ON-POLICY ADVERSARIAL FINE-TUNING FOR LONG-HORIZON AUTOREGRESSIVE VIDEO GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Autoregressive video generation enables efficient long-horizon synthesis by repeatedly conditioning on its own outputs. However, this feedback allows small perturbations to a conditioning image to influence generation far beyond the initial frames. We investigate this vulnerability in distilled autoregressive video models. Our analysis links long-horizon sensitivity to local prediction discrepancies and their propagation, motivating corrective supervision during generation. We propose teacher-guided on-policy adversarial fine-tuning to improve the input robustness of distilled models while retaining their few-step, closed-loop long-video generation capabilities. A frozen self-teacher provides corrective supervision along the model's own adversarial rollouts, with regularization preserving its original generative behavior. We further design an efficient attack surrogate to sustain frequent adversarial exposure during adaptation. Across long-video evaluations, our approach reduces mean latent perturbation response by 22.3–31.4% relative to the pretrained baseline, with a 14.5% reduction in later frames under attacks specifically targeting late-video generation. VBench-I2V further shows improved quality under attack with near-unchanged clean scores. These results demonstrate that distilled autoregressive generators can learn greater input robustness without additional inference cost.

1 INTRODUCTION

Autoregressive video generation supports real-time, controllable long-video synthesis, providing a foundation for interactive content creation and predictive world models (Yang et al., 2025b; Chen et al., 2024). Conditioning on generated history extends synthesis beyond a fixed context; diffusion distillation reduces its denoising cost (Salimans and Ho, 2022; Song et al., 2023; Yin et al., 2024a). CausVid distills a bidirectional teacher into a causal generator (Yin et al., 2025), while Self-Forcing trains with the model's own generated history (Huang et al., 2025). Yet sustaining generation does not ensure robustness to changes in its input.

This feedback also creates a robustness challenge: a small perturbation to the conditioning image can alter predictions that become inputs to later predictions. Its influence can outlast the image's presence in the attention window. We study how to reduce the effects of bounded image perturbations throughout long continuation, not only in the initial clip.

Evaluating native Self-Forcing reveals that perturbation effects grow over extended rollouts (Figure 1). Our analysis links this growth to prediction differences propagating through generated history. We also evaluate a more robust baseline, HiAR (Zou et al., 2026). Its hierarchical denoising with noise-matched history not only mitigates generation drift but also exhibits lower perturbation sensitivity. Yet substantial sensitivity remains, making HiAR a useful starting point for learning further robustness.

The training challenge is to add robustness without undoing the distilled model's *few-step, closed-loop long-video generation capabilities*. Under attack, the model continues from its own perturbed predictions, which change as training progresses. Robustness learning must therefore address both the states encountered by the evolving generator and the supervision used to correct them without sacrificing generation detail.

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: **Small input perturbations can alter long-video content.** *Left:* mean latent response to paired random image perturbations (16-image VBench-I2V subset; $8/255$ bound; shaded 95% source-bootstrap intervals). Baseline is native HiAR; ours lowers its mean late response by 54.0%. *Right:* paired outputs at 19.5 s with matched generation noise. Red dashed boxes mark a garment-color mismatch under perturbed input in the baseline; ours preserves the color.

We propose *teacher-guided on-policy adversarial fine-tuning*. A frozen copy of the pretrained generator serves as a self-teacher, providing corrective supervision along fresh adversarial rollouts. The teacher uses the unperturbed image condition within the student's current generation context, correcting its ongoing prediction rather than pulling it toward a separately generated reference trajectory. To make frequent adversarial exposure practical, we use a lightweight attack surrogate with full-generator scoring to refresh inputs as the model evolves. Regularization on normal inputs constrains changes to its original generative behavior. Adaptation thus follows closed-loop generation without changing native inference, further reducing sensitivity in Figure 1.

In summary, we make the following contributions:

- We characterize input perturbations as a long-horizon robustness problem in distilled autoregressive video generation. Our analysis separates locally introduced prediction differences from their propagation through generated history, explaining how their influence can persist during continuation.
- We introduce teacher-guided on-policy adversarial fine-tuning that provides corrective self-teacher guidance in the student's own generated context. Efficient surrogate-based attack refreshes and normal-input regularization support robustness learning while preserving native inference.
- We demonstrate sustained robustness gains in twenty-second videos: mean latent response falls by 22.3–31.4%, while VBench-I2V shows improved attacked quality with near-unchanged clean scores. A teacher-target control reveals that lower sensitivity can conceal a loss of generation detail, establishing the importance of supervision design alongside adversarial exposure and normal preservation.

2 Method

2.1 Perturbations in a causal video generator

Problem. Let $G_\theta(c, p; \omega)$ generate a video from source image $c \in [0,1]^d$, prompt p, and sampling randomness ω. An attacker supplies $c + \delta$, with $\delta \in \Delta_\epsilon(c) = \{\delta : \|\delta\|_\infty \le \epsilon, c + \delta \in [0,1]^d\}$, without changing the prompt or inference procedure. Our goal is to reduce the resulting change over long continuation while retaining normal generation. Deployment requires only the delivered image, not its clean counterpart or the teacher.

Why the disturbance can persist. Index execution updates in causal order and let h_n include generated history, pending denoising queries, and inherited cache state. A first-order expansion of clean and perturbed executions gives

$$\Delta h_n = A_n \Delta h_{n-1} + B_n \delta + o(\|\delta\|), \qquad \Delta h_N \simeq \Phi_{N:1} \Delta h_0 + \sum_{k=1}^{N} \Phi_{N:k+1} B_k \delta, \tag{1}$$

where $\Phi_{N:k+1} = A_N \cdots A_{k+1}$. The sum separates new source-induced changes from their subsequent propagation. When the source is no longer read directly, $B_n = 0$, but changes already encoded in generated history can remain. Aligned paths can accumulate; different paths can also attenuate or cancel.

HiAR writes generated context at the native transition's output noise level:

$$H_{j,b} = \alpha_j u_{j,b} + b_j \xi_{j,b}, \qquad \alpha_j = 1 - \sigma_{t_{j+1}}, \quad b_j = \sigma_{t_{j+1}}. \tag{2}$$

With coupled noise, $\Delta H_{j,b} = \alpha_j \Delta u_{j,b}$. Intermediate noise levels therefore attenuate the discrepancy presented to the next predictor relative to clean-history conditioning. The predictor's history sensitivity also depends on noise level, and the final clean stage has $\alpha_j = 1$. Noise matching provides an attenuation route, not immunity. We therefore correct predictions along the student's actual generation, before their discrepancies propagate through this fixed inference structure.

2.2 On-policy guidance from a frozen self-teacher

At stage/block node $i = (j, b)$, let $s_i = (q_i, H_i, p, \Gamma_i)$ contain the current noisy query, raw generated history, prompt, and causal replay metadata. The latter records source visibility and write order, including ancestors of inherited cache entries beyond the visible window. The student's clean-latent prediction is $F_{\theta,i}(s_i, a)$ for delivered image a. We draw s_i from a fresh rollout of the current student and query the frozen original generator with the unperturbed source:

$$T_i(s_i, c) = F_{\theta_0,i}(s_i, c), \qquad s_i \sim d_{\theta,a}. \tag{3}$$

Here $d_{\theta,a}$ is the current student's rollout-state distribution. Student and teacher see the same query and raw history, but each reconstructs its own KV cache under its own parameters and source condition. Thus the target answers the student's current denoising problem rather than asking it to jump to an independently generated clean trajectory. For robust updates, $a = c + \delta$; for clean updates, $a = c$. In the proposed context-matched model, both use the prediction loss

$$\mathcal{L}_m(\theta) = \mathbb{E}_{c,p,\omega,i}\left[\frac{1}{n_i}\|F_{\theta,i}(s_i^m, a_m) - \operatorname{sg} T_i(s_i^m, c)\|_2^2\right], \quad s_i^m \sim d_{\theta,a_m}, \quad m \in \{\text{robust}, \text{clean}\}. \tag{4}$$

Here n_i is the predicted latent's dimensionality. Robust updates reduce source-conditioned discrepancies on perturbed student states; clean updates constrain departure from the original generator on normal student states. The main schedule uses two robust updates per clean update. Our target control changes only the robust label to the teacher's own clean-trajectory prediction, coupled by source, noise, stage and block. Both alternatives retain fresh student rollouts, isolating target choice rather than on-policy versus off-policy state exposure.

