BADWORLD: ADVERSARIAL ATTACKS ON WORLD MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Visual world models (VWMs) generate interactive futures from an initial observation and a sequence of user actions. As these models are increasingly used as simulators, an important question is whether their predicted dynamics remain stable when the initial observation is slightly perturbed. To answer this question, we introduce BADWORLD, a label-free attack for autoregressive VWMs. Rather than optimizing over full future rollouts, BADWORLD directly perturbs the model's early denoising velocity. The resulting errors enter the autoregressive history and compound over subsequent generations. We further introduce trajectory-adaptive optimization, which mines hard control sequences to improve the attack across different user actions. We evaluate BADWORLD on VWMs with both continuous and discrete controls. Across models, visually small perturbations lead to incomplete denoising, structural degradation, and reduced control consistency. On an autonomous-driving world model, perturbations optimized only through the video-generation stream also substantially increase trajectory prediction error. These results show that coherent rollouts under clean inputs do not necessarily imply robust world dynamics, revealing an important weakness of current autoregressive world models.

1 INTRODUCTION

Visual world models are moving from passive video generators to interactive simulators. Given a single context image and a sequence of user-defined actions, they can synthesize action-conditioned future videos (Hafner et al., 2019; 2023). This ability makes them useful for interactive games (Bruce et al., 2024; He et al., 2025; Mao et al., 2025; Wang et al., 2026b), robotics (Li et al., 2025; Brohan et al., 2022; Bar et al., 2025; O'Neill et al., 2024), and autonomous driving (Caesar et al., 2020; Feng et al., 2025; Hu et al., 2023; 2026; Wang et al., 2024b; Zhang et al., 2025). As these models generate increasingly coherent rollouts, a common belief has emerged: visual world models may have implicitly learned physical and geometric rules of the world (Meng et al., 2024; Motamed et al., 2026; Liu et al., 2024a; Kang et al., 2024).

This progress raises a fundamental robustness question: *are the learned dynamics of current world models stable under small input perturbations?* This question is especially important for safety-critical applications, where fragile rollouts may undermine simulation, prediction, or planning. In this work, we use adversarial perturbations as a stress test for this form of temporal robustness.

However, existing adversarial attacks on generative models do not directly fit world models. Prior work mainly targets text-to-image personalization (Xu et al., 2024; Choi et al., 2025; Liu et al., 2024b; Van Le et al., 2023; Liang et al., 2023; Wang et al., 2024a; Shan et al., 2023), image-to-image editing (Salman et al., 2023; Wang et al., 2025a; Shen et al., 2025), and conventional image-to-video or video generation pipelines (Wang et al., 2025b; Li et al., 2024; Xu et al., 2026; Gui et al., 2025; Vu et al., 2026; Chowdhury et al., 2025; Long et al., 2026). These attacks are usually designed for fixed generation conditions, reference outputs, or localized spatial-temporal degradation. In contrast, autoregressive world models are interactive, history-dependent, and control-conditioned (Teng et al., 2025; Zhu et al., 2026b; He et al., 2025; Bruce et al., 2024; Wang et al., 2026b; Mao et al., 2025), which leads to two central challenges.

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.

- **C1: Missing future supervision.** A world-model adversary only observes a single context image, without paired future videos, ground-truth trajectories, or predefined correct rollouts. Therefore, reference-based adversarial losses are not directly applicable.
- **C2: Unknown future controls.** World-model rollouts depend on future camera paths, navigation commands, or discrete user actions. An attack optimized for one fixed trajectory may fail when the control signal changes.

To address these challenges, we propose **BadWorld**, a label-free adversarial framework for autoregressive world models. Given a pretrained world model and a single clean context image, BadWorld learns an imperceptible perturbation that drives future rollouts into out-of-distribution behaviors, without requiring paired future videos or knowledge of the user's future actions. Specifically, BadWorld consists of two technical components.

- **S1: Self-supervised velocity attack.** To address **C1**, we attack the model's predicted velocity space instead of comparing outputs with unavailable ground truth. We use the model's own denoising dynamics as supervision, together with an early-denoising approximation and a simple context-based history proxy. This yields a label-free attack that requires neither future videos nor action annotations.
- **S2: Trajectory-adaptive optimization.** To address **C2**, we formulate the attack as a bi-level optimization problem. The outer loop searches for hard trajectories under which the current perturbation is least effective, while the inner loop updates the perturbation against these trajectories. This produces a more control-agnostic adversarial image that is harder to bypass by simply changing the action sequence.

We evaluate BadWorld on representative autoregressive world models with both continuous camera control and discrete action control. Our results show that current world models are surprisingly fragile: adversarial images remain visually close to clean inputs, yet the generated rollouts can suffer from incomplete denoising, structural collapse, semantic drift, and loss of control consistency. Across models and metrics, Velocity-Min consistently produces strong degradation, while trajectory-adaptive optimization further improves robustness across difficult trajectories.

Crucially, we show that such attacks can be easily extended to trajectory planning in autonomous driving. The resulting trajectories can be severely unsafe. This reveals a critical robustness risk for deploying world models in safety-critical systems.

Our contributions are summarized as follows: (1) We formalize adversarial attacks on autoregressive world models; (2) We propose BadWorld, a label-free adversarial framework that leverages predicted velocity to overcome missing future supervision, together with trajectory-adaptive optimization to handle diverse future controls; (3) We show that representative interactive world models are highly fragile; imperceptible perturbations severely corrupt visual rollouts and disrupt autonomous driving trajectory planning, with profound implications for both safety and privacy.

2 Related Work

Controllable Visual World Models Recent advancements have evolved video generation from passive synthesis into interactive Visual World Models (VWMs) (Blattmann et al., 2023; Wan et al., 2025; Polyak et al., 2024). Unlike standard generators, autoregressive VWMs internalize physical dynamics and simulate futures conditioned on temporal history and user inputs, ranging from continuous camera trajectories (Zhu et al., 2026b) to discrete action tokens (He et al., 2025). Ensuring the control adherence and stability of these models is crucial for their safe deployment in interactive and safety-critical domains.

Adversarial Attacks on Generative Models Adversarial attacks have transitioned from static image models (Liang et al., 2023; Salman et al., 2023) to video generation, exposing vulnerabilities like backdoor triggers (Wang et al., 2025b) and drag-based manipulation (Xu et al., 2026). While recent defenses evaluate standard image-to-video (I2V) pipelines (Gui et al., 2025; Chowdhury et al., 2025), the robustness of autoregressive VWMs remains a largely unexplored. Existing methods address condition-level or search-level attacks (Guo et al., 2026a;b), but none explore image-space attacks against interactive, history-dependent VWMs. To address this gap, we introduce the first adversarial framework tailored to the unique autoregressive architecture of video world models. A more comprehensive discussion of related work is provided in Appendix A.1.

3 BACKGROUND

This section provides the necessary background for studying the robustness of video world models. We first review autoregressive video generation and its flow-matching formulation, which together establish the foundation for defining our adversarial objective.

3.1 PRELIMINARY: AUTOREGRESSIVE VIDEO GENERATION FOR WORLD MODELS

In the study, we focus on the autoregressive (AR) model for world generation. Unlike bidirectional generation, the AR approach decomposes a video sequence into K segments (chunks) $\{z_1, \ldots, z_K\}$. Given context frame $x \in \mathbb{R}^{H\times W\times 3}$, the model factorizes the joint distribution:

$$p(z_{1:K}|x, \tau_{1:K}) = \prod_{i=1}^{K} p(z_i|z_{<i}, x, \tau_i) \tag{1}$$

The control τ_i typically encodes camera motions or actions. Each segment z_i is generated conditioned on context x, history $z_{<i}$ (via sliding window), control τ_i, and an optional prompt p.

Most recent AR world models instantiate each chunk generator with diffusion or flow-matching (Lipman et al., 2022) dynamics. For a target chunk z_i, Flow Matching constructs an interpolated latent between the data and Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$, as that $z_t^i = (1-t)z_i + t\epsilon$. The velocity network v_θ is trained to predict the transport direction $v^* = \frac{dz_t^i}{dt} = \epsilon - z_i$ by minimizing

$$\mathcal{L}(\theta) = \mathbb{E}_{i,t,\epsilon}\left[\left\|v_\theta(z_t^i, t \mid z_{<i}, x, \tau_i) - v^*\right\|_2^2\right]. \tag{2}$$

During inference, z_i is generated via sequential self-rollout. While capturing temporal dependencies, this AR structure is sensitive to error accumulation, amplifying small perturbations over the rollout.

3.2 PROBLEM SETUP AND CHALLENGES

Given a pretrained autoregressive world model $\mathcal{G}_\theta$, we aim to construct an adversarial context image

$$x_{\text{adv}} = x + \delta, \qquad \|\delta\|_\infty \le \eta, \tag{3}$$

such that the generated rollout videos under the same control sequence becomes substantially different from the clean rollout. The ideal rollout-level objective is

$$\delta^* = \arg\max_{\|\delta\|_\infty \le \epsilon} \mathcal{D}(\mathcal{G}_\theta(x+\delta, \tau_{1:K}), \mathcal{G}_\theta(x, \tau_{1:K})), \tag{4}$$

where $\mathcal{G}\theta(\cdot, \tau_{1:K})$ is the video generated under controls $\tau_{1:K}$, and $\mathcal{D}$ measures the difference between the adversarial and clean videos. Intuitively, the attack finds a small perturbation δ within budget η that maximizes this difference. A large difference indicates OOD behavior, such as visual artifacts, temporal inconsistency, motion collapse, or semantic drift. However, Eq. 4 presents two challenges.

C1: Missing future supervision. The first challenge is that the attacker only observes the context image x and has no future ground-truth video. However, optimizing Eq. 4 requires costly rollout generation and backpropagation. The attack must therefore be *label-free*.

