: five uniformly sampled frames from the clean (top) and protected (bottom) generations.

D.2 Challenging Case Analysis

Across the more challenging settings, we observe an intriguing pattern: the visual quality of the generated videos seems to be slightly easier to recover than the identity, with representative cases being strong purifiers such as IMPRESS and GrIDPure (Fig. 5(b)). To better understand why this happens, we further inspect the videos generated from these purified images. As shown in Fig. 9(b)(c), these videos exhibit smoother textures and fewer visible artifacts than those generated from the protected images without purification, yet the generated faces still differ noticeably from the reference in facial structure and details. We conjecture that this stems from the way these purifiers operate: they typically rely on generative priors to remove the perturbation and re-synthesize fine details and textures in the process, such that the perceptual quality is improved. However, the re-synthesized content comes from the prior rather than from the original image. When the purification is strong enough to remove the perturbation, it also discards much of the original detail, including the facial features that determine the identity, so that the identity similarity remains low. In other words, the purified videos look more natural but still do not depict the person in the reference image. Thus, we consider that such quality recovery does not constitute an effective bypass of the protection.

A similar phenomenon is observed when transferring to LTX-Video (Fig. 5). One possible reason is that LTX-Video adopts a VAE with a very high compression ratio (HaCohen et al., 2024), which discards much high-frequency information during encoding and may thus act similarly to a strong generative purifier. Our qualitative results in Fig. 9(a) also confirms that the generated identity remains largely disrupted, suggesting the effectiveness of our method.

D.3 LIMITATIONS & FUTURE WORK

Limitations. Our work still has the following limitations, which we aim to address in future work. *First*, there remains a gap between our theoretical analysis and the practical protection effect. Proposition 1 characterizes the readout within a single attention head, whereas deriving an end-to-end guarantee on the final video through stacked layers, residuals, iterative denoising, and VAE decoding is notoriously difficult, if not impossible, given that dot-product self-attention is not globally Lipschitz continuous (Kim et al., 2021) and "identity / video quality collapse" itself lacks a formal mathematical formulation. Nevertheless, our intervention experiment and the consistent results across diverse models, datasets, and settings empirically validate the mechanism revealed by our analysis. Therefore, we regard our analysis as a principled guidance that motivates the design of LETHE, rather than as a definitive guarantee of the protection. *Second*, similar to prior findings on protective perturbations (Liang et al., 2023; Song et al., 2026), the effectiveness of LETHE depends on the budget and on the surrogate-target similarity. Regarding the budget, the protection becomes weaker when ϵ is very small, possibly because modern video diffusion models are large and trained on massive data, so that overly small perturbations (e.g., $< 4/255$) may be insufficient to induce effective disruption. Regarding the models, the protection is also somewhat weaker when transferred to models with substantially different architectures, which has long been a challenge for adversarial perturbations in general (Gu et al., 2023). That said, LETHE at $8/255$ already achieves protection comparable to that of existing methods at $16/255$, and it also provides state-of-the-art effectiveness on models not involved in the optimization. To the best of our knowledge, no existing I2V protection remains effective at budgets as small as $4/255$, and a gap remains toward protection that is both fully imperceptible and effective. We acknowledge that LETHE does not fully close this gap either. Nevertheless, we believe that it takes a meaningful step toward this goal. Beyond substantially disrupting unauthorized I2V generation across a wide range of settings, LETHE also offers a new perspective on why protective perturbations work by revealing how subsequent frames read from the reference image. We hope that our analysis, findings, method, and re-annotated datasets can motivate further investigation toward reliable protection of personal images against I2V generation.

Future Directions. There are various potential directions worthy of exploration for future work. First, one interesting empirical observation is that, although some models differ substantially from our surrogate models in how they incorporate the reference image, LETHE still exhibits notable transferability to them (Fig. 5(a)). For example, SVD adopts a UNet architecture and injects the reference latent into every frame via channel concatenation, which falls outside the attention-based reading pathway analyzed in Section 3. We *hypothesize* that this is possibly because the reduced diversity of the reference representation induced by LETHE is not specific to the surrogate models, since tokens carrying reference information are still aggregated through softmax attention in these models (e.g., the spatial self-attention in SVD). Verifying this explanation and characterizing the conditions under which such transferability holds are valuable directions for future work. Besides, the transferability of LETHE may be further improved by ensembling multiple surrogate models or adopting more advanced gradient aggregation techniques (Chen et al., 2024; Zhang et al., 2024). Finally, our analysis of the attention-based reading mechanism does not, in principle, depend on a specific architecture: whenever conditioning information is represented as tokens and read through softmax attention, the readout lies within the convex hull of the corresponding values. Therefore, we believe that this perspective may extend to other reference-conditioned generation tasks, such as 3D generation (Zhao et al., 2026; Hong et al., 2026), which we leave for future work.