Matching native transitions. Prediction matching supervises the generator's native transition rather than requiring a score estimate at an arbitrary noise level. For the next latent $y = \alpha_j F_{\theta,i}(s, a) + b_j \xi$, student and teacher share the same standard Gaussian noise law and differ only in translation. Their next-latent transition kernels therefore satisfy

$$W_2^2\big(K_{\theta,i}^Z(\cdot \mid s, a), K_{0,i}^Z(\cdot \mid s, c)\big) = \alpha_j^2 \|F_{\theta,i}(s, a) - T_i(s, c)\|_2^2. \tag{5}$$

The per-node loss is therefore $W_2^2/(n_i \alpha_j^2)$: a stage-reweighted transport cost, including the deterministic final stage. The identity follows by coupling the same noise (upper bound) and applying Jensen's inequality to any coupling (matching lower bound). It concerns the next-latent marginal, not the full cache, and uses the existing generator rather than the score models required by DMD (Yin et al., 2024b). Conditional transition matching is related to diffusion OPD (Li et al., 2026); our design applies it to perturbed autoregressive states with a trusted source condition.

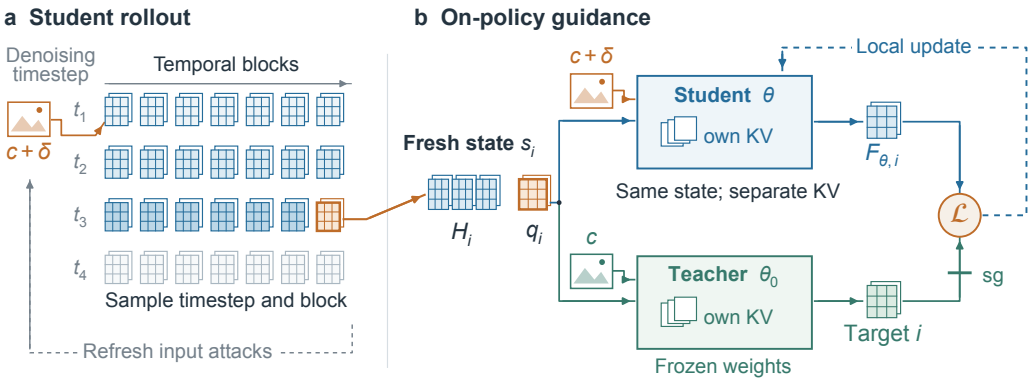


Figure 2: **Teacher-guided on-policy adversarial fine-tuning.** *(a) Student rollout:* sample a denoising timestep and temporal block from a fresh adversarial rollout. *(b) On-policy guidance:* student and frozen teacher use the same query and generated history, but perturbed and clean images, respectively. Prediction matching locally updates only the student (sg: stop-gradient); each model builds its own KV cache. Attacks are refreshed during training; native student inference is unchanged.

2.3 Why local guidance can help long continuation

Fix paired inputs, prompt and noise, and order the M prediction nodes causally. For input $c+\zeta$, $\zeta \in \{0, \delta\}$, denote the student prediction by u_i^ζ, its teacher residual by $\eta_i^\zeta = u_i^\zeta - T_i(s_i^\zeta, c)$, and its norm by $r_i^\zeta = \|\eta_i^\zeta\|$. Let $e_i = \|u_i^\delta - u_i^0\|$ measure paired response. The identity

$$u_i^\delta - u_i^0 = \eta_i^\delta - \eta_i^0 + T_i(s_i^\delta, c) - T_i(s_i^0, c) \tag{6}$$

separates local residuals from the teacher's response to differing histories.

Proposition 1 (Residuals along causal paths)**.** *Suppose the teacher composed with raw-state construction satisfies* $\|T_i(s_i^\delta, c) - T_i(s_i^0, c)\| \le \sum_{k<i} L_{ik} e_k$ *with nonnegative* L_{ik}*. Then*

$$e \preceq G(r^\delta + r^0), \qquad G = (I - L)^{-1} = \sum_{\ell=0}^{M-1} L^\ell. \tag{7}$$

Proof. Equation (6) and the triangle inequality give $e \preceq r^\delta + r^0 + Le$. Substitution terminates since the causal matrix L is strictly lower triangular and $L^M = 0$. □

The branches reduce r^δ and r^0 where the student encounters them; G collects downstream influence without requiring contraction or changed inference. Under an analogous sensitivity bound relative to original clean execution u^*, $e^{\zeta,*} \preceq G_* r^\zeta$, with $e_i^{\zeta,*} = \|u_i^\zeta - u_i^*\|$ and $G_* = (I - L_*)^{-1}$. Robust guidance targets attacked-reference deviation; clean regularization anchors normal behavior. Generalization beyond training states is tested empirically, not guaranteed by this bound.

Local correction versus clean-trajectory recovery. For a clean teacher trajectory s_i^*, the alternative residual is $u_i^\zeta - T_i(s_i^*, c) = \eta_i^\zeta + [T_i(s_i^\zeta, c) - T_i(s_i^*, c)]$. It mixes local correction with a context-dependent target offset. Our target control tests the resulting effect on generation detail.

2.4 Efficient adversarial exposure

A refresh snapshots the current student and searches for an input that increases its deviation from its own clean generation:

$$\max_{\delta \in \Delta_\epsilon(c)} \mathrm{MSE}(Z_{\theta_r}(c+\delta, p; \omega_r), Z_{\theta_r}(c, p; \omega_r)) \,. \tag{8}$$

Training uses a seven-block surrogate with current student weights: each block predicts a clean latent from its initial noise at timestep 1000, writing the prediction at timestep 0 as history for later

blocks, without intermediate re-noising. Two projected sign-gradient steps of $4/255$ maximize the surrogate's own-clean MSE within $\epsilon = 8/255$. The initial and final PNGs are rescored by the native four-stage generator using Eq. (8), retaining the better candidate. Every six parameter updates we refresh the perturbations using new source images and current weights. Every parameter update independently regenerates student states and teacher targets; only the attacked input is reused. This is adversarially selected exposure for the teacher loss, not exact minimax optimization of that same loss.

We sample a denoising stage and block for local parameter-gradient replay. The preceding computation is executed with the current student, and only future stages are omitted. Gradients pass through the selected query, source write and latest history write; older history dependencies are detached. All generator parameters are trainable. This retains on-policy forward states while avoiding full-rollout parameter backpropagation. Evaluation attacks instead differentiate the complete causal graph over their specified horizon.

3 Experiments

We evaluate sustained robustness, generation quality and the training components.

3.1 Setup and measures

Models and data. The student and frozen teacher start from original HiAR EMA; all 1.42B student parameters are trained. We use AdamW (learning rate 10^{-5}, betas $.9/.999$, epsilon 10^{-8}, zero weight decay), gradient clipping at one, and batch size eight on eight A800 GPUs. BF16 execution uses FP32 master weights and optimizer states. The 200-update schedule contains 133 robust and 67 clean updates, 33 attack refreshes, and 50 samples of each native timestep $[1000, 937, 833, 625]$. Refreshes cover 264 source assignments from a 3,840-item filtered OpenVid pool (Nan et al., 2025); 240 parent-disjoint items are reserved for validation. We use existing scene captions, center-crop first frames to the target aspect ratio, and resize to 832×480. Training uses seven generated blocks to limit rollout cost; evaluation extends to 27 blocks to test long continuation. Controls retain initialization and source/noise/ timestep schedules except for their stated intervention.

