PHYSREFLEX: COUNTERFACTUAL CYCLIC REFINEMENT FOR PROMPT-FAITHFUL PHYSICAL EVENTS IN VIDEO GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Modern video generators achieve impressive visual fidelity, yet still struggle to faithfully capture the physical event specified by the prompt. This failure is particularly subtle when different event outcomes produce similar-looking motion, making current-state refinement potentially insufficient for distinguishing whether the intended dynamics are being formed. Existing methods address this challenge through physics-aware training, explicit physical priors, or inference-time guidance and refinement. However, training-free inference methods typically correct generation based on the current sampling state, without explicitly examining how different physical evolutions unfold from that state. We introduce PhysReflex, a training-free framework for counterfactual cyclic refinement. At selected sampling steps, PhysReflex briefly advances the current latent under the target condition, then traces it back under both the target and a counterfactual condition. The discrepancy between the returned states provides an evolution-aware signal for correcting the target sampling trajectory. We further adapt the number of refinement cycles according to changes in the predicted clean-video latent, avoiding unnecessary or excessive correction. Experiments on VideoPhy2 and PhyWorldBench demonstrate consistent improvements over baselines across comprehensive evaluations of physical plausibility, motion consistency, and semantic alignment.

1 INTRODUCTION

Text-to-video models increasingly produce clips that are sharp, temporally smooth, and broadly consistent with a prompt (Ho et al., 2022; Singer et al., 2023; Yang et al., 2025b). Yet they often respond weakly to the part of a prompt that specifies how an event unfolds. A request for a car to brake slowly can yield nearly the same gradual deceleration as a request for the car to brake abruptly; a request for one bounce can produce several; and a request for a weak impact can produce the same visible response as a strong one. Each output may look plausible in isolation, but the model fails to reliably distinguish and express the event attribute specified by the prompt. The resulting problem is therefore not simply physical implausibility: it is a lack of prompt-faithful control over the dynamics of the requested event.

We call this failure *physical-event under-conditioning*. It is distinct from object- or scene-level misalignment, as well as from physical implausibility. Physical plausibility provides a useful constraint, but it does not determine which physically valid event should occur. A slowly braking car and an abruptly braking car can both be physically plausible. Improving plausibility alone can therefore leave the central distinction in the prompt unresolved.

Existing approaches to physical video generation often address this problem through additional training, parameter-efficient adaptation, or external physical simulators and priors. These strategies can be effective, but they require domain-specific data, simulator design, or optimization and may alter the capabilities of the underlying video generator. They also make it difficult to isolate whether a model is genuinely sensitive to the event attribute expressed in the prompt. A training-free approach offers a complementary setting: it preserves the frozen generator and its native visual prior, introduces no new physical supervision, and places the burden of event control entirely on the inference procedure. This setting is especially useful for diagnosing and correcting prompt sensitivity

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.

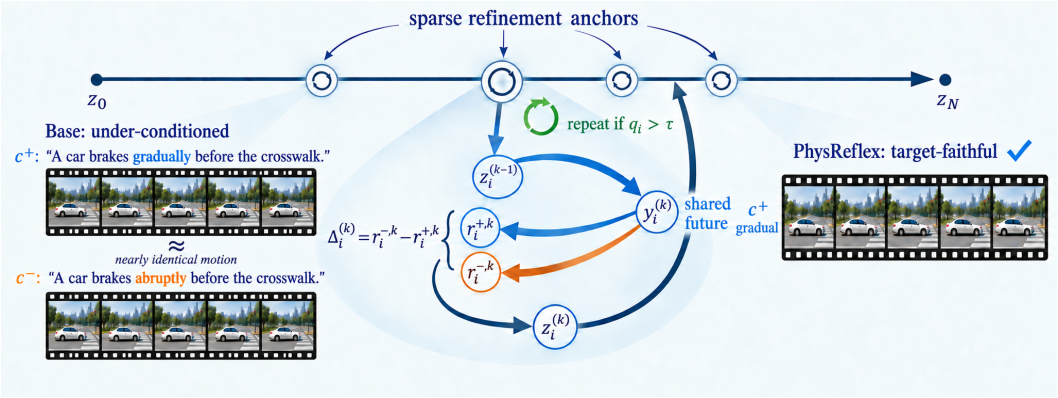

Figure 1: **Motivation.** Two prompts can describe different physical events—for example, gradual versus abrupt braking—yet the base generator may produce nearly indistinguishable motion. The figure illustrates the gap between physical plausibility and event-specific prompt fidelity: both clips can look reasonable, while neither reliably expresses the requested distinction. PhysReflex addresses this gap by using a physically valid counterfactual to expose the event response of the frozen generator; the detailed local cycle and adaptive allocation rule are introduced in Sec. 3.

in models that already produce plausible videos but fail to distinguish between physically admissible event outcomes. Despite this practical and diagnostic value, training-free event-level physical control remains comparatively underexplored.

This paper studies inference-time correction for this failure mode. Given a target condition c^+ and a matched counterfactual condition c^- that preserves the scene while changing one physically admissible event attribute, we ask whether the frozen generator provides a sufficiently distinct conditional response to favor the requested outcome. The direct conditional difference at the current noisy state is a natural baseline, but it can be weak and can be confounded by the state that the model has already formed. Our key construction is a local *counterfactual cycle*: from a sampling anchor, we first move along the target-conditioned denoising trajectory for a short interval, then integrate back from the same endpoint under both conditions. We apply the difference between these two returns at the original noise level, after which the main target trajectory resumes.

The cycle changes both where the comparison is made and which local state it probes. The target return provides a same-condition reference for the counterfactual return, allowing common round-trip drift to be separated from the event-specific contrast. This construction differs from subtracting target and counterfactual predictions at the anchor, and it does not maintain a second counterfactual trajectory throughout the entire generation process. Crucially, the denoising time at which the requested event becomes visually determined is not fixed across prompts or samples. A probe applied too early may precede the state change that determines the event, whereas a probe applied too late may have little influence because the trajectory has already committed to an outcome. We therefore use a response-sensitive gate: after the initial probe at a selected anchor, we measure whether the target clean prediction remains responsive to the correction and allocate another probe only when that signal and the remaining budget justify it. The gate is thus an adaptive allocation of event-control computation, rather than a generic physical-error estimator.

Our central claim is correspondingly specific: PHYSREFLEX improves *prompt-faithful physical event realization*, while physical plausibility can improve at the same time. On VideoPhy2, PHYSREFLEX produces consistent gains under both Qwen3-VL-32B and Gemini 2.5 Flash Lite. The largest improvement appears in semantic alignment (SA), increasing by 8.34 and 7.50 percentage points over direct CFG, respectively; physical commonsense (PC) and joint success improve as well. These results support the view that PHYSREFLEX addresses a prompt-sensitivity gap in event realization rather than merely producing smoother or more visually plausible videos.