C2: Unknown future controls. A second challenge comes from the interactive nature of world models. Rollouts heavily depend on user controls (e.g., camera paths or actions) which are unknown at attack time. A perturbation optimized for one fixed trajectory may therefore fail when the control sequence changes. A robust attack should remain effective across diverse future controls.

These two challenges guide the design of our method. To address C1, we replace the costly rollout-level supervision with a self-supervised velocity-level attack objective that perturbs the model's denoising dynamics. To address C2, we introduce trajectory-adaptive optimization that actively searches for hard control trajectories and optimizes the perturbation against them.

4 METHODOLOGY

Following the two challenges identified, BADWORLD consists of two main components. For C1, we design a label-free velocity attack using the model's denoising dynamics as supervision. For C2, we introduce trajectory-adaptive rollout optimization, which finds difficult control trajectories and optimizes perturbations across different control signals.

4.1 LABEL-FREE VELOCITY ATTACK OBJECTIVE

To address missing future supervision (**C1**), we avoid optimizing the rollout discrepancy (Eq. 4), which lacks ground truth and requires prohibitive backpropagation through time. Instead, we attack

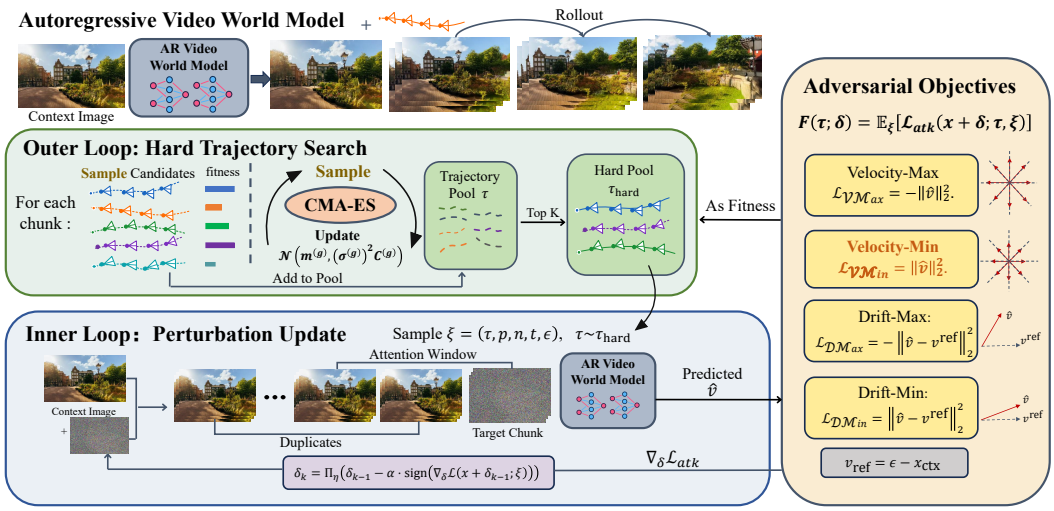

Figure 1: **BadWorld Pipeline.** An autoregressive video world model generates rollouts conditioned on context and camera (top). The outer loop employs CMA-ES for hard trajectory mining (middle), while the inner loop updates the adversarial perturbation under label-free settings (bottom). Four velocity-based objectives are provided (right), with Velocity-Min ($\mathcal{L}_{\text{VMin}}$) serving as the default.

the *velocity space directly*. Corrupting velocity disrupts local denoising, which inherently compounds across autoregressive steps.

Formally, we define a state tuple representing the generation context:

$$\xi = (\tau, p, n, t, \epsilon), \tag{5}$$

where τ is a control signal, p is an optional prompt, $n \in \{1, \ldots, K\}$ denotes the target chunk, t is a denoising timestep, and $\epsilon \sim \mathcal{N}(0, I)$ is Gaussian noise. Given the adversarial context $x + \delta$, the model predicts a velocity

$$\hat{v}_\delta = v_\theta(\epsilon, t \mid h_n^\delta, \tau, p), \tag{6}$$

where h_n^δ denotes the history condition used for the n-th autoregressive chunk. Using this notation, we define the general attack objective to find the perturbation δ^* that minimizes an adversarial loss:

$$\delta^* = \arg \min_{\|\delta\|_\infty \le \eta} \mathbb{E}_\xi \left[\mathcal{L}_{\text{atk}}(x + \delta; \xi) \right]. \tag{7}$$

Here, $\mathcal{L}_{\text{atk}}(x + \delta; \xi)$ specifies the harmful velocity property induced by the adversarial context. In practice, we estimate the expectation by random sampling and update δ using projected gradient descent (PGD) (Madry et al., 2017). Specifically, we instantiate $\mathcal{L}_{\text{atk}}$ through two types of objectives.

Velocity magnitude objectives. The first type of objective attack the *magnitude* of predicted velocity:

$$\mathcal{L}_{\text{VMax}}(x + \delta; \xi) = -\|\hat{v}_\delta\|_2^2, \qquad \mathcal{L}_{\text{VMin}}(x + \delta; \xi) = \|\hat{v}_\delta\|_2^2. \tag{8}$$

By maximizing this magnitude, Velocity-Max ($\mathcal{L}_{\text{VMax}}$) forces aggressive denoising, which injects unnatural contrast and severe distortions into the chunk. Conversely, Velocity-Min ($\mathcal{L}_{\text{VMin}}$) suppresses the update norm, causing the model to retain initial noise and yield incomplete structures. Because of the autoregressive nature of the model, these localized denoising failures compound, leading to severe degradation over long-horizon rollouts.

Velocity direction objectives. The second type of objective attacks the *velocity direction*. We define a context-anchored reference velocity as $v^{\text{ref}} = \epsilon - x_{\text{ctx}}$, where x_{ctx} is the encoded original image. This reference represents *completely static video* that simply preserves the initial frame. Using this reference velocity, we define:

$$\mathcal{L}_{\text{DMax}}(x + \delta; \xi) = -\|\hat{v}_\delta - v^{\text{ref}}\|_2^2, \qquad \mathcal{L}_{\text{DMin}}(x + \delta; \xi) = \|\hat{v}_\delta - v^{\text{ref}}\|_2^2. \tag{9}$$

Intuitively, Drift-Max ($\mathcal{L}_{\text{DMax}}$) pushes the predicted velocity away from this static baseline video, encouraging semantic drift and temporal inconsistency. Conversely, Drift-Min ($\mathcal{L}_{\text{DMin}}$) pulls the velocity toward the static reference, suppressing scene evolution and inducing motion collapse.

Label-free realization. Evaluating these objectives requires querying the predicted velocity $\hat{v}_\delta$. In standard flow-matching, computing the interpolated latent $z_t^n = (1 - t)z_n + t\epsilon$ and the target velocity

$v^* = \epsilon - z_n$ both require the future ground-truth chunk z_n. Since z_n is unavailable in our label-free setting, we introduce two practical approximations.

Early-denoising approximation. When the timestep t approaches 1, the interpolated latent is overwhelmingly dominated by Gaussian noise:

$$z_t^n = (1-t)z_n + t\epsilon \approx \epsilon. \tag{10}$$

By restricting queries to early denoising timesteps, we can input ϵ directly, entirely bypassing z_n. Attacking these early stages remains highly effective: early disruptions alter the global structure, seeding fundamental errors that drastically amplify throughout the autoregressive rollout.

Context-based history proxy. Autoregressive generation also requires a history condition $z_{<n}$. Since the true rollout is unavailable, we approximate it by repeating the encoded adversarial context:

$$h_n^\delta = \mathrm{Repeat}(E(x+\delta), L_n), \tag{11}$$

where L_n is the required history length for the n-th chunk. This proxy provides a reliable pathway for the perturbation δ to influence the velocity prediction. Combined with the early-denoising approximation, this establishes the fully label-free velocity query (Eq. 6), allowing us to optimize the objectives (Eq. 8–Eq. 9) without ground-truth future videos or explicit annotations.

4.2 Trajectory-Adaptive Bi-Level Optimization

Although the self-supervised objectives in Sec. 4.1 address **C1**, interactive world models remain sensitive to control signals τ. Perturbations optimized for a single action sequence may overfit to its motion pattern and fail under unpredictable controls (**C2**). To ensure generalization across control sequences, we propose *trajectory-adaptive bi-level optimization*. Our key intuition is that an attack that remains effective against hard trajectories is more likely to generalize to diverse user controls.

Min-Max Formulation. Let $\tau = \{c_1, \ldots, c_T\}$ denote a continuous camera trajectory. We frame our trajectory-adaptive attack as a min-max optimization game:

$$\min_{\|\delta\|_\infty \le \eta} \max_{\tau \in \mathcal{T}} F(\tau;\delta), \qquad F(\tau;\delta) = \mathbb{E}_{\xi\sim\Omega(\tau)}\left[\mathcal{L}_{\mathrm{atk}}(x+\delta;\xi)\right], \tag{12}$$

where $\mathcal{T}$ is the feasible trajectory space and $\Omega(\tau)$ denotes the query state distribution for candidate τ. The inner minimization updates the perturbation δ under current hard trajectories, while the outer maximization identifies trajectories that resist the attack and yield the highest loss $F(\tau;\delta)$.

To parameterize the search space $\mathcal{T}$ efficiently, we discretize the continuous control at each frame into a compact vector preserving the primary degrees of freedom:

$$c_t = (\psi_t, f_t, s_t), \qquad \tau = [\psi_1, f_1, s_1, \ldots, \psi_T, f_T, s_T], \tag{13}$$

where ψ_t (yaw), f_t (forward translation), and s_t (lateral shift) define the viewpoint transformation.