Latent-response evaluation. Outputs contain 27 blocks (about 20 seconds), with a 21-latent sliding context and source eviction. Eighty OpenVid development sources support controls; 160 separately fixed OpenVid validation sources confirm the unchanged model. Both sets are disjoint from training by parent and first-frame hash. Figure 1 instead uses a fixed sixteen-image VBench-I2V subset for the paired random-perturbation comparison; it is not the OpenVid confirmation set. On OpenVid, each model receives clean inputs, paired random perturbations, and its own attack, with matched prompts and generation noise. Random perturbations sample each pixel channel uniformly from $\{-8, \dots, 8\}/255$, then clip to the image range. White-box PGD uses fixed, known generation noise, $\epsilon = 8/255$, step $1/255$, a paired random start, and best quantized candidate selection. PGD20/40 optimize seven blocks with a 22-latent non-evicting context before long replay; only the first six clean blocks coincide with deployed inference. PGD10 instead attacks blocks 21–27 through the deployed full history on the first forty development sources. Evaluation attacks differentiate the native four-stage generator, encoder and complete causal ancestors, without training-time gradient truncation. These comparisons use one training seed per configuration and fixed generation noise; 95% whole-source bootstrap intervals describe source variation, not training-seed or cross-domain uncertainty.

Response and preservation. Let $Z_{\theta,b}$ denote the final latent of block b. For a block set $\mathcal{B}$, define

$$S_{\theta,\mathcal{B}}(\delta) = \frac{1}{|\mathcal{B}|} \sum_{b \in \mathcal{B}} \frac{\|Z_{\theta,b}(c+\delta) - Z_{\theta,b}(c)\|_2^2}{n_b}, \tag{9}$$

$$Q_{\theta,\mathcal{B}}(\delta) = \frac{1}{|\mathcal{B}|} \sum_{b \in \mathcal{B}} \frac{\|Z_{\theta,b}(c+\delta) - Z_{\theta_0,b}(c)\|_2^2}{n_b}, \quad N_{\theta,\mathcal{B}} = \frac{1}{|\mathcal{B}|} \sum_{b \in \mathcal{B}} \frac{\|Z_{\theta,b}(c) - Z_{\theta_0,b}(c)\|_2^2}{n_b}. \tag{10}$$

Prompt and noise are paired and suppressed in the notation. S isolates the input's effect on each model; Q measures attacked-output deviation from original clean behavior; N measures normal

Table 1: **Long-video robustness and design controls.** Eighty development sources; model-specific prefix PGD20 and 27-block outputs. Late: blocks 21–27; otherwise all blocks. Updates count actual optimizer steps: the no-clean control skips 67 scheduled positions.

		Own-model PGD20			Random	Clean
Model / intervention	Updates	$S \downarrow$	Late $S \downarrow$	$Q \downarrow$	$S \downarrow$	$N \downarrow$
Original HiAR	0	1.9397	1.8052	1.9397	.3529	0
Ours	200	1.3539	1.3273	1.4193	.2162	.4949
Clean-trajectory target	200	.9500	.8971	1.2232	.0898	.5639
Teacher states: clean branch	200	1.4184	1.3763	1.4002	.1825	.5251
Teacher states: robust branch	200	1.4709	1.4651	1.5320	.2784	.4899
Random training perturbations	200	1.8244	1.8126	1.8764	.2204	.5244
No clean-preservation updates	133	1.3103	1.2640	1.5095	.1382	.5468
Longer training	400	1.4043	1.4931	1.4517	.1641	.6441

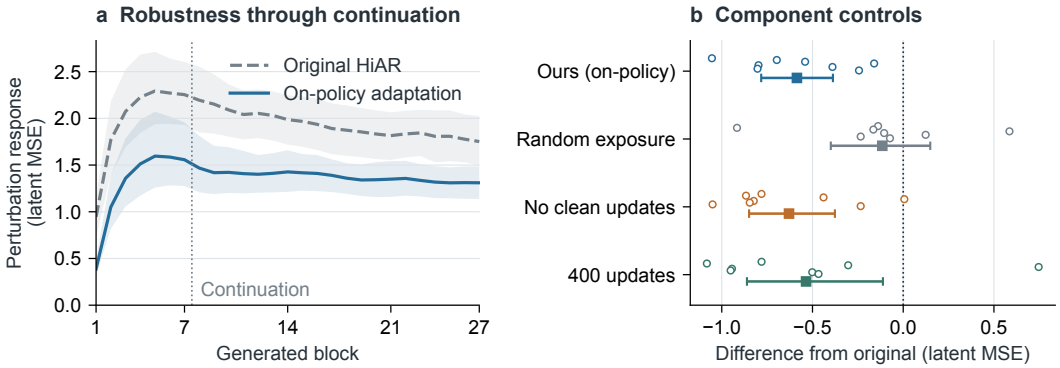


Figure 3: **The benefit persists beyond the attacked prefix.** Eight-source subset. (a) Own-PGD20 response; the dotted line marks the optimized prefix. (b) Full-video differences from original HiAR: circles are sources, squares means. Bands show pointwise 95% source-bootstrap intervals; bars show paired 95% intervals for mean differences; negative differences favor adaptation.

drift. All means weight sources equally. We separately assess input alignment, temporal smoothness, motion activity and visual quality; latent distance alone is not a measure of semantic correctness.

3.2 Long-video robustness

Adaptation reduces twenty-second response by **30.2%**, improving all eighty development sources (Table 1). Gains persist through feedback and context turnover: early, middle and late reductions are 34.6%, 29.7% and **26.5%** (Figure 3); attacked-reference deviation also falls 26.8%.

Additional sources. The fixed 160-source confirmation retains **22.3%** overall and **20.6%** late reductions, improving 150/160 and 130/160 sources (Tables 3–4). The full-video difference has 95% interval $[-.4727, -.2405]$; Q decreases 20.1%. Excluding ten sources previously seen in a separate monitoring run leaves 21.1% overall and 19.2% late reductions, without reselection.

Normal behavior and random perturbations. Random-input S falls 38.7% on development sources, but random-input Q rises 38.9%; clean drift is .4949. On confirmation sources, random S falls from .4564 to .1546 while Q rises to .5411. Reduced sensitivity therefore does not imply unchanged normal behavior.

3.3 Overall video quality

Beyond latent response, we evaluate original HiAR and Ours using official VBench-I2V implementations and dimension-specific image subsets (Huang et al., 2024b). Each image–prompt pair yields

Table 2: **Clean quality is retained; attacked quality improves.** VBench-I2V scores (0–100). Official subsets: 246 subject images, 109 background images, and 355 pooled images for Aesthetic/Imaging Quality. † Motion activity, not a standalone quality measure.

Input	Model	I2V Subject ↑	I2V Background ↑	Motion Smoothness ↑	Dynamic Degree †	Aesthetic Quality ↑	Imaging Quality ↑
Clean	Original HiAR	96.80	97.50	98.00	50.00	62.00	69.00
	Ours	96.70	97.40	97.90	49.19	61.80	68.80
Adaptive attack	Original HiAR	90.90	93.10	95.80	40.98	55.30	61.90
	Ours	**94.96**	**96.04**	**96.72**	**48.05**	**59.39**	**66.80**

Table 3: **Robustness across sources and attack protocols.** Model-specific attacks; 27-block outputs. Prefix attacks optimize blocks 1–7; late attacks optimize 21–27 through full history. Only the first two rows isolate attack iterations; the last also changes source subset.

Sources	Attack target / budget	Original S	Ours S	Full reduction	Late reduction
Development ($n = 80$)	Prefix / PGD20	1.9397	1.3539	30.20%	26.47%
Development ($n = 80$)	Prefix / PGD40	2.4726	1.6972	31.36%	25.95%
Confirmation ($n = 160$)	Prefix / PGD20	1.6065	1.2487	22.27%	20.58%
Development ($n = 40$)	Late / PGD10	.9181	.7873	14.26%	14.52%

five 20-second videos, with seeds paired across models and clean/attacked inputs. Own-model prefix PGD20 (Section 3.1) maximizes latent deviation; VBench measures the resulting quality using unperturbed official images as I2V references, rather than supplying attack objectives.

The five quality and alignment scores stay within 0.2 points of the original on clean inputs and all improve under attack (Table 2), including +4.06 in I2V Subject and +4.90 in Imaging Quality. Dynamic Degree remains near its clean level (48.05 versus 49.19), supporting retained motion alongside improved content alignment and visual quality.