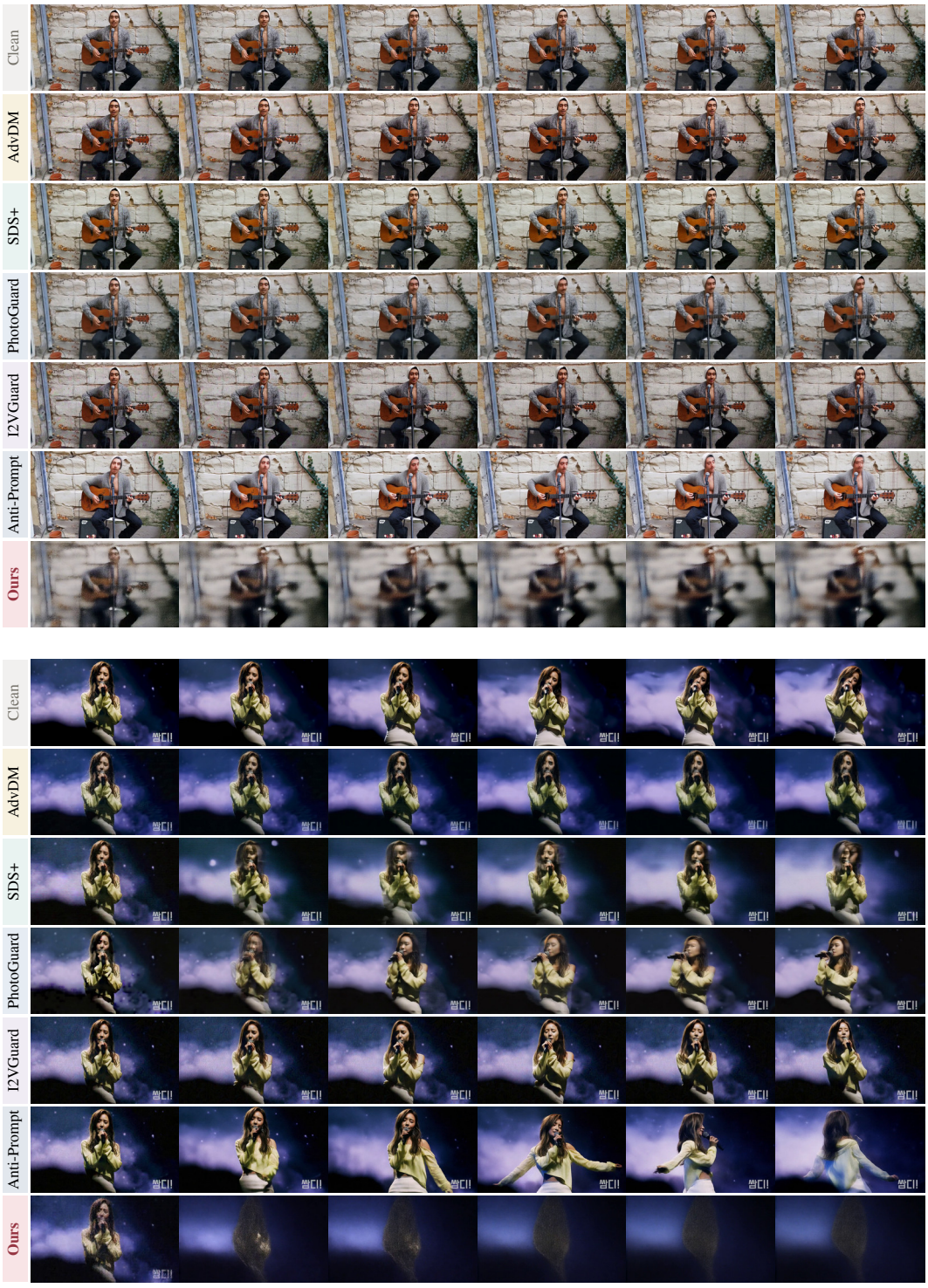


Figure 10: **Additional qualitative results on Wan2.1** ($\epsilon = 16/255$). Most baselines only introduce mild artifacts or alter the motion, while LETHE erases the subject or dissolves it into the background.