Outer loop: hard trajectory mining. The outer maximization in Eq. 12 aims to discover "hard" trajectories that resist the perturbation. However, optimizing this sequence is non-trivial; the autoregressive generation process is largely non-differentiable with respect to the control signals.

To search without gradients, we employ the CMA-ES (Hansen, 2016). Unlike independent random sampling, CMA-ES adapts a multivariate Gaussian distribution to capture coherent camera motion patterns that effectively challenge the current perturbation.

Specifically, at each generation g, we maintain a search distribution over the trajectory vector τ:

$$q^{(g)}(\tau) = \mathcal{N}\Big(m^{(g)}, (\sigma^{(g)})^2 C^{(g)}\Big), \qquad \tau_k^{(g)} \sim q^{(g)}(\tau), \quad k = 1, \ldots, \lambda. \tag{14}$$

where $m^{(g)}$, $\sigma^{(g)}$, and $C^{(g)}$ denote the mean, step-size, and covariance matrix. Each sampled trajectory is projected onto $\mathcal{T}$ and scored by $F(\tau;\delta)$ (Eq. 12). CMA-ES then ranks the candidates and updates its distribution: The mean m update shifts the search toward harder trajectories, while the covariance C update increases the probability of sampling successful correlated directions. This is crucial for camera control, where hard cases often arise from coherent temporal motions rather than independent frame changes (details refer to Appendix).

After the final CMA generation, we collect evaluated candidates and retain the top-K trajectories:

$$\mathcal{P}_{\mathrm{hard}}^{(n)} = \mathrm{TopK}_{\tau\in\mathcal{S}^{(n)}} F(\tau;\delta), \tag{15}$$

where $\mathcal{S}^{(n)}$ is the set of trajectories evaluated for chunk n. We maintain a separate pool $\mathcal{P}_{\mathrm{hard}}^{(n)}$ for each chunk because trajectory difficulty varies across autoregressive steps.

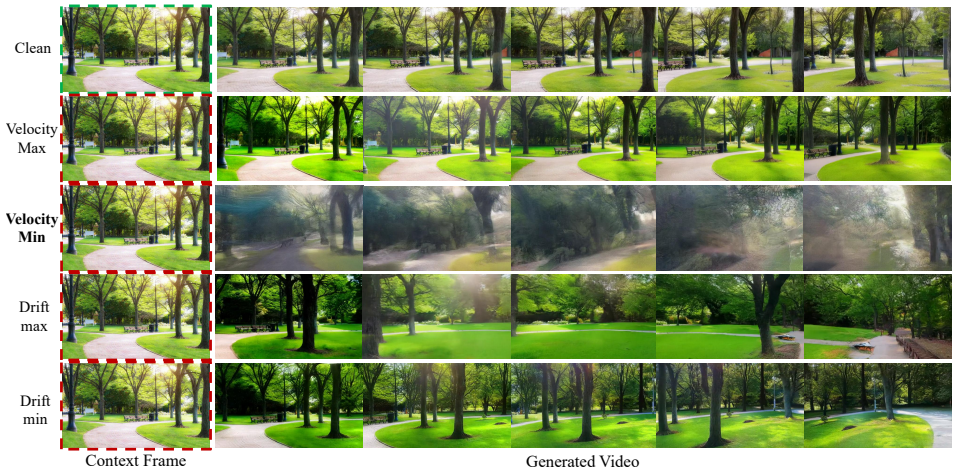

Figure 2: Qualitative comparison on Astra. Velocity-Max and Drift objectives induce color distortions. Velocity-Min causes severe geometric collapse.

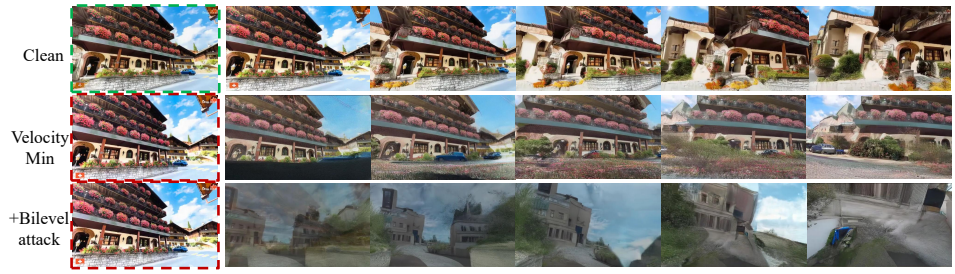

Figure 3: Bi-level Attack performance.

Inner loop: perturbation update. Given the hard trajectories, we update the adversarial perturbation. At each step, for a selected chunk n, we sample a trajectory τ from $\mathcal{P}_{\text{hard}}^{(n)}$ and construct the query state ξ_τ. The perturbation δ is updated via Projected Gradient Descent (PGD):

$$\delta_{k+1} = \Pi_\eta \left(\delta_k - \alpha \operatorname{sign} \left(\nabla_\delta \mathcal{L}_{\text{atk}}(x + \delta_k; \xi_\tau) \right) \right). \tag{16}$$

We alternate between hard-trajectory mining and perturbation updates to produce an adversarial context image that induces unstable rollouts across diverse controls.

5 Experiments

5.1 Experimental Setup

Models. We evaluate on two representative open-source autoregressive video world models:

- *Astra* (Zhu et al., 2026b): Built on Wan2.1 (Wan et al., 2025), Astra is fine-tuned for causal denoising and continuous camera pose conditioning, making it ideal for studying camera-driven adversarial behaviors.
- *Matrix-Game 2.0* (He et al., 2025): A lightweight, causal few-step generative model distilled from SkyReelsV2-I2V-1.3B (Chen et al., 2025). Utilizing discrete action signals, it provides a complementary platform to test robustness under categorical control.

Dataset Setup. Since no established benchmark exists for adversarial attacks on video world models, we construct two 100-image datasets tailored for each model. For Astra, we sample the first frame from natural landscape videos in the SpatialVID dataset (Wang et al., 2026a) and resized to 832×480. For Matrix-Game 2.0, we extract *Grand Theft Auto (GTA)* gameplay frames and resize to 640×352. To align with the training domain of Matrix-Game-2.0 and enhance clean generation quality, we apply Flux Kontext (Labs et al., 2025) to standardize primary object appearances. Specifically, vehicles are transformed into black sedans, and human subjects are edited to wear black clothing.

Baselines. To the best of our knowledge, no prior work has studied adversarial attacks on autoregressive video generation for world models. Following I2VGuard (Gui et al., 2025), we therefore adopt a noise baseline by adding random noise to the context frame, matching the mean and variance of the perturbations produced by our method.

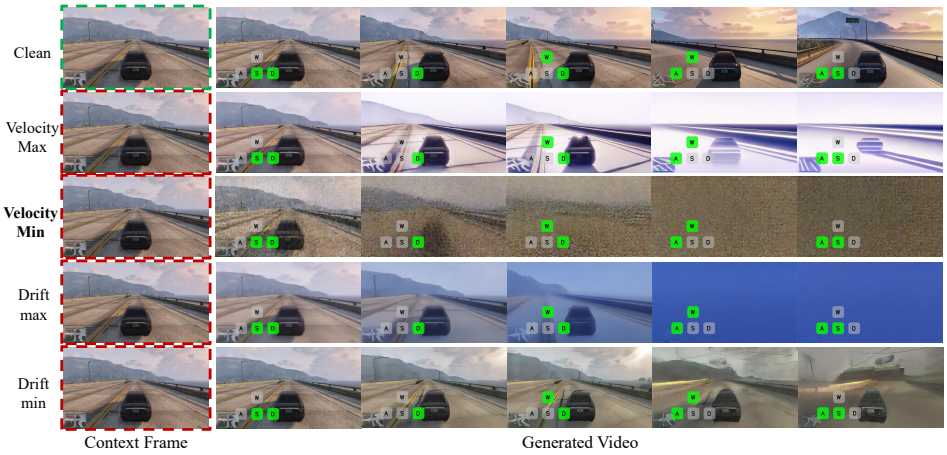

Figure 4: Qualitative comparison on Matrix-Game-2.0. Velocity-Max and Drift-Max inflate visual contrast. Drift-Min produces near-static videos. Velocity-Min causes complete structural collapse.

Evaluation Metrics. To assess adversarial impact, we employ VBench(Huang et al., 2024), VBench++(Huang et al., 2025), CLIP (Hessel et al., 2021) and MEt3R(Asim et al., 2025) to evaluate four dimensions. First, we measure **contextual consistency** by accessing *i2v subject* and *i2v background*, which quantifies how well the generated video preserves identity and environmental details from the source image. We exclude i2v-subject metrics for Astra, as the dataset lacks explicit foreground subjects. Second, we evaluate **video quality** by reporting *background consistency*, *aesthetic quality*, and *imaging quality* to capture visual degradation. Third, we evaluate **semantic preservation** by computing the *CLIP-I* (Hessel et al., 2021) score between the first and last frames. Fourth, we measure **geometry reliance** using *MEt3R* (Asim et al., 2025), which evaluates geometric consistency by reconstructing dense 3D geometry for each frame pair and measuring feature mismatch after cross-view projection.

Implementation Details. We execute all attacks on an A800 80GB GPU, simulating realistic label-free scenarios by restricting optimization to early denoising timesteps and duplicating context frames. During inference, we autoregressively generate three diverse videos per input using default model configurations. Comprehensive implementation details regarding model-specific conditionings and training hyperparameters are deferred to the Appendix.

5.2 Main Results on Different Attack Objectives

To compare the adversarial objectives introduced in Sec. 4.1, we attack 100 images per model under a noise budget of $\eta = 0.05$, evaluating each image across three distinct camera or action sequences.