3.4 Attack-budget and late-horizon tests

Increasing optimization effort. PGD40 resumes the twentieth iterate and selects from all 41 candidates. It raises full-video response by 27.5% for HiAR and 25.4% for Ours, yet retains **31.4%** overall and 26.0% late reductions (Table 3). The gain survives attacks stronger than the training surrogate and the initial twenty-step evaluation.

Attacking the end of the video directly. Under full-history late-target PGD10, response falls **14.5%** late and 14.3% overall (Table 4). Thirty of forty sources improve; source 2's late response rises 2.7%. Late Q is essentially unchanged (+0.1%), separating reduced sensitivity from closer agreement with original clean behavior. The final update still improves twenty adapted-model attacks; these are finite-budget tests, not convergence or worst-case certificates.

3.5 Teacher guidance must preserve generation

We replace only the robust teacher label with its prediction along an independently generated clean trajectory, coupled by source, noise, timestep and block. Current-student rollouts and clean updates remain unchanged: both alternatives are on-policy, but their targets differ.

Smaller distances, poorer generation. Clean-trajectory targets reduce response another 29.8% on development and 24.5% on confirmation sources, with lower Q as well. Yet all 240 sources show late blur or ghosting at fixed sampled times, including clean inputs (Figure 4). Lower latent distances conceal a generation-preservation failure.

Context matching retains technical quality. MUSIQ (Ke et al., 2021) uses VBench's SPAQ checkpoint, with the longer side resized to 512 pixels while preserving aspect ratio (Huang et al.,

Table 4: **Benefits persist across the continuation.** Window-wise latent MSE on additional sources and under direct late attacks. For original HiAR, $Q = S$ and $N = 0$.

Evaluation	Blocks	Original S	Ours S	Reduction	Ours Q	Ours N
Confirmation	1–7	1.4718	1.0953	25.58%	1.0896	.3768
($n = 160$, PGD20)	8–20	1.6697	1.3093	21.58%	1.3432	.5441
	21–27	1.6238	1.2896	20.58%	1.3691	.6456
	1–27	1.6065	1.2487	22.27%	1.2842	.5271
Late objective	1–7	.6443	.5389	16.35%	.5835	.3495
($n = 40$, PGD10)	8–20	.9555	.8281	13.33%	.9392	.6130
	21–27	1.1227	.9597	14.52%	1.1239	.7143
	1–27	.9181	.7873	14.26%	.8949	.5709

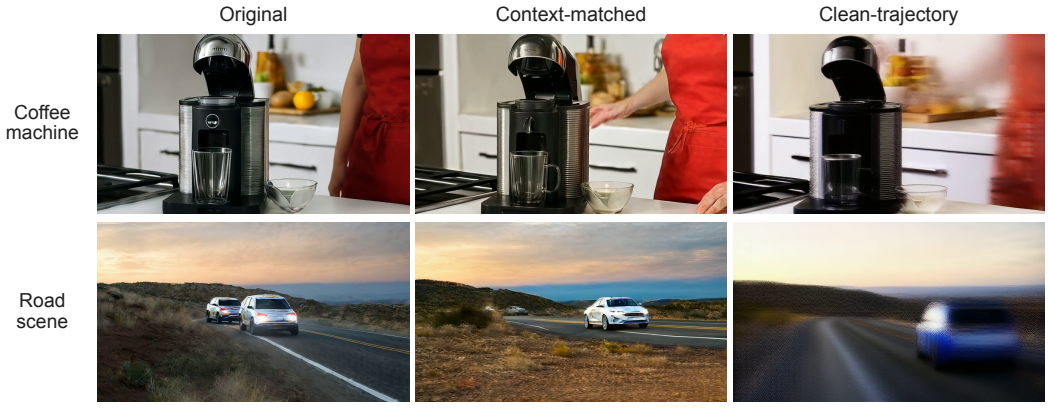


Figure 4: **Reduced response can conceal a loss of generation detail.** Clean outputs at 18 s from the first source of each set, selected by index. Independent-trajectory targets produce blur and repeated edges; context matching retains sharper detail. Frames are unenhanced.

2024a). Trajectory targets score lower on every source, clean and attacked (Table 5). Their clean-score deficit is 17.80 points (95% interval $[-22.52, -13.31]$ for trajectory minus Ours), whereas Ours differs from original HiAR by $+0.09$ ($[-1.06, 1.36]$). From 2 to 18 seconds, the trajectory-target mean drops 69.30→46.69; Ours remains 73.79→75.64. Correcting predictions in the student's own context retains detail that the independent-trajectory target loses.

3.6 What each training component contributes

Adversarial exposure, not noise alone. Replacing optimized inputs with random perturbations leaves only a 5.9% response reduction, versus 30.2% for Ours, with unchanged source allocation, fresh rollout rule and normal updates. Ours improves on this control on 70/80 sources: attack selection adds value beyond noise augmentation.

On-policy states in both branches. Replacing fresh student states with frozen-teacher states raises mean response 4.8% in the clean-branch control and 8.6% in the robust-branch control. The donor rolls out on that branch's input; the student predicts at its query/history and the teacher labels the same state using the clean image. All other settings are fixed, but each student refreshes its own attacks. Ours has the best observed mean; paired intervals include zero (Table 6). These are complete adaptive training comparisons, not fixed-attack state interventions.

Clean preservation anchors the original behavior. Removing clean updates lowers response another 3.2%, but increases normal drift by 10.5% and attacked-output deviation from the original clean reference by 6.4%. This control retains all 133 robust updates and skips the 67 clean optimizer steps (Table 1), isolating the additional preservation schedule rather than matching total updates. Clean guidance anchors behavior, not merely sensitivity.

Table 5: **Target choice separates stability from generation quality.** Own-PGD20, 27-block outputs; context-matched denotes Ours. MUSIQ-SPAQ averages frames at 2/10/18 s within each source, then sources equally (4,320 frames). It measures technical image quality, not semantics.

		Latent distance ↓			MUSIQ ↑	
Sources	Teacher target	Own S	Own Q	Clean N	Clean	Own attack
Development ($n = 80$)	Original HiAR	1.9397	1.9397	0	75.50	72.45
	Context-matched	1.3539	1.4193	.4949	75.69	76.26
	Clean trajectory	.9500	1.2232	.5639	55.81	48.59
Confirmation ($n = 160$)	Original HiAR	1.6065	1.6065	0	74.83	74.54
	Context-matched	1.2487	1.2842	.5271	74.87	75.16
	Clean trajectory	.9433	1.1155	.5242	58.11	52.99

Table 6: **Student-state training has the lowest observed mean.** Ours minus frozen-state control on eighty sources; negative favors Ours. Intervals use 20,000 paired whole-source bootstrap resamples.

Frozen-state branch	Full ΔS	95% interval	Late ΔS	95% interval
Clean	$-.0645$	$[-.2228, .0814]$	$-.0489$	$[-.2599, .1444]$
Robust	$-.1170$	$[-.3724, .1248]$	$-.1377$	$[-.4095, .1078]$

More short-rollout training is not uniformly better. Resuming weights and Adam states to 400 updates, with fresh sources and unchanged loss/optimizer, improves early response 11.6% but worsens full response 3.7% and late response 12.5%. Training length must be selected at the intended continuation horizon, not the training clip alone.

4 Related work

Efficient video generation. Diffusion and latent-space sampling (Ho et al., 2020; Song et al., 2021; Rombach et al., 2022) underpin modern video synthesis (Ho et al., 2022a;b; Blattmann et al., 2023a; Guo et al., 2024; Blattmann et al., 2023b; Yang et al., 2025a; Kong et al., 2024; Team Wan et al., 2025). Diffusion Forcing uses token-wise noise in causal prediction (Chen et al., 2024); Self-Forcing and HiAR enable few-step generation with model-generated context (Huang et al., 2025; Zou et al., 2026). We target input robustness in this closed-loop regime.