The contributions are:

- We introduce PhysReflex, a training-free local counterfactual cycle that compares target and physically valid alternative returns from a shared target-conditioned future, rather than using only an anchor difference or a persistent second trajectory.
- We add a response-sensitive gate that allocates another probe when the event-forming state remains controllable, while keeping the generator and main sampling schedule frozen.
- We evaluate PhysReflex on VideoPhy2 and PhyWorldBench and find the clearest completed gains in semantic alignment to the requested event. The evaluation also tracks physical commonsense and joint success, allowing prompt adherence and plausibility to be assessed together.

2 Related Work

Generative backbones. Latent diffusion reduces the cost of high-resolution synthesis (Rombach et al., 2022); Align Your Latents extends this approach through temporal modeling (Blattmann et al., 2023). Diffusion Transformers provide a scalable latent backbone (Peebles & Xie, 2023), while CogVideoX combines video compression with text-conditioned transformers (Yang et al., 2025b).

Physical plausibility and event fidelity. Video benchmarks such as VBench, EvalCrafter, and T2V-CompBench separate visual quality, motion, and text alignment (Huang et al., 2024; Liu et al., 2024b; Sun et al., 2025). VideoPhy and VideoPhy-2 evaluate physical commonsense and action-centric event understanding, while PhyGenBench, Physics-IQ, T2VPhysBench, and PhyWorldBench cover complementary physical phenomena (Bansal et al., 2025; 2026; Meng et al., 2025; Motamed et al., 2026; Guo et al., 2025; Gu et al., 2026). VBench-2.0 extends evaluation toward intrinsic faithfulness, while Physics Question Scene Graph supports fine-grained localization of physical violations (Zheng et al., 2025; Pothiraj et al., 2026). These benchmarks motivate separating whether motion is physically acceptable from whether it realizes the particular event described in the prompt. Our work targets this latter distinction without treating plausibility and event fidelity as competing objectives.

Physics-aware video generation. PhysGen uses rigid-body simulation, PhysVid learns local physical conditioning, PhyT2V performs language-guided prompt refinement, and VLIPP plans motion with a vision-language model (Liu et al., 2024a; Pathak et al., 2026; Xue et al., 2025; Yang et al., 2025a). DiffPhy, PSIVG, and MMPhysVideo inject physical context through training, simulation, or multimodal priors (Zhang et al., 2025; Foo et al., 2026; Lin et al., 2026). VideoJAM jointly learns appearance and motion predictions and uses motion guidance during inference (Chefer et al., 2025). These approaches target physical or motion consistency. In contrast, we focus on the case where several outcomes are plausible but only one matches the requested event attribute.

Conditional and counterfactual guidance. Classifier-free guidance (CFG) combines conditional and unconditional predictions (Ho & Salimans, 2022); related training-free methods use energy, arbitrary guidance functions, or weaker references (Karras et al., 2024; Bansal et al., 2023; Yu et al., 2023; Ye et al., 2024). Enhancing Physical Plausibility constructs physics-violating alternatives and uses Synchronized Decoupled Guidance (SDG) with separate trajectories (Hao et al., 2025). Visual-semantic chain-of-events conditioning likewise plans causal transitions and applies guidance against physics-violating counterfactual prompts (Wang et al., 2026). Our counterfactual is deliberately physically valid and changes the event attribute requested by the user. More importantly, we compare target and counterfactual returns from a shared target-shaped future, rather than subtracting fields only at the anchor or maintaining a global counterfactual trajectory. SDG addresses plausibility by asking how target and implausible alternatives diverge throughout sampling; PhysReflex asks how two admissible event conditions respond to the same target-shaped local future, then discards both probe paths. The former is global dual-trajectory guidance; the latter is local anchored response measurement.

Inference-time refinement and cycles. DDIM, DPM-Solver, and UniPC improve the sampling computation–quality trade-off (Song et al., 2021; Lu et al., 2022; Zhao et al., 2023); verifier-guided noise search explores additional inference computation (Ma et al., 2025). SDEdit establishes noise–denoise cycles for training-free editing (Meng et al., 2022). CycleDiffusion extends cyclic

consistency across diffusion domains (Wu & De la Torre, 2022). Restart, Video-T1, Zigzag sampling, Weak-to-Strong Diffusion with Reflection, U-Turn chains, and Self-Refining Video Sampling demonstrate that extra inference can improve frozen generators (Xu et al., 2023; Liu et al., 2025; Bai et al., 2025; 2026; Kang et al., 2026; Jang et al., 2026). Our contribution is the event-conditioned return contrast and its prompt-fidelity objective, rather than cycling alone.

3 METHOD

We first formalize the failure that PhysReflex addresses.

Let s denote the shared scene context: entities, camera, location, and unchanged action description. Let a^+ be the target event attribute and a^- a contrasting but physically admissible attribute. The target and counterfactual prompts are $c^+ = (s, a^+)$ and $c^- = (s, a^-)$. For example, a^+ may request gradual braking and a^- abrupt braking. The pair is a semantic construction; it is not assumed to factorize the text embedding.

Let $\mathcal{V}_{\text{phys}}(s)$ contain videos that are physically plausible in the shared context, and let $\mathcal{V}_{\text{event}}(s, a^+)$ contain videos that express the requested attribute. The desired output lies in

$$V \in \mathcal{V}_{\text{phys}}(s) \cap \mathcal{V}_{\text{event}}(s, a^+). \tag{1}$$

A video may belong to $\mathcal{V}_{\text{phys}}(s)$ but not to $\mathcal{V}_{\text{event}}(s, a^+)$; this is the failure that our method targets. The primary objective is therefore event realization: the video must express the particular physical outcome requested by c^+. Physical plausibility is a compatibility constraint and may improve as a consequence, but it is not the signal that defines the target. The counterfactual is not a negative prompt for visual quality and is not required to violate a physical law; it is a controlled alternative that makes the desired event distinction explicit.

3.1 OVERVIEW

The method follows a simple inference-time loop. We start from a sparse anchor on the main sampling trajectory, compare a target prompt with a physically valid event alternative, and use the resulting local response to correct the target state. The probe paths are temporary: after the correction, only the target-conditioned trajectory continues. Because the event-forming state can occur at different denoising times, the response gate decides whether another probe is worthwhile.

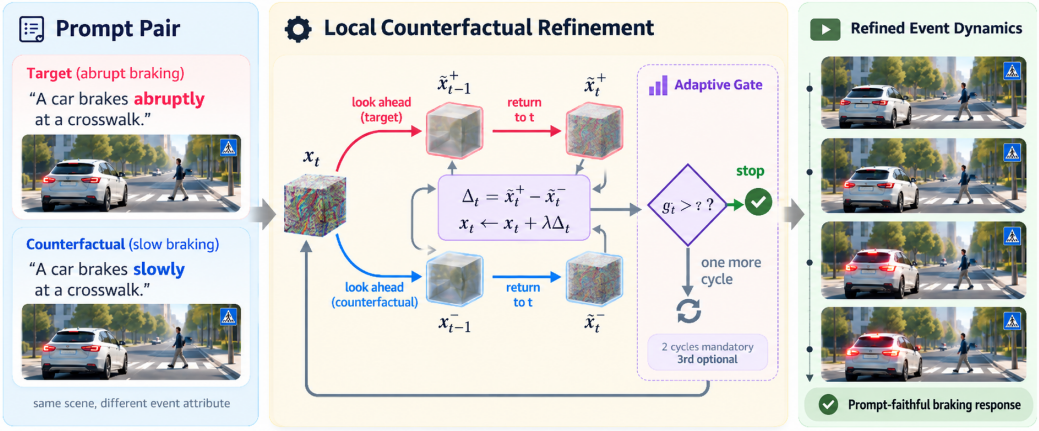

Figure 2: **PhysReflex overview.** A target prompt and a physically valid counterfactual differ in one event attribute while preserving the scene. From the current latent x_t, PhysReflex looks ahead under the target condition, returns from the shared future under both conditions, and uses their matched response difference to update the target state. An adaptive gate stops after the mandatory cycles or requests one additional cycle when the response remains informative, producing a more prompt-faithful event trajectory.