Comparison on Astra. As shown in Tab. 1 (best results are bolded, and second-best results are underlined), *Velocity-Min* is the most effective objective, achieving the lowest scores across nearly all metrics. Compared to the clean baseline, it induces the sharpest decline in visual fidelity, reducing aesthetic quality from 0.501 to 0.405 and imaging quality from 0.690 to 0.513. This trend is further evidenced by MEt3R, where *Velocity-Min* reaches 0.264, representing a 74.8% increase over the clean baseline. *Drift-Max* follows as a strong runner-up, notably reducing imaging quality to 0.564. Conversely, *Velocity-Max* causes only moderate degradation, while *Random Noise* and *Drift-Min* have negligible impact, closely mirroring the clean baseline. For qualitative comparison, Fig.2 shows that *Velocity-Min* triggers incomplete denoising and grayish artifacts, whereas other objectives primarily alter color temperature.

Comparison on Matrix-Game-2.0. As detailed in Tab. 1, Matrix-Game-2.0 exhibits a broader vulnerability, where most objectives serve as effective attacks. *Velocity-Min* remains the most potent, significantly reducing background consistency to 0.960, imaging quality to 0.442, and achieving the highest MEt3R of 0.204. Other strategies like *Drift-Max* and *Velocity-Max* also demonstrate clear effectiveness by degrading imaging quality and CLIP-I. As visualized in Fig. 4, *Velocity-Min* induces severe structural disintegration and noise, while *Drift-Max* and *Velocity-Max* primarily generate frames with unnaturally high color saturation. *Drift-Min* tends to produce near-static videos that fail to capture the intended motion.

Method	I2V Background ↓	Background Consistency ↓	Aesthetic Quality ↓	Imaging Quality ↓	CLIP-I↓	MEt3R↑
Clean	0.978	0.915	0.501	0.690	0.813	0.151
Random Noise	0.973	0.912	0.490	0.680	0.810	0.150
Self-supervised Attack Objectives						
Velocity-Max	0.956	0.891	0.493	0.657	0.779	0.154
Velocity-Min	**0.930**	**0.845**	**0.405**	**0.513**	**0.714**	**0.264**
Drift-Max	0.938	0.866	0.466	0.564	0.750	0.207
Drift-Min	0.968	0.908	0.475	0.601	0.800	0.166
Trajectory Adaptive Bi-level Attack						
VMin+Bilevel	0.925	0.842	0.396	0.524	0.712	0.265

Table 1: Comparison of attack objectives against Astra.

Method	I2V Subject ↓	I2V Background ↓	Background Consistency ↓	Aesthetic Quality ↓	Imaging Quality ↓	CLIP-I ↓	MEt3R ↑
Clean	0.897	0.897	0.973	0.512	0.648	0.796	0.098
Random Noise	0.878	0.895	0.974	0.522	0.647	0.785	0.101
Self-supervised Attack Objectives							
Velocity-Max	0.788	0.795	0.964	0.443	0.491	0.631	0.119
Velocity-Min	**0.708**	0.750	**0.960**	0.352	0.442	0.584	**0.204**
Drift-Max	0.725	**0.743**	0.965	**0.340**	**0.364**	**0.537**	0.159
Drift-Min	0.868	0.868	0.964	0.478	0.623	0.631	0.101

Table 2: Comparison of attack objectives against Matrix-Game-2.0 (GTA).

Summary. In summary, while Astra shows high selectivity with *Velocity-Min* being the only truly effective objective, Matrix-Game-2.0 is susceptible to a wider range of perturbations. Across both benchmarks, *Velocity-Min* consistently remains the strongest objective for degrading video quality.

We therefore adopt the strongest objective, *Velocity-Min*, for all subsequent experiments. Furthermore, since baseline attacks on Matrix-Game-2.0 often trigger a near-complete collapse that masks performance nuances, we utilize the more challenging Astra benchmark to conduct ablation studies (5.3) and validate the efficacy of our bi-level optimization framework (5.4).

5.3 Ablation Studies

Following the experimental setup described in Sec. 5.1, we conduct three ablation studies on Astra (Zhu et al., 2026b) to evaluate the impact of different budgets and the effectiveness of our designs.

Ablation on attack budget. We first examine the impact of the perturbation budget η on attack performance. As shown in Tab. 3, increasing η from 0.03 to 0.10 leads to a consistent decline in video quality metrics and a corresponding rise in MEt3R scores, indicating more effective disruption. However, larger budgets also make the adversarial perturbations more visually perceptible. To achieve a balance between attack potency and imperceptibility, we select $\eta = 0.05$ as our default setting.

Ablation on diffusion timesteps. We assess timestep selection by comparing our *early-denoising* strategy (Sec.4.1) against uniform timestep sampling from 0 to 1000, denoted as "-Timesteps Selection" in Tab. 4. Without this early focus, attack efficacy collapses: Aesthetic Quality rises to 0.514 and MEt3R drops to 0.167. Targeting initial structural formation is thus essential to effectively neutralize generative priors.

Ablation on history simulation. To verify the effectiveness of our *context-based history proxy* (Sec. 4.1), we compare it against a *self-rollout* approach, denoted as "+ Self-Rollout" in Tab. 4. Since performing rollouts at every attack step is computationally prohibitive, we use an interval-based scheme: every 50 steps, we generate four videos from current adversarial context under distinct trajectories to update a history pool. During the subsequent interval, we sample from this pool as the history reference. This sophisticated approach yields negligible improvements, with metrics like MEt3R and CLIP-I showing slight improvement. Given the heavy rollout overhead, duplicating context frames is a more efficient and effective strategy.

5.4 Performance of Trajectory-Adaptive Bi-level Optimization

Following the baseline protocol, we evaluate the bi-level attack on 100 Astra images across 3 trajectories each. As shown in Tab. 1, trajectory-adaptive optimization improves most metrics over Velocity-Min, though gains are moderate on the full benchmark. This is expected, as Velocity-Min is already highly effective for most samples, leaving limited room for further degradation.

Budget	Step Size	I2V Back. ↓	Back. Consis. ↓	Aesthetic ↓	Imaging Qual. ↓	CLIP-I ↓	MEt3R ↑
$\eta = 0.03$	0.003	0.945	0.860	0.429	0.563	0.710	0.272
$\eta = 0.05$*	0.005	0.930	0.845	0.405	0.513	0.714	0.264
$\eta = 0.10$	0.005	**0.903**	**0.794**	**0.387**	**0.507**	**0.678**	**0.291**

Table 3: Performance under different attack budgets. * indicates the default setting.

Method	I2V Back. ↓	Back. Consis. ↓	Aesthetic ↓	Imaging Qual. ↓	CLIP-I ↓	MEt3R ↑
Velocity-Min	**0.930**	0.845	**0.405**	**0.513**	0.714	0.264
+Self-Rollout	0.932	**0.843**	0.410	0.514	**0.701**	**0.271**
-Timesteps Selection	0.976	0.913	0.514	0.715	0.801	0.167

Table 4: Ablation on label-free designs.

Method	I2V Back. ↓	Back. Consis. ↓	Aesthetic ↓	Imaging Qual. ↓	CLIP-I↓	MEt3R↑
Velocity-Min	0.935	0.845	0.440	**0.556**	0.720	0.249
+ Bi-level Attack	**0.931**	**0.835**	**0.425**	**0.556**	**0.711**	**0.256**

Table 5: Hard Sample Performance Comparison.

To assess robustness in challenging cases, we target the 10 most resilient samples under Velocity-Min, evaluating each against 7 distinct trajectories (Tab. 5). Our method improves five of six metrics, reducing Aesthetic Quality from 0.440 to 0.425. As shown in Fig. 3, bi-level optimization induces stronger noise and geometric distortion where the baseline struggles. These results show larger gains on resilient samples, averaged over a broader set of seven camera trajectories.

5.5 Real World Safety Risks

To assess practical safety risks beyond visual degradation, we evaluate if BadWorld disrupts downstream trajectory planning in autonomous-driving world models.

Experimental Setup. We attack Epona (Zhang et al., 2025), an autoregressive world model with separate video and trajectory diffusion transformers, across 20 driving clips. Given a 10-frame context, we perturb only the final input frame. Crucially, we optimize this perturbation solely via the video-generation stream, without any trajectory supervision or planning-stream access.

For video generation, we report FVD_{10} and FID against the ground-truth future video. For trajectory planning, we measure the L2 errors (at 1, 2, 3s, and average) against the ground-truth trajectory. We additionally compare the clean and adversarial outputs directly using video LPIPS, trajectory ΔADE, and the distance between their final predicted positions.

	Video vs. GT		Trajectory Error vs. GT (m)			
Input	FVD_{10} ↑	FID ↑	L2@1s↑	L2@2s↑	L2@3s↑	Avg. L2↑
Clean	786.87	42.19	0.41	1.02	1.68	1.03
Adversarial	**3710.91**	**352.67**	**1.19**	**2.90**	**4.94**	**3.01**

Table 6: Video and trajectory errors against ground truth on Epona. Higher values indicate stronger attack.

Comparison	LPIPS	ΔADE (m)	Final Sep. @10s (m)
Clean–Adv.	0.601	1.11	**24.47**

Table 7: Clean–adversarial output discrepancy on Epona.