On-policy teacher guidance. Student-generated states connect to imitation learning and distillation (Ross et al., 2011; Agarwal et al., 2024). Flow-OPD uses teacher-based rewards (Fang et al., 2026); DiffusionOPD matches student transitions (Li et al., 2026). D-OPSD adapts distilled models using privileged teachers, including frozen-base guidance (Jiang et al., 2026); OPSD-V uses real-video history and EMA teacher adapters for autoregressive video (Liu et al., 2026). Our frozen teacher instead retains the student's generated context and uses a clean source to supervise robustness.

Adversarial robustness. Our input adversary builds on gradient-based robust training (Goodfellow et al., 2015; Madry et al., 2018; Zhang et al., 2019) and clean-teacher robust distillation (Goldblum et al., 2020), now with generated history. Unlike AFD's teacher–student discrimination (Luo et al., 2026), it acts on source images. I2VGuard and Anti-I2V disrupt generation this way (Gui et al., 2025; Vu et al., 2026); we instead train input robustness. Native-model gradients and larger budgets address evaluation pitfalls (Athalye et al., 2018; Croce and Hein, 2020).

5 Conclusion

Teacher-guided on-policy adversarial fine-tuning improves long-video robustness without changing distilled inference. VBench-I2V shows higher attacked quality with near-unchanged clean scores. The target control demonstrates why guidance must respect generated context: lower sensitivity can conceal degraded detail. Adversarial exposure and normal regularization complement this guidance.

AI USE STATEMENT

We used generative AI tools only to refine the writing and improve the clarity and readability of the manuscript. All content was reviewed and verified by the authors.

ETHICS STATEMENT

This work studies input robustness in pretrained image-to-video generators using existing captioned video data. Source footage may depict people, but the analysis does not identify individuals or infer sensitive personal attributes. Image perturbation methods have dual-use implications: they can reveal vulnerabilities, but can also be used to interfere with generation. Our intended application is improving the reliability of authorized generation. The reported robustness is restricted to the evaluated perturbation settings and does not establish general safety, authenticity, or content-policy compliance. Use and redistribution of source footage and pretrained assets remain subject to their respective terms.

REPRODUCIBILITY STATEMENT

Sections 2–3 specify the learning objective, theoretical assumptions, teacher context, local-gradient convention, optimizer, training schedule, evaluation protocols, and quantitative comparisons. The appendices provide expanded derivations, implementation details, per-source breakdowns, and additional visual diagnostics. Input images, sampling noise, checkpoint identities, and saved outputs were matched when constructing each comparison; latent-response results were recomputed from the saved latent tensors rather than inferred from run completion.

REFERENCES

Rishabh Agarwal et al. On-Policy Distillation of Language Models: Learning from Self-Generated Mistakes. *ICLR*, 2024. https://arxiv.org/abs/2306.13649.

Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples. *ICML*, 2018. https://arxiv.org/abs/1802.00420.

Andreas Blattmann et al. Align your Latents: High-Resolution Video Synthesis with Latent Diffusion Models. *CVPR*, 2023. https://arxiv.org/abs/2304.08818.

Andreas Blattmann et al. Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Datasets. arXiv:2311.15127, 2023. https://arxiv.org/abs/2311.15127.

Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion Forcing: Next-token Prediction Meets Full-Sequence Diffusion. *NeurIPS*, 2024. https://arxiv.org/abs/2407.01392.

Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. *ICML*, 2020. https://arxiv.org/abs/2003.01690.

Zhen Fang et al. Flow-OPD: On-Policy Distillation for Flow Matching Models. arXiv:2605.08063v5, 2026. https://arxiv.org/abs/2605.08063v5.

Micah Goldblum, Liam Fowl, Soheil Feizi, and Tom Goldstein. Adversarially Robust Distillation. *AAAI*, 2020. https://doi.org/10.1609/aaai.v34i04.5816.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and Harnessing Adversarial Examples. *ICLR*, 2015. https://arxiv.org/abs/1412.6572.

Dongnan Gui, Xun Guo, Wengang Zhou, and Yan Lu. I2VGuard: Safeguarding Images against Misuse in Diffusion-based Image-to-Video Models. *CVPR*, pp. 12595–12604, 2025. https://openaccess.thecvf.com/content/CVPR2025/html/Gui_I2VGuard_Safeguarding_Images_against_Misuse_in_Diffusion-based_Image-to-Video_Models_CVPR_2025_paper.html.

Yuwei Guo et al. AnimateDiff: Animate Your Personalized Text-to-Image Diffusion Models without Specific Tuning. *ICLR*, 2024. https://arxiv.org/abs/2307.04725.

Jonathan Ho, Ajay N. Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. *NeurIPS*, 2020. https://arxiv.org/abs/2006.11239.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video Diffusion Models. *NeurIPS*, 2022. https://arxiv.org/abs/2204.03458.

Jonathan Ho et al. Imagen Video: High Definition Video Generation with Diffusion Models. arXiv:2210.02303, 2022. https://arxiv.org/abs/2210.02303.

Ziqi Huang et al. VBench: Comprehensive Benchmark Suite for Video Generative Models. *CVPR*, 2024. https://arxiv.org/abs/2311.17982.

Ziqi Huang et al. VBench++: Comprehensive and Versatile Benchmark Suite for Video Generative Models. arXiv:2411.13503, 2024. https://arxiv.org/abs/2411.13503.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self Forcing: Bridging the Train-Test Gap in Autoregressive Video Diffusion. *NeurIPS*, 2025. https://arxiv.org/abs/2506.08009.

Dengyang Jiang et al. D-OPSD: On-Policy Self-Distillation for Continuously Tuning Step-Distilled Diffusion Models. arXiv:2605.05204v3, 2026. https://arxiv.org/abs/2605.05204v3.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. MUSIQ: Multi-Scale Image Quality Transformer. *ICCV*, 2021. https://arxiv.org/abs/2108.05997.

Weijie Kong et al. HunyuanVideo: A Systematic Framework For Large Video Generative Models. arXiv:2412.03603, 2024. https://arxiv.org/abs/2412.03603.

Quanhao Li et al. DiffusionOPD: A Unified Perspective of On-Policy Distillation in Diffusion Models. arXiv:2605.15055v2, 2026. https://arxiv.org/abs/2605.15055v2.

Hongyu Liu et al. OPSD-V: On-Policy Self-Distillation for Post-Training Few-Step Autoregressive Video Generators. arXiv:2607.08766, 2026. https://arxiv.org/abs/2607.08766.

Yang Luo et al. On-Policy Adversarial Flow Distillation for Autoregressive Video Generation. arXiv:2605.26105, 2026. https://arxiv.org/abs/2605.26105.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards Deep Learning Models Resistant to Adversarial Attacks. *ICLR*, 2018. https://arxiv.org/abs/1706.06083.

Kepan Nan et al. OpenVid-1M: A Large-Scale High-Quality Dataset for Text-to-video Generation. *ICLR*, 2025. https://arxiv.org/abs/2407.02371.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models. *CVPR*, 2022. https://arxiv.org/abs/2112.10752.

Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning. *AISTATS*, 2011. https://proceedings.mlr.press/v15/ross11a.html.

Tim Salimans and Jonathan Ho. Progressive Distillation for Fast Sampling of Diffusion Models. *ICLR*, 2022. https://arxiv.org/abs/2202.00512.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising Diffusion Implicit Models. *ICLR*, 2021. https://arxiv.org/abs/2010.02502.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency Models. *ICML*, 2023. https://proceedings.mlr.press/v202/song23a.html.

Team Wan et al. Wan: Open and Advanced Large-Scale Video Generative Models. arXiv:2503.20314, 2025. https://arxiv.org/abs/2503.20314.

Duc Vu, Anh Nguyen, Chi Tran, and Anh Tran. Anti-I2V: Safeguarding your Photos from Malicious Image-to-video Generation. *CVPR*, pp. 37621–37631, 2026. https://arxiv.org/abs/2603.24570.

Zhuoyi Yang et al. CogVideoX: Text-to-Video Diffusion Models with An Expert Transformer. *ICLR*, 2025. https://arxiv.org/abs/2408.06072.