3.2 SAMPLING NOTATION AND MAIN TRAJECTORY

Let z_i be the video latent at noise level σ_i, with $\sigma_0 > \cdots > \sigma_N$. The frozen generator and the base sampler are represented by a guided field $g_\theta(z, \sigma, c)$ and transition $\mathcal{S}_i(z, c)$, following the flow-matching and rectified-flow formulation (Lipman et al., 2023; Liu et al., 2023). For an Euler flow update,

$$\mathcal{S}_i(z, c) = z + h_i g_\theta(z, \sigma_i, c), \qquad h_i = \sigma_{i+1} - \sigma_i < 0. \tag{2}$$

The refinement anchors $\mathcal{A}$ are sparse sampling indices. A refinement loop is a temporary excursion: it does not advance the main scheduler index, and the main trajectory advances exactly once after refinement.

3.3 COUNTERFACTUAL CYCLIC REFINEMENT

At anchor i, initialize $z_i^{(0)} = z_i$. For probe k, compose m target-conditioned transitions over the local noise interval:

$$y_{i,0}^{(k)} = z_i^{(k-1)}, \qquad y_{i,\ell+1}^{(k)} = y_{i,\ell}^{(k)} + h_{i+\ell} g_\theta(y_{i,\ell}^{(k)}, \sigma_{i+\ell}, c^+), \quad \ell = 0, \dots, m-1, \qquad y_i^{(k)} = y_{i,m}^{(k)}. \tag{3}$$

From the common endpoint $y_i^{(k)}$, integrate the same interval back under each condition:

$$r_i^{a,k} = R_i^a(y_i^{(k)}), \qquad a \in \{+, -\}, \tag{4}$$

where the two returns use identical grids, solver, guidance strength, and endpoint. The correction is

$$\Delta_i^{(k)} = r_i^{-,k} - r_i^{+,k}, \qquad z_i^{(k)} = z_i^{(k-1)} + \eta \mathcal{N}(\Delta_i^{(k)}). \tag{5}$$

Here, $\mathcal{N}(\cdot)$ denotes the fixed RMS normalization used for the correction. The probe paths are discarded after the correction. The main sampler then advances once from the refined anchor under the target condition.

The target return serves as a matched reference. Writing $e_i^{+,k} = r_i^{+,k} - z_i^{(k-1)}$ for the target round-trip residual gives

$$r_i^{-,k} - z_i^{(k-1)} = \Delta_i^{(k)} + e_i^{+,k}. \tag{6}$$

Thus the raw response uses the counterfactual displacement after subtracting the same-condition return displacement; the fixed normalization is applied only when updating the main state. Exact inversion is not assumed; the reference simply removes the explicit common target-return term from this decomposition.

Algorithm 1 PhysReflex: local counterfactual cyclic refinement

Require: Target condition c^+, matched event counterfactual c^-, anchor set $\mathcal{A}$, probe length m, correction scale η, gate threshold τ, and probe budget B.

1: Initialize the main latent trajectory $z_0, \dots, z_N$ under c^+.
2: **for** each anchor $i \in \mathcal{A}$ **do**
3: Set $z_i^{(0)} \leftarrow z_i$ and reserve the budget required by later anchors.
4: Advance m temporary transitions from $z_i^{(0)}$ under c^+ to obtain the shared endpoint y_i.
5: Return from y_i over the same interval under c^+ and c^-, obtaining r_i^+ and r_i^-.
6: Compute the event response $\Delta_i \leftarrow r_i^- - r_i^+$ and update $z_i \leftarrow z_i + \eta \mathcal{N}(\Delta_i)$.
7: Compute the target clean-prediction response q_i and discard both temporary return paths.
8: **if** $q_i > \tau$ and B can pay for another probe **then**
9: Repeat the local probe and correction from the updated anchor.
10: **end if**
11: Resume the main sampler with the target condition c^+.
12: **end for**
13: **return** the target-conditioned video decoded from the single resumed main trajectory.

3.4 MECHANISM AND RELATION TO CONDITIONAL GUIDANCE

A direct counterfactual baseline evaluates both conditions at the anchor and applies

$$\Delta_{\mathrm{dir}}(z_i) = g_\theta(z_i, \sigma_i, c^-) - g_\theta(z_i, \sigma_i, c^+), \tag{7}$$

possibly with a scalar step and normalization. Both fields are evaluated once at the same anchor z_i; the direct baseline neither advances to a shared future nor runs target and counterfactual return branches. The sign of the update can be absorbed into the scalar step according to the solver convention. PhysReflex compares returns after target-conditioned evolution. For a local solver interval, substituting the corresponding forward and backward updates yields

$$\Delta_i^{(k)} = h_i \left[g_\theta(y_i^{(k)}, \sigma_{i+1}, c^+) - g_\theta(y_i^{(k)}, \sigma_{i+1}, c^-) \right]. \tag{8}$$

For a finite probe interval, the two backward branches visit condition-dependent intermediate states, so the correction includes a path-dependent response. The distinction is not the subtraction itself, but that it is evaluated after a target-shaped local future and referenced to a matched return. This construction is designed to expose event attributes that are weak at the noisy anchor. For a local interval of size h, let f_+ and f_- denote the target and counterfactual fields. Writing the target endpoint as $y = x + hf_+(x) + O(h^2)$ and expanding the two return fields around x gives

$$\Delta_{\mathrm{PR}}(x) = h[f_+(x) - f_-(x)] + h^2[J_+(x) - J_-(x)]f_+(x) + O(h^3). \tag{9}$$

The first term is the direct conditional contrast up to the solver sign and scale. The second term is a state-response correction: the target field first moves the latent, and the two conditions are re-evaluated at that target-shaped state. For a finite probe interval, repeated field evaluations produce higher-order Jacobian products and path-dependent terms. This is the same structural principle identified in reflection-based diffusion analyses: a finite forward–backward operation can agree with a direct score difference at first order while retaining nonlinear state-response information at finite displacement (Bai et al., 2026). The expansion motivates our event-specific probe, but it is not an improvement guarantee.