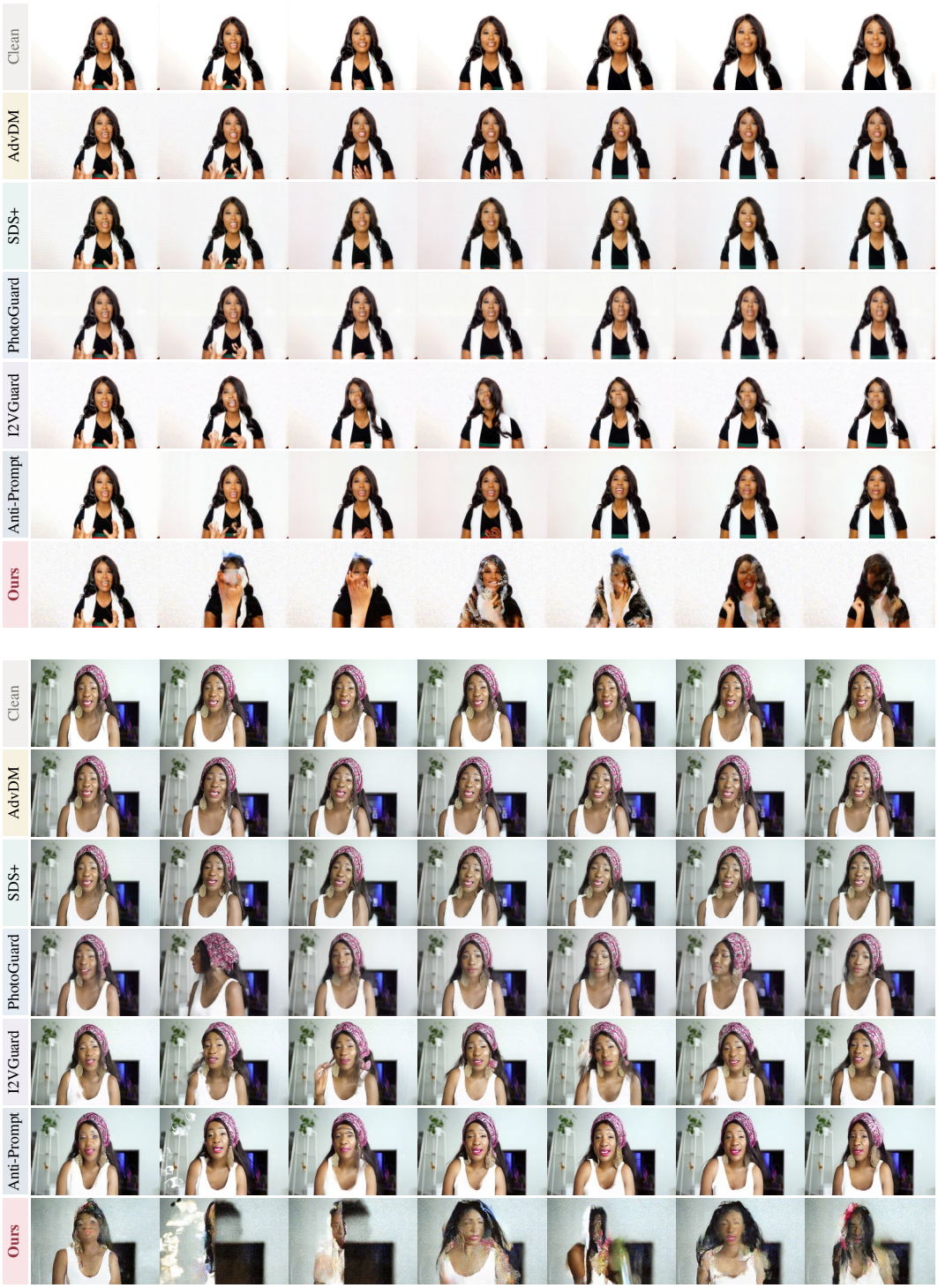


Figure 11: **Additional qualitative results on CogVideoX** ($\epsilon = 16/255$). Most baselines still keep the subject largely recognizable, whereas LETHE smears the face into unrecognizable textures.