Shuai Yang, Wei Huang, Ruihang Chu, Yicheng Xiao, Yuyang Zhao, Xianbang Wang, Muyang Li, Enze Xie, Yingcong Chen, Yao Lu, Song Han, and Yukang Chen. LongLive: Real-time Interactive Long Video Generation. *arXiv preprint arXiv:2509.22622*, 2025.

Tianwei Yin et al. One-step Diffusion with Distribution Matching Distillation. *CVPR*, 2024. https://arxiv.org/abs/2311.18828.

Tianwei Yin et al. Improved Distribution Matching Distillation for Fast Image Synthesis. *NeurIPS*, 2024. https://arxiv.org/abs/2405.14867.

Tianwei Yin et al. From Slow Bidirectional to Fast Autoregressive Video Diffusion Models. *CVPR*, 2025. https://arxiv.org/abs/2412.07772.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P. Xing, Laurent El Ghaoui, and Michael I. Jordan. Theoretically Principled Trade-off between Robustness and Accuracy. *ICML*, 2019. https://arxiv.org/abs/1901.08573.

Kai Zou, Dian Zheng, Hongbo Liu, Tiankai Hang, Bin Liu, and Nenghai Yu. HiAR: Efficient Autoregressive Long Video Generation via Hierarchical Denoising. arXiv:2603.08703, 2026. https://arxiv.org/abs/2603.08703.

A IMPLEMENTATION AND REPRODUCIBILITY

Model and optimizer. The student and independently frozen teacher initialize from the original HiAR EMA; all 1,418,996,800 student DiT parameters are optimized. Native timesteps are $[1000, 937, 833, 625]$, with scheduler sigmas governing re-noising. The source is written at noise level zero and generated history at t_{j+1}, ending with clean output. AdamW uses learning rate 10^{-5}, betas $(0.9, 0.999)$, epsilon 10^{-8}, zero weight decay, gradient clipping at one, and global batch eight. Execution is BF16 with FP32 master parameters, optimizer states, and averaged gradients. All three state-source variants share the randomized stage schedule, with 50 visits per stage: 24/13/15/15 in clean mode and 26/37/35/35 in robust mode.

Input and refresh protocol. Inputs pair OpenVid segment first frames with their existing overall scene-description captions; training and validation are separated by source parent. From the 3,840-item pool, 33 refreshes draw eight new assignments each, covering 264 assignments and 1,064 robust sample visits. The 67 clean updates add 536 visits, with possible overlap in source identity. Each refresh supplies six updates, except the last, which supplies eight. The surrogate takes two projected sign-gradient steps of $4/255$ from a seeded integer-pixel random start within $\epsilon = 8/255$, applying PNG rounding and projection at each step. The full four-stage generator scores the initial and final candidates, selecting the earlier one in a tie. Noise is fixed within an attack and resampled for each training update. Evaluation instead uses full four-stage PGD20 or PGD40 with step $1/255$ and scores every candidate.

Single-stage surrogate operator. The surrogate uses the current student snapshot θ_r. Let a be a candidate RGB input, E the VAE encoder, and u_b the fixed initial Gaussian noise for block b. Using the native clean-prediction query D and cache-writing operation W, its recurrence is

$$K_0 = W_{\theta_r}(E(a), 0, \varnothing; p),$$
$$\widetilde{z}_b = D_{\theta_r}(u_b, 1000, K_{b-1}; p), \qquad K_b = W_{\theta_r}(\widetilde{z}_b, 0, K_{b-1}; p).$$

The seven blocks contain three latents each; the unused final cache write is omitted. Source and history writes use noise level zero, while each query maps timestep 1000 directly to a clean prediction, omitting $937, 833, 625$ and their re-noising transitions. A 22-latent context retains the source and all preceding blocks. Each candidate regenerates this history, and input gradients traverse every block, cache write, and the encoder. The objective is $\mathrm{MSE}(\widetilde{Z}_{\theta_r}(a, p; u), \widetilde{Z}_{\theta_r}(c, p; u))$, with $\widetilde{Z}$ concatenating the seven predictions. The reference is the surrogate's own clean rollout. Candidate selection uses the full four-stage student and its native clean reference, with shared initial and fixed native transition noise. Proxy gradients therefore propose inputs, while native predictions select them.

Training and state replay. Each update selects the robust or clean branch, a shared denoising stage, and a block per sample, then generates fresh student states. Stages are randomly reshuffled in groups of four updates. Robust updates reuse the current attacked input; clean updates draw a normal input. The frozen teacher predicts at the selected query using the same raw generated history and clean source, rebuilding its own KV cache. Replay includes ancestors of retained cache values: source eviction removes direct reads but not source information inherited through history. The student matches this target with Eq. (4); gradients include the selected query, source cache write, and most recent history write, while older history-generation dependencies are detached. Gradients are averaged across eight workers, clipped, and applied with AdamW. This local parameter update preserves on-policy forward states without full-rollout parameter BPTT; attack differentiation retains the complete input-gradient paths. Checkpoints retain model, optimizer, RNG, and sampler states. Deployment uses only the student with native inference.

All perturbations modify one fixed source image. Seven-block prefix attacks use a 22-latent context without eviction, whereas long continuation uses a 21-latent sliding context; the first six clean blocks match, while block 7 can differ. Late-objective attacks use the same 27-block sliding-window operator for optimization and evaluation.

State-source controls. The clean-branch control obtains only clean-update states from the frozen original model on clean inputs; the robust-branch control obtains only robust-update states from that model on attacked inputs. The student predicts at the donor query and history with its delivered

image, and the teacher provides the target at the same state with the clean image. Each rebuilds its own KV cache. The other branch remains on-policy, and all arms refresh attacks against their trainable student. Table 6 uses 20,000 whole-source bootstrap resamples with seed 2026092601.

B CAUSAL ANALYSIS AND TEACHER-TARGET ALIGNMENT

We use the complete raw-state reconstruction and teacher/state-construction Lipschitz bounds of Proposition 1, extended to comparisons with the original clean execution. Let $G = (I - L)^{-1}$ denote the finite causal propagation operator. The analysis relates local prediction residuals to rollout deviations under these bounds; it does not assume that short training covers every long-rollout state.

B.1 PERTURBATION INJECTION AND DOWNSTREAM PROPAGATION

For nonnegative gain g and continuing aligned injection $\beta\delta$, a scalar recurrence accumulates $\beta\delta \sum_{k=0}^{N-1} g^k$: the response saturates for $g < 1$, grows linearly for $g = 1$, and grows geometrically for $g > 1$. This illustrates how generated history can sustain or amplify input differences. After the last direct source read at boundary e, linearized propagation is $\Delta h_B \simeq \Phi_{B:e+1} \Delta h_e$. Reducing Δh_e by factor α while keeping Φ fixed reduces squared response by α^2 at every later horizon. Long-range robustness can therefore improve by reducing injected discrepancies without changing the inference mechanism or its propagation gain.

B.2 TEACHER TRANSITIONS AND SUPERVISION TARGETS

At a fixed query and state, the kernels in Eq. (5) are translations of the same noise law. Coupling identical noise gives cost $\|\alpha_j(F_\theta - T)\|^2$; Jensen's inequality bounds every coupling below by the squared difference of its means, proving equality. The normalized prediction MSE is thus $W_2^2/(n_i\alpha_j^2)$: a local transition matching objective with stage-dependent scaling.

For an independent clean-trajectory target, let $\beta_i = T_i(\bar{s}_i, c) - T_i(s_i^\delta, c)$ and $t_i = \|u_i^\delta - T_i(\bar{s}_i, c)\|$, where $\bar{s}_i$ is the teacher's clean state at the corresponding stage and block. Its signed prediction error is $\eta_i^\delta - \beta_i$. By the triangle inequality, $r_i^\delta \le t_i + \|\beta_i\|$; writing $\gamma_i = \|\beta_i\|$ gives $e \preceq G(t + \gamma + r^0)$. Same-state guidance directly fits the residual in the propagation bound, whereas an independent trajectory introduces a target offset from a different query and history. This offset can also compensate for corrupted history, so its effect depends on the generated states. Indeed, with $v_i = T_i(\bar{s}_i, c)$, $t_i^\delta = \|u_i^\delta - v_i\|$, and $q_i^0 = \|u_i^0 - v_i\|$, the triangle inequality and Eq. (12) yield

$$e \preceq t^\delta + q^0 \preceq t^\delta + Gr^0. \tag{11}$$

An independent-trajectory target can therefore reduce response by directly anchoring predictions to the clean reference. These objectives differ in their treatment of context; Section C evaluates the resulting trade-off between latent response and generation detail.