The method also differs from CFG. CFG contrasts a conditional field with an unconditional or quality-negative field. PhysReflex contrasts two event conditions that share the scene and differ in one admissible attribute. CFG can improve prompt adherence in general; our hypothesis is that a matched event contrast is more selective for the requested physical event, while any plausibility gain is a secondary consequence rather than the optimization target.

SDG and PhysReflex both use conditional contrast, but their state variables answer different questions. SDG maintains x_t^+ and x_t^- from the initial noise, so $f_+(x_t^+) - f_-(x_t^-)$ contains both the current condition contrast and the history induced by two different trajectories. PhysReflex starts from the actual main state x_t, constructs $y = F_+^{(m)}(x_t)$, and compares $B_-^{(m)}(y)$ with $B_+^{(m)}(y)$. The probe has a shared endpoint and a shared noise interval, and neither branch persists after the correction. This gives PhysReflex a local reference frame: it measures conditional response around the event trajectory that the target prompt itself proposes.

The distinction is a design hypothesis, not a theorem that local probing always dominates global guidance. It predicts that local probing is useful when an event depends on state changes such as contact, impact, braking intensity, or bounce count, and that its cost can be concentrated at sparse anchors. The experiments therefore treat direct conditional difference and global dual-trajectory guidance as required mechanism controls rather than optional implementation variants.

3.5 ADAPTIVE EVENT-TIME ALLOCATION

The event-defining transition does not occur at a fixed denoising index across prompts or samples. A first cycle is mandatory at every selected anchor; the gate then decides whether another probe is useful. Let $D_\theta(z, \sigma, c^+)$ be the predicted clean latent. Under the flow parameterization $g_\theta = \epsilon - x_0$,

$$D_\theta(z, \sigma, c) = z - \sigma g_\theta(z, \sigma, c). \tag{10}$$

Before and after the first cycle, compute $A_i = D_\theta(z_i^{(0)}, \sigma_i, c^+)$ and $B_i = D_\theta(z_i^{(1)}, \sigma_i, c^+)$. Define

$$q_i = \frac{\mathrm{RMS}(B_i - A_i)}{\frac{1}{2}[\mathrm{RMS}(A_i) + \mathrm{RMS}(B_i)] + \varepsilon}. \tag{11}$$

The score measures how much the target clean-video estimate changes after the event-conditioned correction, normalized by its own latent scale. A small q_i indicates that the current state is insensitive to another correction, whereas a large q_i indicates that the target event remains controllable at this anchor. If $q_i > \tau$ and the remaining global budget can pay for an additional probe while reserving later mandatory probes, another local cycle is executed from the updated anchor; otherwise the main target trajectory resumes immediately.

The budget check is part of the decision rather than an after-the-fact accounting step. Let M_i denote the minimum cost of the mandatory probes at later anchors. An optional probe is allowed only when its cost can be paid while maintaining the invariant that at least M_i calls remain reserved. This prevents an early, highly responsive event from consuming the computation required to inspect later events. In this sense, the gate adapts along two axes: it responds to the sample-specific time at which the event is still controllable, and it allocates a fixed inference budget toward those states. The gate allocates computation according to continued controllability of the requested event; it is not a physical error estimator and is not intended to replace the benchmark's plausibility criterion.

4 EXPERIMENTS

4.1 BENCHMARKS AND METRICS

We evaluate the same frozen Wan2.2-A14B text-to-video generator (Team Wan et al., 2025) on two physical-video benchmarks. VideoPhy2 contains action-centric prompts whose physical criteria and event descriptions test whether a clip both behaves plausibly and follows the requested event. PhyWorldBench broadens the physical phenomena and category coverage; we evaluate a frozen 120-scenario seed-0 subset with Anti-Physics excluded. Both benchmarks are evaluated with the target prompt, the generated clip, and the benchmark's criterion prompts. We keep the benchmark identity visible because the two suites probe related but different aspects of physical video generation.

We report three measures. Physical commonsense (PC) is the fraction of clips satisfying the benchmark's physical criteria. Semantic alignment (SA) is the fraction satisfying the prompt's object and event requirements; for this paper, it is the most direct measure of whether the requested physical event is realized. Joint success (JOINT) requires both conditions for the same clip. Thus PCasks "is the motion physically acceptable?", while SAasks "did the model depict the event described by the prompt?" The central claim concerns an improvement in SAthat can occur together with an improvement in PC.

4.2 EXPERIMENTAL PROTOCOL

All methods use the same prompts, seeds, Wan2.2-A14B checkpoint, resolution, frame count, negative prompt, and main denoising schedule. Base is the original sampler with direct CFG. Self-Refine (Jang et al., 2026) is run with its published stochastic perturbation and denoising procedure while the uncertainty mask is disabled. PhysReflex keeps the main sampler fixed and adds only the local counterfactual probes described in Sec. 3. We preserve evaluator identity rather than averaging model-based judges: VideoPhy2 is reported with Qwen3-VL-32B and Gemini 2.5 Flash Lite, and the PhyWorldBench subset uses the same two substitute judges with its official Physics prompts.

To test the mechanism rather than only the final score, the complete evaluation protocol defines three mechanism controls. Direct Counterfactual Guidance evaluates the target and alternative fields at the anchor and applies their difference with the same update budget. SDG implements the global dual-trajectory alternative of Enhancing Physical Plausibility with the same prompt pair. PhysReflex is the full local forward–backward probe, while Direct Counterfactual and SDG are the mechanism controls. These controls separate direct first-order contrast from persistent counterfactual trajectories. The controls are defined with matched prompts, seeds, and model-call budgets so that their results can be interpreted as mechanism comparisons rather than as changes in evaluation protocol.

4.3 RESULTS

Table 1 reports the completed VideoPhy2 results. Under Qwen3-VL-32B, PhysReflex improves semantic alignment from 58.33% for direct CFG to 66.67%, an 8.34-point gain. Under Gemini 2.5

Table 1: VideoPhy2 results. Percentages are higher-is-better; PCis physical commonsense, SAis semantic alignment, JOINTrequires both, and Calls is normalized transformer-call cost with Base $= t$.

Evaluator	Method	PC	SA	JOINT	Calls
Qwen3-VL-32B	Base	82.08%	58.33%	46.67%	t
	Self-Refine (Jang et al., 2026)	82.50%	54.58%	43.75%	$1.5t$
	Direct Counterfactual	83.33%	61.67%	44.58%	$1.5t$
	SDG	82.91%	60.00%	43.33%	$2t$
	PHYSREFLEX	**85.41%**	**66.67%**	**49.17%**	$1.5t$
Gemini 2.5 Flash Lite	Base	87.08%	49.17%	43.33%	t
	Self-Refine (Jang et al., 2026)	91.67%	45.00%	41.67%	$1.5t$
	Direct Counterfactual	90.83%	47.50%	44.17%	$1.5t$
	SDG	87.50%	50.83%	42.50%	$2t$
	PHYSREFLEX	**92.50%**	**56.67%**	**48.33%**	$1.5t$

Table 2: PhyWorldBench results on the frozen 120-scenario seed-0 subset. Content denotes semantic alignment and Physics denotes physical commonsense; Anti-Physics is excluded and Qwen/Gemini are substitute judges.