Results. BadWorld severely degrades both both visual prediction and trajectory planning (Tab. 6). Against the ground truth, FID spikes from 42.19 to 352.67, and the average trajectory L2 error (at 1, 2, and 3s) nearly triples from 1.03 m to 3.01 m. Directly comparing clean and adversarial outputs (Tab. 7) yields a video LPIPS of 0.601, a trajectory ΔADE of 1.11 m, and a massive 24.47 m divergence in 10s final predicted positions. These results reveal a serious safety risk: ***visual corruption can propagate to trajectory planning, leading to severely unsafe driving decisions.***

6 Conclusion

In this work, we present BadWorld, an adversarial framework designed for autoregressive visual world models (VWMs). To bypass ground-truth dependencies, our self-supervised velocity attack directly disrupts early denoising dynamics. To handle unpredictable future controls, our trajectory-adaptive bi-level optimization mines hard sequences to forge control-agnostic perturbations. Crucially, this corruption disrupts real-world trajectory planning, exposing severe safety risks for VWM deployment while offering a practical mechanism for privacy protection.

AI USE STATEMENT

Generative AI tools supported manuscript review, experimental planning, code development, and language editing. The authors conducted the experiments, verified the code, results, text, and references, and take full responsibility for the final manuscript.

ETHICS STATEMENT

This work studies world-model robustness through controlled, offline adversarial evaluation. Its purpose is to identify vulnerabilities and inform the development of more reliable models. The driving experiments use recorded data rather than interventions on deployed vehicles. We consider disclosure beneficial for robustness research and defensive evaluation.

REPRODUCIBILITY STATEMENT

Sections 4.1 and 4.2 describe the attack objectives and optimization procedures. Sections 5.1 and 5.5 specify the generation and driving evaluation protocols. Appendices A.5.1, A.5.2, and A.5.3 detail trajectory constraints, sampling, and bi-level optimization. Appendix A.6 provides attack hyperparameters, inference settings, prompt construction, hardware, and runtime details.

REFERENCES

Mohammad Asim, Christopher Wewer, Thomas Wimmer, Bernt Schiele, and Jan Eric Lenssen. Met3r: Measuring multi-view consistency in generated images. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6034–6044. IEEE, 2025.

Jianhong Bai, Menghan Xia, Xiao Fu, Xintao Wang, Lianrui Mu, Jinwen Cao, Zuozhu Liu, Haoji Hu, Xiang Bai, Pengfei Wan, et al. Recammaster: Camera-controlled generative rendering from a single video. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14834–14844. IEEE, 2025.

Amir Bar, Gaoyue Zhou, Danny Tran, Trevor Darrell, and Yann LeCun. Navigation world models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15791–15801. IEEE, 2025.

Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Guanghui Liu, Amit Raj, et al. Lumiere: A space-time diffusion model for video generation. In *SIGGRAPH Asia 2024 conference papers*, pp. 1–11, 2024.

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.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022.

Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In *Forty-first international conference on machine learning*, 2024.

Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 11618–11628. IEEE, 2020.

Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. *Advances in Neural Information Processing Systems*, 37:24081–24125, 2024.

Guibin Chen, Dixuan Lin, Jiangping Yang, Chunze Lin, Junchen Zhu, Mingyuan Fan, Hao Zhang, Sheng Chen, Zheng Chen, Chengcheng Ma, et al. Skyreels-v2: Infinite-length film generative model. *arXiv preprint arXiv:2504.13074*, 2025.

William June Suk Choi, Kyungmin Lee, Jongheon Jeong, Saining Xie, Jinwoo Shin, and Kimin Lee. Diffusionguard: A robust defense against malicious diffusion-based image editing. In *International Conference on Learning Representations*, volume 2025, pp. 27134–27180, 2025.

Rohit Chowdhury, Aniruddha Bala, Rohan Jaiswal, and Siddharth Roheda. Vid-freeze: Protecting images from malicious image-to-video generation via temporal freezing. *arXiv preprint arXiv:2509.23279*, 2025.

Tuo Feng, Wenguan Wang, and Yi Yang. A survey of world models for autonomous driving. *arXiv preprint arXiv:2501.11260*, 2025.

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.

Zhixiang Guo, Siyuan Liang, Andras Balogh, Noah Lunberry, Rong-Cheng Tu, Mark Jelasity, and Dacheng Tao. When world models dream wrong: Physical-conditioned adversarial attacks against world models. *arXiv preprint arXiv:2602.18739*, 2026a.

Zhixiang Guo, Siyuan Liang, Shi Fu, Cheng Guo, Andras Balogh, Mark Jelasity, and Dacheng Tao. Wmattack: Automated attack search for adversarial evaluation of world-model agents. *arXiv preprint arXiv:2605.23220*, 2026b.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *International conference on machine learning*, pp. 2555–2565. PMLR, 2019.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023.

Nikolaus Hansen. The cma evolution strategy: A tutorial. *arXiv preprint arXiv:1604.00772*, 2016.

Xianglong He, Chunli Peng, Zexiang Liu, Boyang Wang, Yifan Zhang, Qi Cui, Fei Kang, Biao Jiang, Mengyin An, Yangyang Ren, et al. Matrix-game 2.0: An open-source real-time and streaming interactive world model. *arXiv preprint arXiv:2508.13009*, 2025.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 7514–7528, 2021.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. Gaia-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080*, 2023.

Xiaotao Hu, Mingkai Jia, Xiaoyang Guo, Qian Zhang, Xiao-xiao Long, and Wei Yin. Drivingworld: Constructing world model for autonomous driving via video gpt. In *International Conference on Pattern Recognition*, pp. 276–291. Springer, 2026.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. *Advances in Neural Information Processing Systems*, 38:167283–167308, 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.

Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, et al. Vbench++: Comprehensive and versatile benchmark suite for video generative models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Team HunyuanWorld. Hy-world 1.5: A systematic framework for interactive world modeling with real-time latency and geometric consistency. *arXiv preprint*, 2025.

Bingyi Kang, Yang Yue, Rui Lu, Zhijie Lin, Yang Zhao, Kaixin Wang, Gao Huang, and Jiashi Feng. How far is video generation from world model: A physical law perspective. *arXiv preprint arXiv:2411.02385*, 2024.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025.

Guanlin Li, Shuai Yang, Jie Zhang, and Tianwei Zhang. Prime: Protect your videos from malicious editing. *arXiv preprint arXiv:2402.01239*, 2024.

Xinqing Li, Xin He, Le Zhang, Min Wu, Xiaoli Li, and Yun Liu. A comprehensive survey on world models for embodied ai. *arXiv preprint arXiv:2510.16732*, 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. *arXiv preprint arXiv:2302.04578*, 2023.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Shaowei Liu, Zhongzheng Ren, Saurabh Gupta, and Shenlong Wang. Physgen: Rigid-body physics-grounded image-to-video generation. In *European Conference on Computer Vision*, pp. 360–378. Springer, 2024a.

Yisu Liu, Jinyang An, Wanqian Zhang, Dayan Wu, Jingzi Gu, Zheng Lin, and Weiping Wang. Disrupting diffusion: Token-level attention erasure attack against diffusion-based customization. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 3587–3596, 2024b.

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.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Xiaofeng Mao, Shaoheng Lin, Zhen Li, Chuanhao Li, Wenshuo Peng, Tong He, Jiangmiao Pang, Mingmin Chi, Yu Qiao, and Kaipeng Zhang. Yume: An interactive world generation model. *arXiv preprint arXiv:2507.17744*, 2025.

Fanqing Meng, Jiaqi Liao, Xinyu Tan, Wenqi Shao, Quanfeng Lu, Kaipeng Zhang, Yu Cheng, Dianqi Li, Yu Qiao, and Ping Luo. Towards world simulator: Crafting physical commonsense-based benchmark for video generation. *arXiv preprint arXiv:2410.05363*, 2024.

Saman Motamed, Laura Culp, Kevin Swersky, Priyank Jaini, and Robert Geirhos. Do generative video models understand physical principles? In *2026 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 948–958. IEEE, 2026.

Abby O'Neill, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, et al. Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6892–6903. IEEE, 2024.

Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 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.

Shawn Shan, Jenna Cryan, Emily Wenger, Haitao Zheng, Rana Hanocka, and Ben Y Zhao. Glaze: Protecting artists from style mimicry by {Text-to-Image} models. In *32nd USENIX Security Symposium (USENIX Security 23)*, pp. 2187–2204, 2023.

Linghui Shen, Mingyue Cui, and Xingyi Yang. Decontext as defense: Safe image editing in diffusion transformers. *arXiv preprint arXiv:2512.16625*, 2025.

Robbyant Team, Zelin Gao, Qiuyu Wang, Yanhong Zeng, Jiapeng Zhu, Ka Leong Cheng, Yixuan Li, Hanlin Wang, Yinghao Xu, Shuailei Ma, et al. Advancing open-source world models. *arXiv preprint arXiv:2601.20540*, 2026.

Hansi Teng, Hongyu Jia, Lei Sun, Lingzhi Li, Maolin Li, Mingqiu Tang, Shuai Han, Tianning Zhang, WQ Zhang, Weifeng Luo, et al. Magi-1: Autoregressive video generation at scale. *arXiv preprint arXiv:2505.13211*, 2025.

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.

Feifei Wang, Zhentao Tan, Tianyi Wei, Yue Wu, and Qidong Huang. Simac: A simple anti-customization method for protecting face privacy against text-to-image synthesis of diffusion models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12047–12056. IEEE, 2024a.

Hanhui Wang, Yihua Zhang, Ruizheng Bai, Yue Zhao, Sijia Liu, and Zhengzhong Tu. Edit away and my face will not stay: Personal biometric defense against malicious generative editing. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23806–23816. IEEE, 2025a.

Jiahao Wang, Yufeng Yuan, Rujie Zheng, Youtian Lin, Jian Gao, Lin-Zhuo Chen, Yajie Bao, Chang Zeng, Yanxi Zhou, Xiao-Xiao Long, et al. Spatialvid: A large-scale video dataset with spatial annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 42592–42603, 2026a.

Ruotong Wang, Mingli Zhu, Jiarong Ou, Rui Chen, Xin Tao, Pengfei Wan, and Baoyuan Wu. Badvideo: Stealthy backdoor attack against text-to-video generation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19075–19084. IEEE, 2025b.