B.3 NORMAL ANCHORING AND OVERALL DEVIATION

Let v_i denote the original model's independent clean execution, and $d_i^\zeta = \|u_i^\zeta - v_i\|$ its distance from execution $\zeta \in \{0, \delta\}$. Under the stated Lipschitz bounds,

$$d^\delta \preceq Gr^\delta, \qquad d^0 \preceq Gr^0, \qquad G = (I - L)^{-1}. \tag{12}$$

For the nonnegative final-block selection and normalization matrix $W_{\mathcal{B}}$,

$$Q_{\theta,\mathcal{B}} \le \|W_{\mathcal{B}} G r^\delta\|_2^2, \qquad N_{\theta,\mathcal{B}} \le \|W_{\mathcal{B}} G r^0\|_2^2. \tag{13}$$

The reference shares the source, prompt, and noise draws, using frozen original parameters and its own clean history. Inserting $T_i(s_i^\zeta, c)$ between u_i^ζ and v_i gives $d_i^\zeta \le r_i^\zeta + \sum_{k<i} L_{ik} d_k^\zeta$. A common componentwise bound on both branches and iteration of this triangular system prove Eq. (12). The row of $W_{\mathcal{B}}$ selecting final node $i = (4, b)$ has entry $1/\sqrt{|\mathcal{B}| n_b}$, all others zero, giving the normalization in Eq. (13).

Stack normalized final-block latents into vectors z^δ, z^0, v. Then $S = \|z^\delta - z^0\|^2$, $Q = \|z^\delta - v\|^2$, and $N = \|z^0 - v\|^2$ obey

$$(\sqrt{Q} - \sqrt{N})^2 \le S \le (\sqrt{Q} + \sqrt{N})^2. \tag{14}$$

Stacking sources with an additional $1/\sqrt{n}$ normalization gives the same inequalities for equal-source means. Identical perturbed and clean outputs can have $S = 0$ while $Q = N$ remains large, and shared residual bias can cancel in S. Normal updates consequently anchor original behavior through N, complementing rather than duplicating the robust objective.

C Supervision target and generation detail

The clean-trajectory control changes only the robust target: the teacher independently generates a clean prefix with the same source, prompt, initial noise, and transition noise, then supplies its pre-transition prediction at the sampled stage and block. The student retains its current attacked rollout and local gradient, and normal updates retain context-matched guidance. Both arms use 200 optimizer updates, the 133/67 robust/normal schedule, 33 own-model attack refreshes, and identical source/noise assignments; each follows its own evolving attacks and student states.

Trajectory targets lower own-attack response by 29.8%/24.5% and original-reference deviation by 13.8%/13.1% on the development/confirmation sets (Table 5); normal drift changes by 13.9%/-0.5%. On confirmation sources, late S is .8867 versus 1.2896 for context matching, and random S is .0810 versus .1546. The quality comparison below shows why these distance reductions alone do not determine target choice.

Across all three endpoints, the comparison uses 4,320 frames from 1,440 clean and own-attack videos over 240 sources, sampled at 2, 10, and 18 seconds. Every source exhibits softened subject detail or repeated edges with trajectory targets at later sampled times under both input conditions, although some stationary backgrounds remain sharp. Figure 5 shows the first source of each set, selected by index.

Technical image quality. MUSIQ (Ke et al., 2021) quantifies technical image quality using the SPAQ checkpoint and VBench imaging-quality preprocessing (Huang et al., 2024a): aspect-preserving resizing to a 512-pixel longer side, followed by averaging over the three frames per source and equally across sources. All endpoints share the metric, frame indices, and decoding procedure; clean and own-attack outputs are scored separately. This fixed-frame analysis complements the full VBench evaluation.

Trajectory targets lower normal-frame quality in all 240 sources: the mean is 57.34, versus 75.14 for context matching and 75.05 for the original model. The paired difference from context matching is -17.80 points (95% source-bootstrap interval $[-22.52, -13.31]$; 20,000 resamples), whereas context matching differs from the original by $+0.09$ points ($[-1.06, 1.36]$). Under own-model attacks, trajectory targets again score lower on every source, with a mean difference of -24.01 points ($[-28.87, -19.06]$). Their normal score falls from 69.30 at 2 seconds to 46.69 at 18 seconds, while context matching changes from 73.79 to 75.64. Thus the loss of generation detail intensifies during continuation despite lower latent response. Averaging incompatible continuations is one possible explanation for this effect; the measured result supports context-matched guidance as the better response–quality trade-off.

D Training controls and attack-strength evaluation

The controls distinguish adversarial exposure, normal preservation, and additional training within the same adaptation framework. Table 1 specifies their actual optimizer-update counts.

Attack exposure. The random-exposure arm replaces optimized inputs with their unoptimized random candidates, retaining source order, noise and stage/block schedules, fresh student states, teacher-target construction, and normal updates. Ours lowers own-PGD response by 25.79% and also reduces normal displacement, supporting optimized adversarial exposure over generic noise training.



Figure 5: **Teacher targets affect normal continuation.** Clean-input frames from the first development and confirmation sources show blurred or repeated subject edges with clean-trajectory targets. Images retain their original aspect ratio without enhancement.

Normal preservation. The loss ablation retains 133 robust updates and skips all 67 normal positions, including Adam updates. It matches robust source/noise/stage allocations rather than total optimizer updates; each arm refreshes its own attacks and states. Removing normal updates increases reference drift while slightly lowering own-clean response, establishing the normal branch's role in anchoring original behavior.

Training scale. The longer-training variant restores the 200-update model's parameters, Adam states, and per-rank random states, then continues another 200 updates with fresh source allocations and the same optimizer, local loss, and 2:1 mode cycle. This tests additional learning and data

exposure together. Early improvement but late degradation favors selecting training length by long-continuation performance.

Attack optimization. PGD40 resumes the actual twentieth iterate and selects the best of all candidates 0–40 under the same short-horizon objective. Long-rollout response increases by 27.48% for the original and 25.36% for Ours, while the relative benefit remains similar. Figure 6 shows that optimization remains active; the comparison tests a stronger finite-budget attack rather than assuming convergence.

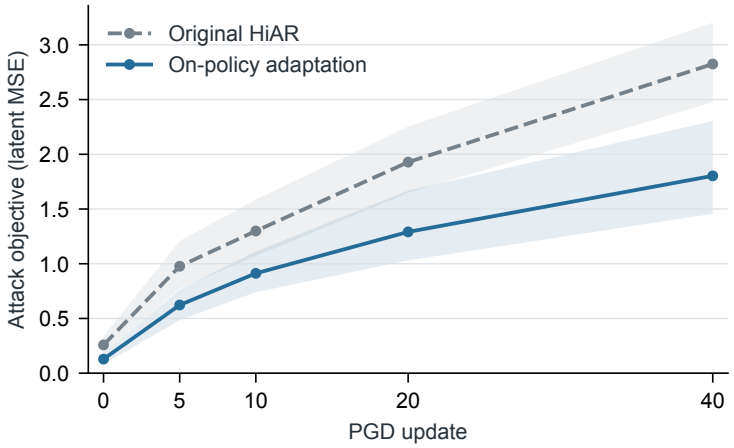


Figure 6: **Attack optimization remains active at larger budgets.** Eight-source subset. Mean actual-candidate objectives at updates 0, 5, 10, 20 and 40, with pointwise 95% source-bootstrap intervals; values are not best-so-far. Gradients traverse all four native stages over seven blocks.