Evaluator	Method	Content	Physics	Joint	Calls
Qwen3-VL-32B	Base	76.67%	53.33%	49.17%	t
	Self-Refine (Jang et al., 2026)	75.83%	52.50%	51.67%	$1.5t$
	Direct Counterfactual	78.33%	50.83%	50.00%	$1.5t$
	SDG	77.50%	55.00%	52.50%	$2t$
	PHYSREFLEX	**80.83%**	**60.83%**	**55.83%**	$1.5t$
Gemini 2.5 Flash Lite	Base	83.33%	50.83%	49.17%	t
	Self-Refine (Jang et al., 2026)	86.67%	51.67%	50.83%	$1.5t$
	Direct Counterfactual	85.50%	48.33%	49.17%	$1.5t$
	SDG	87.50%	53.33%	52.50%	$2t$
	PHYSREFLEX	**88.33%**	**58.33%**	**54.17%**	$1.5t$

Flash Lite, the corresponding gain is from 49.17% to 56.67%. Physical commonsense also increases, from 82.08% to 85.41% and from 87.08% to 92.50%, respectively. The joint score rises at the same time. These are consistent gains across two evaluators, despite their different absolute calibration.

Self-Refine is a useful additional-computation control. It improves physical commonsense relative to Base under both evaluators, but its semantic alignment and joint score decrease. This contrast is relevant to our claim: generic stochastic refinement can change the sample, yet does not reliably make the generated event more faithful to the prompt. Among the completed Base, Self-Refine, and PhysReflex runs, PhysReflex is the only method that improves all three measures over Base for both evaluators.

Table 2 reports the same metrics for the controlled PhyWorldBench subset using matched generations and the two evaluator passes. The protocol and evaluator identities are fixed in advance, and the subset is kept separate from official full-benchmark leaderboard results.

4.4 ANALYSIS

The VideoPhy2 results support the intended mechanism. The largest and most stable gains are in SA: +8.34 points with Qwen3-VL-32B (58.33%→66.67%) and +7.50 points with Gemini 2.5 Flash Lite (49.17%→56.67%), exceeding the corresponding PCgains of 3.33 and 5.42 points. This pattern points to event-level prompt sensitivity rather than generic smoothing. Direct Counterfactual and SDG improve some individual entries but do not match PhysReflex on the joint outcome, while Self-Refine changes samples without reliably improving event alignment. The two evaluators agree on the direction despite different calibration.

The PhyWorldBench subset tests whether this effect transfers beyond VideoPhy2. Relative to Base, PhysReflex improves Content by 4.16 and 5.00 points, Physics by 7.50 points for both judges, and

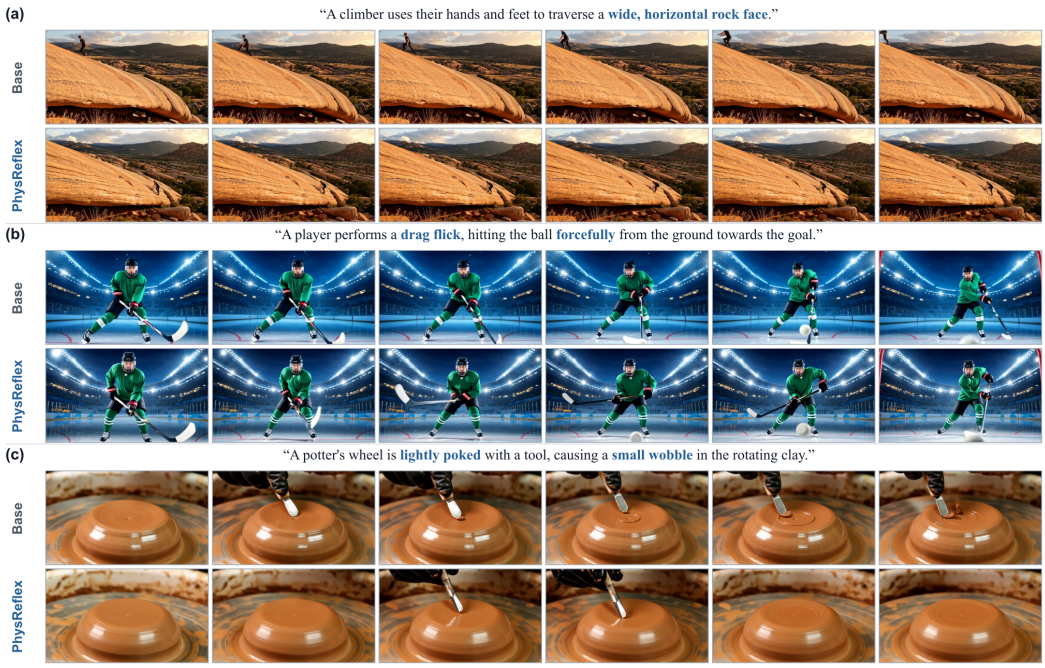

Figure 3: **Qualitative comparison of prompt-faithful physical events.** For each prompt, six temporally ordered frames are shown for Base (top) and PhysReflex (bottom). In (a), the base model fails to coordinate both hands and feet as required for horizontal traversal, whereas PhysReflex realizes the requested contact pattern. In (b), the base model exhibits implausible ball motion during the drag flick, while PhysReflex produces a more coherent hit-and-response sequence. In (c), the base model overreacts to a light poke with excessive clay splashing, whereas PhysReflex preserves the requested small wobble.

Joint by 6.66 and 5.00 points for Qwen and Gemini, respectively. The 76.67% Base Content score under Qwen therefore reflects a real headroom regime: event specification is weaker than the corresponding Gemini score, and the method raises it to 80.83%. Because this is a fixed seed-0 subset with substitute judges, these numbers support transfer under the stated protocol rather than a claim about the official full-benchmark leaderboard. Across both suites, improvements in event alignment and physical criteria occur together, but the metrics remain benchmark proxies rather than direct measurements of acceleration or trajectory geometry.

Figure 3 shows synchronized examples in which PhysReflex better realizes the requested contact or response while preserving the scene. Failures are retained in Appendix Fig. 4; probe calls and optional repeats are included in the cost accounting.

5 Discussion and Conclusion

Limitation. PhysReflex can amplify only distinctions represented by the frozen generator. Results depend on matched counterfactual prompts, and the finite-step response is a mechanism hypothesis rather than a guarantee. Model-based scores are benchmark proxies, not direct measurements of acceleration or contact impulse; failures are retained in Appendix Fig. 4.

Broader impact. More faithful event control can aid visual explanation but also make fabricated demonstrations more convincing. Outputs should be marked synthetic and not treated as evidence.

Conclusion. PhysReflex couples physical admissibility with event selection: its local counterfactual cycle measures response after a target-shaped future, while the gate allocates probes where the event remains controllable. VideoPhy2 gains are clearest in semantic alignment and coexist with improved physical and joint scores.