Yuqi Wang, Jiawei He, Lue Fan, Hongxin Li, Yuntao Chen, and Zhaoxiang Zhang. Driving into the future: Multiview visual forecasting and planning with world model for autonomous driving. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14749–14759. IEEE, 2024b.

Zile Wang, Zexiang Liu, Jiaxing Li, Kaichen Huang, Baixin Xu, Fei Kang, Mengyin An, Peiyu Wang, Biao Jiang, Yichen Wei, et al. Matrix-game 3.0: Real-time and streaming interactive world model with long-horizon memory. *arXiv preprint arXiv:2604.08995*, 2026b.

Zeqi Xiao, Yushi Lan, Yifan Zhou, Wenqi Ouyang, Shuai Yang, Yanhong Zeng, and Xingang Pan. Worldmem: Long-term consistent world simulation with memory. *Advances in Neural Information Processing Systems*, 38:49632–49652, 2026.

Zhen Xing, Qijun Feng, Haoran Chen, Qi Dai, Han Hu, Hang Xu, Zuxuan Wu, and Yu-Gang Jiang. A survey on video diffusion models. *ACM Computing Surveys*, 57(2):1–42, 2024.

Jingyao Xu, Yuetong Lu, Yandong Li, Siyang Lu, Dongdong Wang, and Xiang Wei. Perturbing attention gives you more bang for the buck: Subtle imaging perturbations that efficiently fool customized diffusion models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24534–24543. IEEE, 2024.

Shuhan Xu, Siyuan Liang, Hongling Zheng, Yong Luo, Han Hu, Lefei Zhang, and Dacheng Tao. Ctrlattack: A unified attack on world-model control in diffusion models. *arXiv preprint arXiv:2603.13435*, 2026.

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*, volume 2025, pp. 83048–83077, 2025.

Kaiwen Zhang, Zhenyu Tang, Xiaotao Hu, Xingang Pan, Xiaoyang Guo, Yuan Liu, Jingwei Huang, Li Yuan, Qian Zhang, Xiao-Xiao Long, et al. Epona: Autoregressive diffusion world model for autonomous driving. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 27220–27230. IEEE, 2025.

Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all. *arXiv preprint arXiv:2412.20404*, 2024.

Hongzhou Zhu, Min Zhao, Guande He, Hang Su, Chongxuan Li, and Jun Zhu. Causal forcing: Autoregressive diffusion distillation done right for high-quality real-time interactive video generation. *arXiv preprint arXiv:2602.02214*, 2026a.

Yixuan Zhu, Jiaqi Feng, Wenzhao Zheng, Yuan Gao, Xin Tao, Pengfei Wan, Jiwen Lu, and Jie Zhou. Astra: General interactive world model with autoregressive denoising. In *International Conference on Learning Representations*, volume 2026, pp. 79167–79184, 2026b.

A APPENDIX

In this supplement, we provide further details on BadWorld, including objective explanations, robustness evaluations, implementation specifics, and additional experimental results. Specifically, Sec. A.1 provides an extended review of related work, Sec. A.2 offers a deeper explanation of the velocity magnitude objective, Sec. A.3 evaluates adversarial robustness against image preprocessing and black-box transferability across models, Sec. A.4 details trajectory sampling algorithms and experimental hyperparameters, Sec. A.7 presents extension experiments on Matrix-Game-2 variants, and Sec. B provides further qualitative visuals of our attack.

A.1 RELATED WORK

Controllable Visual World Models Recent advancements in diffusion and flow-matching frameworks (Ho et al., 2020; Lipman et al., 2022; Huang et al., 2026; Chen et al., 2024; Zhu et al., 2026a) have shifted video generation from passive synthesis toward interactive simulation (Blattmann et al., 2023; Chen et al., 2025; Bar-Tal et al., 2024; Wan et al., 2025; Kong et al., 2024; Polyak et al., 2024; Zheng et al., 2024; Xing et al., 2024). Unlike standard generators, visual world models (VWMs) aim to internalize learned physics and transition dynamics, evolving in response to agent actions and temporal history. A prominent trend involves adapting pretrained models into autoregressive frameworks for viewpoint-aware rollouts. Within this landscape, control is typically implemented through three pathways: explicit continuous camera trajectories (Bai et al., 2025; Zhu et al., 2026b; Xiao et al., 2026), discrete action tokens (He et al., 2025; HunyuanWorld, 2025; Team et al., 2026), or implicit text descriptions (Mao et al., 2025). We focus on these autoregressive denoising VWMs, selecting models from both continuous and discrete paradigms to investigate their adversarial robustness. This assesses whether these interactive environments reliably maintain consistency and control adherence, which is essential for safety-critical deployment.

Adversarial Attacks on Generative Models Adversarial research in generative modeling typically focuses on two objectives: assessing structural vulnerabilities and enhancing privacy protection. Early work predominantly targeted image generation, employing imperceptible pixel-level perturbations to disrupt generation process. Studies on Text-to-Image (T2I) personalization employ training data poisoning to prevent unauthorized model fine-tuning (Xu et al., 2024; Liu et al., 2024b; Van Le et al., 2023; Liang et al., 2023; Wang et al., 2024a), while Image-to-Image (I2I) attacks perturb input context images to trigger abnormal outputs for privacy protection (Salman et al., 2023; Wang et al., 2025a; Shen et al., 2025; Choi et al., 2025). Building upon prior image-based work, adversarial paradigms have recently extended to video generation, where temporal dynamics introduce new attack surfaces. Specifically, identified threats include backdoor triggers in text prompts (Wang et al., 2025b), jailbreaking techniques (Li et al., 2024), and adversarial trajectories in drag-based I2V systems (Xu et al., 2026). Recent studies further investigate world-model attacks via physical-condition perturbation in driving scenes (Guo et al., 2026a) or automated attack search for world agents (Guo et al., 2026b). However, they address condition-level or search-level attacks, while we study image-space attacks on video world models. Most relevant to our work are approaches that target the input images in I2V pipelines (Gui et al., 2025; Vu et al., 2026; Chowdhury et al., 2025; Long et al., 2026). While effective, these frameworks are primarily evaluated on standard diffusion-based backbones like SVD (Blattmann et al., 2023), CogVideoX (Yang et al., 2025), or Wan2.1 (Wan et al., 2025). A critical gap remains in evaluating the adversarial robustness of autoregressive video world models, particularly those conditioned on interactive signals like camera trajectories or action sequences. To bridge this gap, we propose the first adversarial attack framework tailored to unique architectures of autoregressive video world models.

A.2 Explanation of the Velocity Magnitude Objective

This section further explains the velocity magnitude objectives introduced in the main paper.

In flow-matching (Lipman et al., 2022) frameworks, the denoising network predicts a velocity field $\hat{v}_\theta$ specifying the direction and rate of latent-state change at each step. To isolate the effect of velocity magnitude, we perform a velocity scaling experiment during inference. Specifically, we modulate the predicted velocity by a positive scalar s, yielding $\hat{v}'_\theta = s \cdot \hat{v}_\theta$. This operation preserves the prediction direction and changes only its magnitude. The latent state is then updated via $z_{t-\Delta t} = \mathrm{SchedulerStep}(s \cdot \hat{v}_\theta, t, z_t)$ using the scaled velocity.

Empirical results in Fig. 5 and 6 show that the magnitude of $\hat{v}_\theta$ directly determines visual quality. When velocity is under-scaled ($s < 1$), the update is too weak to drive the latent state toward the data manifold. This yields under-denoised outputs that appear gray, blurry, and structurally disordered. Conversely, over-scaling velocity ($s > 1$) pushes latents toward extreme values. While amplifying contrast, it often introduces overshooting artifacts such as extreme saturation and pixel distortion. These observations indicate that a precise velocity norm is vital for a stable denoising trajectory.

Building on these observations, we find that at equivalent scaling intensity, reducing velocity magnitude disrupts video coherence more severely than increasing it. For adversarial purposes, Velocity-Min proves superior to Velocity-Max in triggering rapid structural collapse. We hypothesize that this vulnerability stems from autoregressive world models: weakened updates prevent the model from reaching stable manifolds, causing subtle denoising errors to accumulate and compound through the temporal history window. This also motivates our choice of Velocity-Min as the final attack objective.



Figure 5: Scale the velocity magnitude for Astra.



Figure 6: Scale the velocity magnitude for Matrix-Game-2.0.

A.3 Robustness and Transferability

Following the setup in Sec. A.6, we evaluate the robustness and transferability of our method. We randomly select 10 context images and generate videos across three distinct camera trajectories.

Robustness to Image Preprocessing.

We examine the resilience of our adversarial perturbations against common image preprocessing techniques, including Gaussian noise, JPEG compression, and Total Variation (TV) denoising. As shown in Tab. 8, Velocity-Min remains highly effective under light Gaussian noise ($\sigma = 0.5$), maintaining a strong MEt3R score of 0.255 and significantly low imaging quality. While more aggressive preprocessings like JPEG and TV denoising partially mitigate the attack, our method still induces measurable degradation compared to the clean baseline. These results demonstrate that our perturbations are not easily neutralized by standard image-level filters, confirming their practical robustness.

Method	I2V Background ↓	Background Consistency ↓	Aesthetic Quality ↓	Imaging Quality ↓	CLIP-I↓	MEt3R↑
Clean	0.978	0.915	0.501	0.690	0.813	0.151
Velocity-Min	0.922	0.816	0.405	0.506	0.701	0.266
+gaussianσ0.5	0.924	0.823	0.407	0.514	0.687	0.255
+gaussian σ0.7	0.927	0.846	0.437	0.575	0.721	0.245
+JPEG 75	0.969	0.896	0.477	0.670	0.715	0.164
+TV λ0.04	0.960	0.882	0.454	0.587	0.749	0.191

Table 8: Robustness to Image Preprocessing.