E FULL-HISTORY LATE-SEGMENT ATTACKS

The first forty development sources receive separate own-model attacks against the original generator and Ours. The objective is own-clean latent MSE over blocks 21–27 of the deployed 27-block, four-stage generator with a 21-latent sliding window. Attack optimization and replay use this same operator, differentiating through the RGB encoder and every causal ancestor of the late objective. PGD takes ten updates of $1/255$ within $8/255$ from paired random starts, selecting the earliest best quantized PNG among all eleven candidates against each model's own clean reference.

Table 4 reports window-wise means in the main text; the table below shows the late window for four of the forty sources.

Table 7: Four example sources from the preselected forty-source late-attack subset, including the source whose late response increases. These are own-endpoint attacks, not a shared optimized input.

Source (fixed index)	Original late S	Ours late S	Reduction	Ours late Q
0	.6819	.4867	28.62%	.7719
1	1.5526	1.2222	21.28%	1.3950
2	.7366	.7565	−2.70%	.9558
3	1.5198	1.3733	9.64%	1.3728

Thirty of forty sources improve in both full-rollout and late S; source 2 increases by 11.88% and 2.70%, respectively. This extends the average benefit to attacks optimized directly for late outputs. Absolute scores differ in scope from the eighty-source PGD20/40 evaluation because the source subset, target horizon, and iteration count differ.

F FROZEN-SOURCE CONFIRMATION

The 160 confirmation sources were sampled from the filtered OpenVid validation pool before their Ours outputs were evaluated. They are disjoint from the 3,840 training items and eighty development sources by record, parent, and first-frame identity. Each model receives its own PGD20 attack; clean/random inputs and a single fixed generation-noise realization are paired.

Figure 1 instead uses sixteen VBench-I2V images, with shared random perturbations and generation noise across Self-Forcing, HiAR, and Ours. This cohort has no identical clean-input PNGs in the OpenVid confirmation set and is a subset of the full VBench-I2V suite in Table 2.

Window-wise S, Q and N appear in Table 4. The corresponding Q reductions are 25.97%, 19.55%, 15.68% and 20.06% for blocks 1–7, 8–20, 21–27 and 1–27, respectively.

Ten of the 160 sources had appeared in a separate training-monitoring run, but not in training or development. Alongside the prespecified set, the following sensitivity analysis excludes those ten without replacement or further model selection.

Table 8: Confirmation after excluding previously monitored sources. The remaining 150 sources retain reductions in both response and reference deviation.

Window	Original S	Ours S	S reduction	Improved	Q reduction
Full rollout	1.6274	1.2840	21.10%	140/150	18.70%
Late	1.6537	1.3361	19.20%	120/150	14.17%

With 20,000 whole-source bootstrap resamples, the S-difference intervals remain negative: $[-.4635, -.2243]$ overall and $[-.4538, -.1730]$ late. These intervals characterize source-level variation within the same domain.

For the full 160-source set, random-S falls from .4564 to .1546 while random-Q increases to .5411, distinguishing reduced sensitivity from preservation of the original trajectory.

G QUALITATIVE OBSERVATIONS

Figure 7 illustrates reduced attack-induced structural changes alongside residual normal-generation drift. Visual review covers all 325 decoded-frame thumbnails in each of the 320 clean development videos and fixed 4/10/19.5-second frames from all 960 confirmation videos. Own-attack comparisons use each model's separately optimized PNG.

H NATIVE-SYSTEM SENSITIVITY

On the sixteen paired random-input sources, Self-Forcing and HiAR have late means of $.5587/.3304$ (blocks 21–27). The paired difference is $-.2283$, with a 95% percentile interval $[-.3641, -.0795]$ from 20,000 whole-source bootstrap resamples. Figure 8 shows every source, including the three without lower late response under HiAR. Both systems use their native weights and inference schedules. Under these same inputs and generation noise, adapted HiAR reduces the full-video mean from .2508 to .1004 and the late mean from .3304 to .1519: relative reductions of 60.0% and 54.0%. All sixteen sources improve in both ranges, with paired-source 95% intervals $[-.2136, -.0935]$ and $[-.2503, -.1132]$, respectively. These are shared random-input comparisons, distinct from own-model attacks.

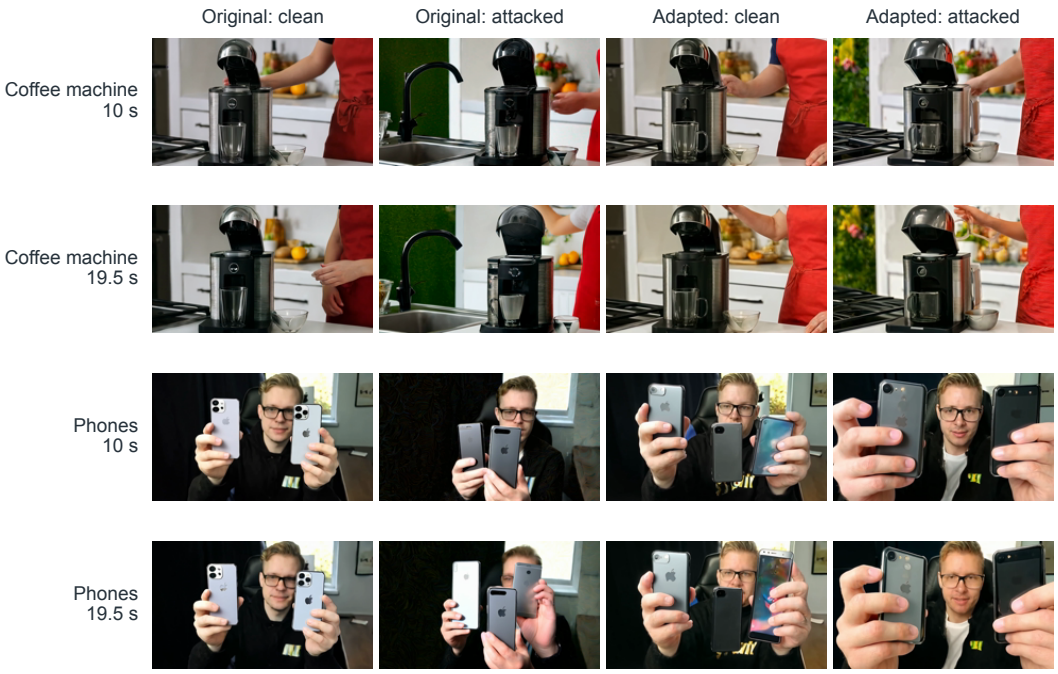


Figure 7: **Perturbation attenuation and normal-generation drift.** Clean and own-PGD20 frames at 10 and 19.5 seconds; "Adapted" is Ours. Adaptation avoids the attacked baseline's added faucet but retains some background changes; the phone example shows overlapping surfaces in the adapted clean output. Frames are shown without retouching.

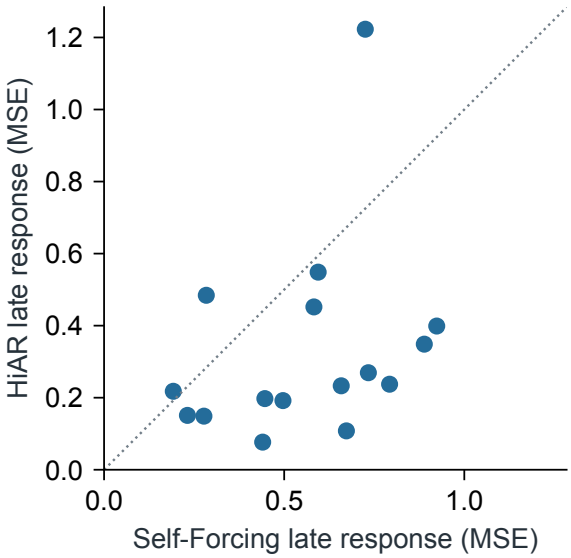


Figure 8: **Native-system sensitivity across sources.** Each point compares mean response over blocks 21–27 for one of sixteen sources under the same random $8/255$ perturbation. Points below the equality line favor HiAR (13/16 sources). Both systems use their native weights and schedules.