ETHICS STATEMENT

This work studies text-to-video generation and does not collect personal data or infer private mental states. Generated videos can be misused to depict fabricated physical events as real; we recommend clear provenance, disclosure of synthetic content, and compliance with benchmark and model licenses. The evaluation uses publicly released benchmark prompts and model-based judgments; no human subjects are involved in the reported experiments.

REPRODUCIBILITY STATEMENT

The paper specifies the generator, sampler, resolution, frame count, seed protocol, prompt pairing rule, refinement equations, anchor schedule, budget gate, evaluator prompts, and metric definitions. The accompanying repository records frozen manifests, source hashes, per-video checksums, evaluator responses, and runtime accounting. VideoPhy2 results in Table 1 use matched prompt–seed pairs across methods. PhyWorldBench is reported as a 120-scenario seed-0 subset with substitute evaluators and is explicitly separated from official full-benchmark scores.

AI USE STATEMENT

Generative AI assisted with organization, language editing, consistency checks, and schematic drafting. The authors remain responsible for the method, experimental design, numerical results, citations, and all claims. AI-generated text and figures were reviewed against the implementation and source records before inclusion.

REFERENCES

Lichen Bai, Shitong Shao, Zikai Zhou, Zipeng Qi, Zhiqiang Xu, Haoyi Xiong, and Zeke Xie. Zigzag diffusion sampling: Diffusion models can self-improve via self-reflection. In *International Conference on Learning Representations*, 2025.

Lichen Bai, Masashi Sugiyama, and Zeke Xie. Weak-to-strong diffusion with reflection. In *International Conference on Learning Representations*, 2026.

Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Soumyadip Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Universal guidance for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 843–852, 2023.

Hritik Bansal, Zongyu Lin, Tianyi Xie, Zeshun Zong, Michal Yarom, Yonatan Bitton, Chenfanfu Jiang, Yizhou Sun, Kai-Wei Chang, and Aditya Grover. VideoPhy: Evaluating physical commonsense for video generation. In *International Conference on Learning Representations*, 2025.

Hritik Bansal, Clark Peng, Yonatan Bitton, Roman Goldenberg, Aditya Grover, and Kai-Wei Chang. VideoPhy-2: A challenging action-centric physical commonsense evaluation in video generation. In *International Conference on Learning Representations*, 2026.

Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Blattmann_Align_Your_Latents_High-Resolution_Video_Synthesis_With_Latent_Diffusion_Models_CVPR_2023_paper.html.

Hila Chefer, Uriel Singer, Amit Zohar, Yuval Kirstain, Adam Polyak, Yaniv Taigman, Lior Wolf, and Shelly Sheynin. VideoJAM: Joint appearance-motion representations for enhanced motion generation in video models. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025. URL https://proceedings.mlr.press/v267/chefer25a.html.

Lin Geng Foo, Mark He Huang, Alexandros Lattas, Stylianos Moschoglou, Thabo Beeler, and Christian Theobalt. Physical simulator in-the-loop video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4301–4311, 2026.

Jing Gu, Xian Liu, Yu Zeng, Ashwin Nagarajan, Fangrui Zhu, Daniel Hong, Yue Fan, Qianqi Yan, Kaiwen Zhou, Ming-Yu Liu, and Xin Eric Wang. PhyWorldBench: A comprehensive evaluation of physical realism in text-to-video models. In *International Conference on Learning Representations*, 2026.

Xuyang Guo, Jiayan Huo, Zhenmei Shi, Zhao Song, Jiahao Zhang, and Jiale Zhao. T2VPhysBench: A first-principles benchmark for physical consistency in text-to-video generation. *arXiv preprint arXiv:2505.00337*, 2025. URL https://arxiv.org/abs/2505.00337.

Yutong Hao, Chen Chen, Ajmal Saeed Mian, Chang Xu, and Daochang Liu. Enhancing physical plausibility in video generation by reasoning the implausibility. *arXiv preprint arXiv:2509.24702*, 2025. URL https://arxiv.org/abs/2509.24702.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. URL https://arxiv.org/abs/2207.12598.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models. In *Advances in Neural Information Processing Systems*, 2022.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21807–21818, 2024.

Sangwon Jang, Taekyung Ki, Jaehyeong Jo, Saining Xie, Jaehong Yoon, and Sung Ju Hwang. Self-refining video sampling. In *International Conference on Machine Learning*, 2026.

Hyunmo Kang, Noam Itzhak Levi, Corinna Elena Wegner, Daniel J. Korchinski, and Matthieu Wyart. Sampling data with chains of forward-backward diffusion steps. *arXiv preprint arXiv:2605.27006*, 2026. URL https://arxiv.org/abs/2605.27006.

Tero Karras, Miika Aittala, Tuomas Kynkäänniemi, Jaakko Lehtinen, Timo Aila, and Samuli Laine. Guiding a diffusion model with a bad version of itself. In *Advances in Neural Information Processing Systems*, 2024.

Shubo Lin, Xuanyang Zhang, Wei Cheng, Weiming Hu, Gang Yu, and Jin Gao. MMPhysVideo: Physically plausible video generation through joint RGB-perception modeling. *arXiv preprint arXiv:2604.02817*, 2026. URL https://arxiv.org/abs/2604.02817.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

Fangfu Liu, Hanyang Wang, Yimo Cai, Kaiyan Zhang, Xiaohang Zhan, and Yueqi Duan. Video-T1: Test-time scaling for video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18671–18681, 2025.

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

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations*, 2023.

Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. EvalCrafter: Benchmarking and evaluating large video generation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22139–22149, 2024b.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. DPM-Solver: A fast ODE solver for diffusion probabilistic model sampling in around 10 steps. In *Advances in Neural Information Processing Systems*, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/260a14acce2a89dad36adc8eefe7c59e-Abstract-Conference.html.

Nanye Ma, Shangyuan Tong, Haolin Jia, Hexiang Hu, Yu-Chuan Su, Mingda Zhang, Xuan Yang, Yandong Li, Tommi Jaakkola, Xuhui Jia, and Saining Xie. Scaling inference time compute for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2523–2534, 2025.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2022.

Fanqing Meng, Jiaqi Liao, Xinyu Tan, Quanfeng Lu, Wenqi Shao, Kaipeng Zhang, Yu Cheng, Dianqi Li, and Ping Luo. Towards world simulator: Crafting physical commonsense-based benchmark for video generation. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 43781–43806. PMLR, 2025.

Saman Motamed, Laura Culp, Kevin Swersky, Priyank Jaini, and Robert Geirhos. Do generative video models understand physical principles? In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 948–958, 2026. URL https://arxiv.org/abs/2501.09038.

Saurabh Pathak, Elahe Arani, Mykola Pechenizkiy, and Bahram Zonooz. PhysVid: Physics aware local conditioning for generative video models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 41847–41858, 2026. URL https://arxiv.org/abs/2603.26285.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Peebles_Scalable_Diffusion_Models_with_Transformers_ICCV_2023_paper.html.