Black-box Transfer across Models and Limitations.

To evaluate the black-box transferability of our method, we conduct cross-model evaluations between Matrix-Game-2.0 (He et al., 2025) and Astra (Zhu et al., 2026b). Given their different default input resolutions, we resize the generated adversarial context images to match the target model's configuration prior to inference. As shown in Tab. 9: (1) Trained on Matrix-Game, test on Astra (Left): Our method demonstrates clear transferability. The video quality metrics notably decline compared to the clean baseline. Specifically, Background Consistency drops from 0.878 to 0.846, and Imaging Quality decreases from 0.553 to 0.515. This confirms the attack's effectiveness on an unseen architecture. (2) Trained on Astra, Test on Matrix-Game (Right): The adversarial transferability is severely limited, showing negligible performance degradation. This asymmetry highlights a key limitation: excessive resizing required to bridge resolution gaps likely disrupts the delicate pixel-level perturbations, thereby degrading transferability in black-box scenarios.

	Matrix-Game → Astra			Astra → Matrix-Game		
Method	Back. Consis.	Aesthetic	Imaging Qual.	Back. Consis	Aesthetic	Imaging Qual.
Clean	0.878	0.492	0.553	0.955	0.488	0.659
Velocity-Min	0.846	0.466	0.515	0.954	0.485	0.645

Table 9: Black-box Transfer across Models.

A.4 Implementation Details

A.5 Trajectory Sampling

A.5.1 Camera Trajectory Formulation

To ensure the adversarial search space is both expressive and tractable, we represent the camera control sequence using a compact, low-dimensional parameterization. At any target frame t, the control signal is defined as a triplet $c_t = (\psi_t, f_t, s_t)$, representing the yaw angle, forward displacement, and lateral shift, respectively. This triplet is uniquely mapped to a relative camera pose matrix $P_t \in \mathbb{R}^{3\times4}$:

$$P_t = \begin{bmatrix} \cos\psi_t & 0 & \sin\psi_t & s_t \\ 0 & 1 & 0 & 0 \\ -\sin\psi_t & 0 & \cos\psi_t & -f_t \end{bmatrix}$$

The first 3×3 block encodes the yaw rotation around the vertical axis, while the translation vector captures horizontal and forward motions. For an autoregressive chunk comprising $T = 8$ frames, the complete continuous trajectory is vectorized as $\tau = [\psi_1, f_1, s_1, \ldots, \psi_T, f_T, s_T]^\top \in \mathbb{R}^{3T}$.To ensure the optimized trajectories remain physically plausible and strictly within the model's in-distribution control space, we enforce bound constraints $\mathcal{B}$ on both the absolute magnitudes and the temporal derivatives of the sequence. Specifically, the frame-wise controls are bounded by $\psi_{\max} = 0.05$, $f_{\max} = 0.05$, and $s_{\max} = 0.025$. To prevent erratic, discontinuous camera jumps, the step-wise variations are strictly bounded by $\Delta_\psi = 0.03$, $\Delta_f = 0.03$, and $\Delta_s = 0.015$. All sampled or optimized trajectories are projected onto this feasible convex set.

A.5.2 Stochastic Trajectory Sampling via Random Walk

For the baseline self-supervised objectives (Sec.4.1) where adaptive mining is not employed, we approximate the expectation over control signals through stochastic trajectory sampling. At each step of the Projected Gradient Descent (PGD) (Madry et al., 2017), a fresh trajectory is generated to prevent the adversarial perturbation from overfitting to a static camera motion. The initial frame control c_1 is drawn uniformly from the frame-wise feasible range. Subsequent frames follow a Gaussian random walk to ensure temporal coherence:

$$c_t = c_{t-1} + \eta_t, \quad \eta_t \sim \mathcal{N}(0, \Sigma_{rw})$$

where $\Sigma_{rw} = \mathrm{diag}(\sigma_\psi^2, \sigma_f^2, \sigma_s^2)$. The resulting trajectory is subsequently clipped to satisfy the variation and range constraints defined in Sec. A.5.1. This unbiased stochastic exploration ensures the generated perturbation maintains generalizability across a wide distribution of camera motions.

A.5.3 Trajectory-Adaptive Bi-Level Execution

For the trajectory-adaptive bi-level attack (Sec. 4.2), during the outer optimization loop, we search for trajectories that maximize the empirical adversarial loss. To account for the stochasticity of the denoising process, the fitness of a candidate trajectory τ is evaluated via Monte Carlo approximation over M evaluation contexts (fixing the history length, timestep, and latent noise):

$$F(\tau; \delta) = \frac{1}{M} \sum_{j=1}^{M} \mathcal{L}_{atk}(x + \delta; \tau, \xi_j)$$

For efficiency, we default to $M = 1$. We maintain separate hard trajectory pools $\mathcal{P}_{\mathrm{hard}}$ tailored for different autoregressive history lengths, recognizing that the sensitivity of the model to specific camera motions diverges as the rollout progresses. During the inner loop, the PGD update dynamically queries these pools. For standard short-horizon steps, the attack defaults to the random-walk sampler to maintain broad robustness. For extended horizons where error accumulation is critical, the attack samples directly from $\mathcal{P}_{\mathrm{hard}}$, forcing the inner minimization to prioritize control sequences under which the current perturbation is least effective.

A.5.4 COVARIANCE MATRIX ADAPTATION EVOLUTION STRATEGY (CMA-ES)(HANSEN, 2016)

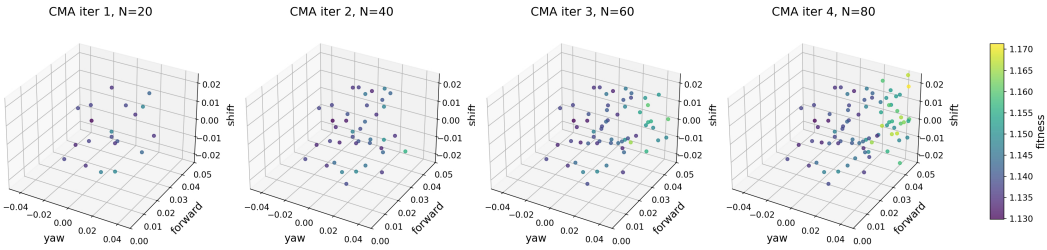

Figure 7: CMA trajectory updates.

The outer maximization over $\tau \in \mathbb{R}^{3T}$ presents a challenging non-convex optimization problem over a continuous sequence. Importantly, computing exact gradients through the multi-step autoregressive generation process is computationally prohibitive and prone to gradient shattering. We thus employ CMA-ES(Hansen, 2016), a derivative-free evolutionary algorithm, to efficiently navigate this space. At generation g, CMA-ES maintains a multivariate Gaussian search distribution parameterized by a mean trajectory $m^{(g)}$, a global step size $\sigma^{(g)}$, and a covariance matrix $C^{(g)}$:

$$q^{(g)}(\tau) = \mathcal{N}\left(m^{(g)}, (\sigma^{(g)})^2 C^{(g)}\right)$$

A population of λ candidate trajectories is sampled, projected into the feasible region, and evaluated against the fitness function $-F(\tau; \delta)$. Let $\{\tau_{i:\lambda}^{(g)}\}_{i=1}^{\mu}$ denote the top μ candidates sorted by fitness. The distribution mean is updated toward the weighted average of these elite candidates:

$$m^{(g+1)} = \sum_{i=1}^{\mu} w_i \tau_{i:\lambda}^{(g)}$$

where $w_i > 0$ are the recombination weights. A critical advantage of CMA-ES for sequence mining is the adaptation of the covariance matrix $C^{(g)}$, which captures the strong temporal correlations inherent in hard camera motions (e.g., a coordinated continuous pan and forward zoom). The covariance is updated by integrating both the current elite population and an accumulated evolution path

$$C^{(g+1)} = (1 - c_{\text{cov}})C^{(g)} + c_{\text{cov}} \sum_{i=1}^{\mu} w_i y_{i:\lambda}^{(g)} \left(y_{i:\lambda}^{(g)}\right)^{\top} \tag{17}$$

where $y_{i:\lambda}^{(g)} = (\tau_{i:\lambda}^{(g)} - m^{(g)})/\sigma^{(g)}$ is the normalized step. As shown in Fig.7, the trajectories are successively updated to higher velocity norm. Upon termination of the CMA-ES generations, the globally top-ranked candidates across the evaluation history are distilled into the hard trajectory pool $\mathcal{P}_{\text{hard}}$ for the inner PGD optimization.

A.6 ADDITIONAL EXPERIMENTAL DETAILS

Training. As discussed in Sec.4.1, the context frame available to users does not come with paired ground-truth videos or viewpoints annotations. To simulate realistic attack scenarios, we introduce two key adjustments: (i) we restrict training to the early denoising phase (timesteps ranging from 950 to 1000 by default) where target frames can be approximated as pure noise, and (ii) we simulate historical context by duplicating the context frame.

For conditioning, Astra uses 30 prompt variants generated via GPT-4o-mini, each semantically aligned with the context frames; we randomly sample one prompt per attack and a unique camera pose per latent frame. Matrix-Game 2.0 requires no text prompts, so we randomly sample discrete actions per frame.

For each model, all four attack objectives are evaluated under identical training schedules (Tab. 10). We conduct training on an A800 80GB GPU.

Model	VAE Ratio	Frames per Chunk	Attack Budget	Step Size	Steps
Astra	4	8	0.05	0.005	300
Matrix-Game 2.0	4	3	0.05	0.004	300

Table 10: Training hyperparameters for different models.