Atin Pothiraj, Jaemin Cho, Yue Zhang, Elias Stengel-Eskin, and Mohit Bansal. Physics question scene graph: Fine-grained evaluation of physical plausibility in text-to-video generation. In *European Conference on Computer Vision*, pp. 301–318, 2026.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Rombach_High-Resolution_Image_Synthesis_With_Latent_Diffusion_Models_CVPR_2022_paper.html.

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-A-Video: Text-to-video generation without text-video data. In *International Conference on Learning Representations*, 2023.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=St1giarCHLP.

Kaiyue Sun, Kaiyi Huang, Xian Liu, Yue Wu, Zihan Xu, Zhenguo Li, and Xihui Liu. T2V-CompBench: A comprehensive benchmark for compositional text-to-video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8406–8416, 2025.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai

Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. URL https://arxiv.org/abs/2503.20314.

Zixuan Wang, Yixin Hu, Wen Li, Feng Chen, Yan Liu, Duo Peng, and Yinjie Lei. Physically plausible video generation via visual-semantic chain-of-events conditioning. *arXiv preprint arXiv:2609.00656*, 2026. URL https://arxiv.org/abs/2609.00656.

Chen Henry Wu and Fernando De la Torre. Unifying diffusion models' latent space, with applications to CycleDiffusion and guidance. *arXiv preprint arXiv:2210.05559*, 2022. URL https://arxiv.org/abs/2210.05559.

Yilun Xu, Mingyang Deng, Xiang Cheng, Yonglong Tian, Ziming Liu, and Tommi Jaakkola. Restart sampling for improving generative processes. In *Advances in Neural Information Processing Systems*, 2023.

Qiyao Xue, Xiangyu Yin, Boyuan Yang, and Wei Gao. PhyT2V: LLM-guided iterative self-refinement for physics-grounded text-to-video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18826–18836, 2025.

Xindi Yang, Baolu Li, Yiming Zhang, Zhenfei Yin, Lei Bai, Liqian Ma, Zhiyong Wang, Jianfei Cai, Tien-Tsin Wong, Huchuan Lu, and Xu Jia. VLIPP: Towards physically plausible video generation with vision and language informed physical prior. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12360–12370, 2025a.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. CogVideoX: Text-to-video diffusion models with an expert transformer. In *International Conference on Learning Representations*, 2025b.

Haotian Ye, Haowei Lin, Jiaqi Han, Minkai Xu, Sheng Liu, Yitao Liang, Jianzhu Ma, James Zou, and Stefano Ermon. TFG: Unified training-free guidance for diffusion models. In *Advances in Neural Information Processing Systems*, 2024.

Jiwen Yu, Yinhuai Wang, Chen Zhao, Bernard Ghanem, and Jian Zhang. FreeDoM: Training-free energy-guided conditional diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 23174–23184, 2023.

Ke Zhang, Cihan Xiao, Jiacong Xu, Yiqun Mei, and Vishal M. Patel. Think before you diffuse: Infusing physical rules into video diffusion. *arXiv preprint arXiv:2505.21653*, 2025. URL https://arxiv.org/abs/2505.21653.

Wenliang Zhao, Lujia Bai, Yongming Rao, Jie Zhou, and Jiwen Lu. UniPC: A unified predictor-corrector framework for fast sampling of diffusion models. In *Advances in Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=hrkmlPhp1u.

Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Lulu Gu, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, Yu Qiao, and Ziwei Liu. VBench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. *arXiv preprint arXiv:2503.21755*, 2025. URL https://arxiv.org/abs/2503.21755.

Algorithm 2 Full PhysReflex sampling procedure

Require: target prompt c^+, matched counterfactual prompt c^-, initial noise z_N, decreasing noise grid $\{\sigma_N, \ldots, \sigma_0\}$, anchor set $\mathcal{A}$, probe length m, correction scale η, threshold τ, and total probe budget B.

Ensure: target-conditioned latent z_0 and decoded video V.

Initialize the persistent trajectory from z_N and the target condition c^+.
for $i = N-1, \ldots, 0$ **do**
  Advance the persistent trajectory to anchor z_i using the ordinary target step.
  if $i \in \mathcal{A}$ and the mandatory probe fits the remaining budget **then**
    Set $z_i^{(0)} \leftarrow z_i$ and set the cycle counter $k \leftarrow 0$.
    repeat
      Copy $z_i^{(k)}$ into a private buffer and integrate m temporary target steps to obtain the shared endpoint $y_i^{(k)}$.
      Integrate from $y_i^{(k)}$ back over the same grid under c^+ and c^-, obtaining $r_i^{+,k}$ and $r_i^{-,k}$.
      Compute $\Delta_i^{(k)} \leftarrow r_i^{-,k} - r_i^{+,k}$ and update $z_i^{(k+1)} \leftarrow z_i^{(k)} + \eta \mathcal{N}(\Delta_i^{(k)})$.
      Evaluate the gate response $q_i^{(k)}$ from cached or newly computed target predictions; increment the uncached-call ledger.
      $k \leftarrow k + 1$.
    until $q_i^{(k-1)} \le \tau$ or the reserved budget cannot pay for another cycle
    Replace the persistent anchor by $z_i \leftarrow z_i^{(k)}$ and delete all private probe buffers.
  end if
  Apply exactly one ordinary target step from the (possibly refined) anchor.
end for
Decode z_0 with the frozen VAE to obtain V.
return V, the call ledger, accepted-cycle count, and anchor record.

A Reproducible algorithm and execution flow

This section gives the complete execution order used by the implementation. The main scheduler owns the persistent latent trajectory; every probe uses temporary copies and is discarded before the scheduler resumes. This distinction is essential: a probe is an observation-and-correction operation at an anchor, not an additional branch of the final video.

The order of operations fixes three implementation invariants. First, both return branches start from the same target-shaped endpoint, so their difference is paired at the level of the local state. Second, the persistent scheduler is advanced exactly once per grid interval, regardless of the number of temporary probe calls. Third, the final decode uses only the target-conditioned persistent trajectory; neither the counterfactual latent nor an unaccepted probe path can enter the output.

A.1 Anchor selection and budget reservation

At the beginning of a run, anchors are selected from the denoising grid and stored as integer indices. The mandatory-cycle reservation is computed before the first probe as the cost of one cycle at every later anchor. At anchor i, an optional repeat is admitted only when the remaining budget exceeds this reservation plus the cost of the repeat. Thus the adaptive gate can allocate spare computation to informative anchors while preserving the same mandatory coverage across samples. The implementation records the requested anchors, skipped anchors, accepted cycles, and the reason for every rejected repeat.

A.2 Per-cycle data flow

For each accepted or mandatory cycle, the ledger stores the anchor index, starting sigma, endpoint sigma, target and counterfactual prompt hashes, solver step count, RMS norms of both return states, the normalized correction norm, gate response q_i, and the number of uncached transformer calls.

These fields make it possible to reconstruct whether a reported gain came from a different prompt pair, a different number of cycles, or a changed numerical schedule. The temporary tensors are released after the correction, while the persistent anchor and the ledger entry are retained.