Inference. We follow the default inference configurations (Tab. 11) of the base models and generate videos autoregressively. For each context frame (input image), we produce 3 videos using different camera or action sequences to ensure result diversity.

Model	Frames per Chunk	Steps per Chunk	Target Latent Frames	Fps	Attn. Window
Astra	8	50	33	20	20
Matrix-Game 2.0	3	3	60	5	4

Table 11: Inference settings for both models.

Evaluations. While Astra is evaluated via standard VBench metrics, we adapt our approach for the longer-form Matrix-Game-2. Specifically, we employ VBench-Long for video quality evaluation and segment the original videos into shorter clips to precisely assess contextual consistency. This length-aware strategy ensures a more rigorous and reliable evaluation.

A.7 Extension Experiments on Matrix-Game-2 Variants

For the Matrix-Game-2 universal variants, we train the adversarial perturbation with an attack budget of 0.05 and a step size of 0.002 for a total of 700 optimization steps. During inference, we follow the default Matrix-Game-2 configuration: the local attention window is set to 6, each chunk contains 3 latent frames, and generation is performed with 3 sampling steps. For each input image, we generate 27 latent frames in total.

As shown in Table 12 and Figure 8, BadWorld effectively generalizes to the Matrix-Game-2 Universal variant. The Velocity-Min objective triggers substantial degradation across all quantitative metrics, notably reducing Imaging Quality from 0.652 to 0.515 and Aesthetic Quality from 0.513 to 0.418. Qualitative results further confirm that the adversarial perturbation successfully induces structural collapse in the rollouts, validating the robustness of our framework across different model configurations.

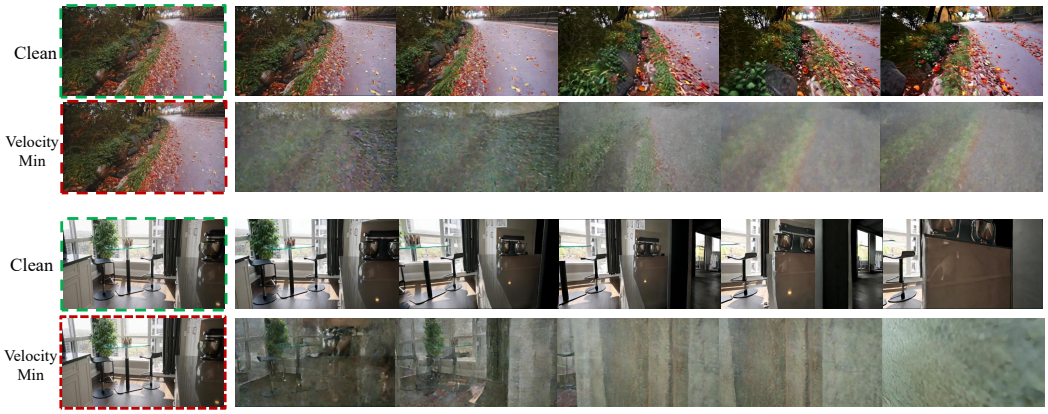

Figure 8: Attack Performance on Matrix-Game-2.0 (Universal).

Method	I2V Background ↓	Background Consistency ↓	Aesthetic Quality ↓	Imaging Quality ↓	CLIP-I↓	MEt3R↑
Clean	0.961	0.932	0.513	0.652	0.755	0.162
Velocity-Min	0.874	0.875	0.418	0.515	0.703	0.223

Table 12: Attack on Matrix-Game-2 (Universal)

B ADDITIONAL QUALITATIVE RESULTS

This section provides additional qualitative results on Astra and Matrix-Game-2.0.

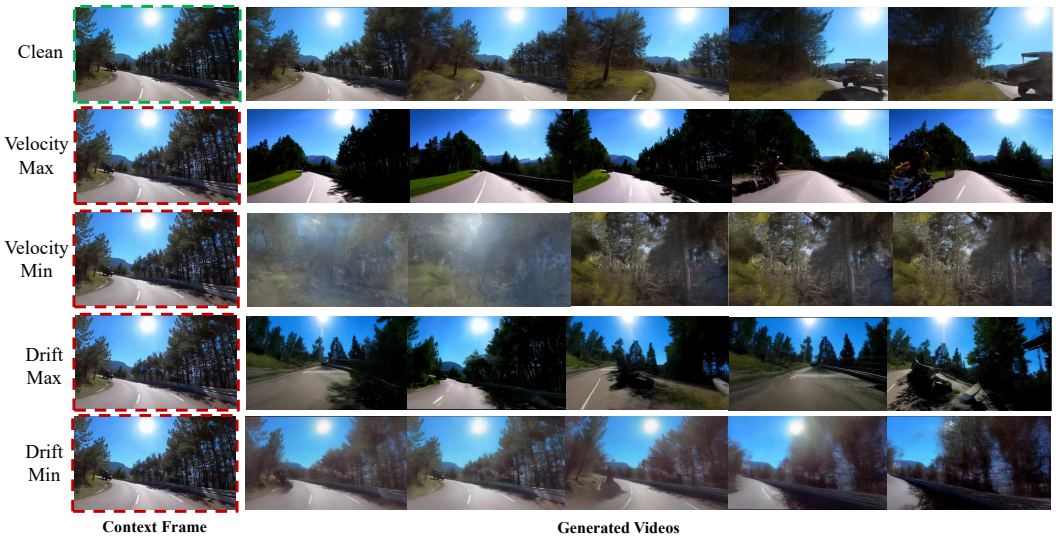

Figure 9: Qualitative Results on Astra.

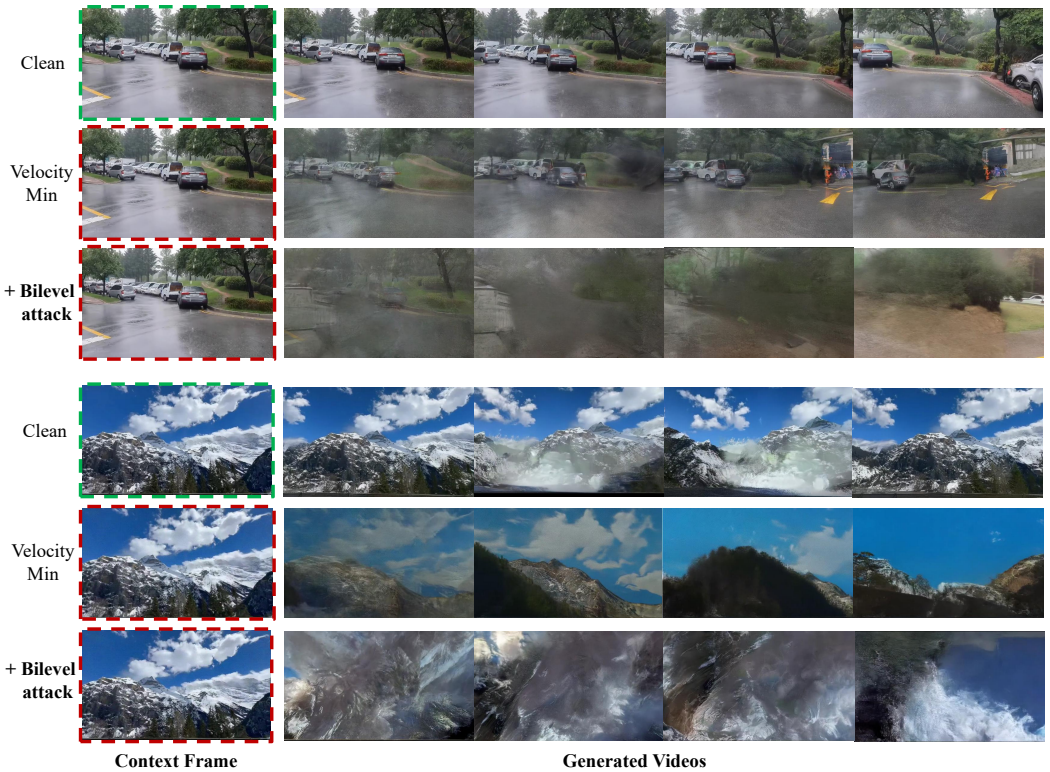

Figure 10: Performance of Bi-Level Attack.

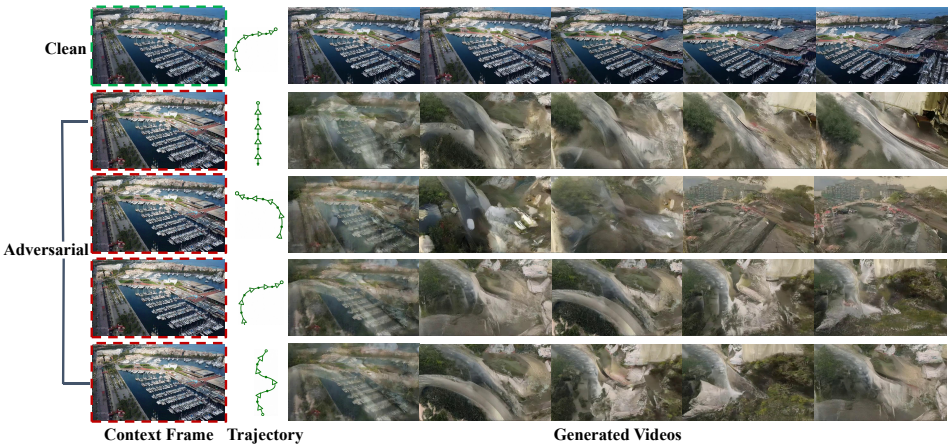

Figure 11: Attack performance unfer different camera trajectories.



Figure 12: Qualitative Results on Matrix-Game-2.0.