B Detailed derivation

For a local Euler interval, let z be the anchor, $y = z + hg_\theta(z, \sigma_i, c^+)$ the target forward endpoint, and let the return use the next-level field. Then

$$r^+ = y - hg_\theta(y, \sigma_{i+1}, c^+), \tag{12}$$

$$r^- = y - hg_\theta(y, \sigma_{i+1}, c^-), \tag{13}$$

$$r^- - r^+ = h\left[g_\theta(y, \sigma_{i+1}, c^+) - g_\theta(y, \sigma_{i+1}, c^-)\right]. \tag{14}$$

The anchor z cancels algebraically. The direct contrast instead evaluates both fields at (z, σ_i). For multiple steps, write $R^a = R_i^{a,(m)} \circ \cdots \circ R_{i+m-1}^{a,(1)}$. Since R^+ and R^- start from the same y but diverge after their first update, $R^-(y) - R^+(y)$ depends on the sequence of intermediate states. This is the path-dependent part of the cycle.

The correction has a useful first-order interpretation. Let $\Phi_a(y)$ denote the return map under condition a. Around a common endpoint y,

$$\Phi_-(y) - \Phi_+(y) \approx [J\Phi_-(y) - J\Phi_+(y)]\,\delta y + \text{direct field contrast}, \tag{15}$$

where δy is the target-conditioned displacement from the anchor. This expression is explanatory rather than a guarantee of improvement: it shows that the cycle can expose how the two conditions respond after the target event has shaped the local state.

C Budget accounting and implementation

A probe never advances the main scheduler. At anchor i, the implementation records the number of uncached guided calls for the forward target branch, target return, counterfactual return, and post-correction target prediction. The cached target prediction is reused by the gate when possible. The reserve M_i ensures that later anchors can still execute their mandatory cycle. The total additional-call budget is therefore an explicit constraint, not an estimate made after generation.

The completed VideoPhy2 comparison uses identical generator settings across Base, Self-Refine, and PhysReflex. The main generation uses Wan2.2-A14B T2V, 40 denoising steps, the same prompt and seed manifests, 832×480 resolution, 81 frames at 16 FPS, and the same negative prompt. The probe implementation does not modify the model weights, VAE, main schedule, or CFG negative condition.

D Counterfactual construction protocol

For each target prompt, we identify one event attribute and replace only that attribute in the counterfactual. The shared subject, objects, scene, viewpoint, and broad action are retained. Both events must be physically admissible under some context; the pair is not constructed by inserting a violation such as impossible gravity. Examples include gradual versus abrupt braking, one versus two bounces, and weak versus strong impact. We reject pairs that change multiple attributes, replace object identity, or introduce a generic quality defect.

The counterfactual is used only during refinement. The final trajectory is always decoded from the target condition, and the counterfactual video is never presented as the desired output. This separation prevents the counterfactual from becoming a second generation objective.

E Evaluation details and reporting

VideoPhy2 scores are reported as percentages over the evaluated prompt set. PCis the proportion satisfying the benchmark's physical-commonsense criteria, SAis the proportion satisfying the se-

mantic/event criteria, and JOINTis the proportion satisfying both for the same sample. Qwen3-VL-32B and Gemini 2.5 Flash Lite are run with the same sampled frames and criterion prompts; their scores are kept separate because the evaluators have different calibration.

For each sample, the release records method, prompt pair, seed, generation checksum, anchor indices, accepted cycle count, uncached model calls, runtime, and evaluator response. The percentages are descriptive benchmark results; we do not attach claims of statistical significance.

F GLOBAL GUIDANCE COMPARISON

Enhancing Physical Plausibility constructs a target condition and a physics-violating alternative. For a common state x_t, a direct contrast can be written as

$$\delta_t^{\text{direct}} = \epsilon_\theta(x_t, c^+, t) - \epsilon_\theta(x_t, c^-, t). \tag{16}$$

Its SDG variant maintains two trajectories from the same initial noise,

$$x_T^+ \to x_{T-1}^+ \to \cdots \to x_0^+, \qquad x_T^- \to x_{T-1}^- \to \cdots \to x_0^-, \tag{17}$$

and uses a normalized difference such as

$$\widehat{\epsilon}^+ = \epsilon^+ + \lambda \frac{\epsilon^+ - \epsilon^-}{\|\epsilon^+ - \epsilon^-\| + \varepsilon}, \qquad \epsilon^+ = \epsilon_\theta(x_t^+, c^+, t), \quad \epsilon^- = \epsilon_\theta(x_t^-, c^-, t). \tag{18}$$

The difference contains trajectory-history effects by construction. PhysReflex instead uses the actual main state x_t as an anchor, obtains $y = F_+^{(m)}(x_t)$, and computes

$$\delta_t^{\text{FCC}} = B_-^{(m)}(y) - B_+^{(m)}(y). \tag{19}$$

The target and alternative returns share y, the interval, and the return solver; the branches are discarded after the update. These constructions are complementary: SDG asks about two independently evolving event trajectories, while PhysReflex measures a local conditional response around one persistent target trajectory.

G UNIFIED MATHEMATICAL COMPARISON OF GUIDANCE MECHANISMS

We now place CFG, direct counterfactual guidance, SDG, and PhysReflex in one notation. Let x_t be a latent at noise level σ_t, let $u_\theta(x_t, t, c)$ denote the model prediction used by the sampler (noise, velocity, or an equivalent solver field), and let $u_\theta^0(x_t, t)$ denote the unconditional prediction. The exact parameterization does not affect the comparison below.

Classifier-free guidance. CFG combines one conditional evaluation and one unconditional evaluation at the *same* state:

$$u_{\text{CFG}}(x_t, t, c^+) = u_\theta^0(x_t, t) + w\left[u_\theta(x_t, t, c^+) - u_\theta^0(x_t, t)\right], \tag{20}$$

where w is the guidance scale. Thus CFG changes the target field before the sampler step. It does not introduce a second event condition, does not construct a shared future, and does not subtract a counterfactual return. With a one-step solver map $S_t(x; u)$, the update is simply $x_{t-1} = S_t(x_t; u_{\text{CFG}})$.

Direct counterfactual difference. Given a matched pair (c^+, c^-), the direct baseline evaluates both event conditions at the current anchor and adds their difference to the target field:

$$d_{\text{dir}}(x_t) = u_\theta(x_t, t, c^+) - u_\theta(x_t, t, c^-), \tag{21}$$

$$u_{\text{dir}}(x_t, t) = u_{\text{CFG}}(x_t, t, c^+) + \lambda \mathcal{N}\left(d_{\text{dir}}(x_t)\right), \tag{22}$$

$$x_{t-1}^{\text{dir}} = S_t\left(x_t; u_{\text{dir}}\right). \tag{23}$$

The two evaluations share the anchor x_t, but there is no target-conditioned look-ahead and no return map. Consequently, the signal is a first-order, same-state conditional contrast; its meaning is tied to whatever event information is already visible at the noisy anchor.

SDG-style persistent dual trajectories. SDG maintains two persistent trajectories from the same initial noise. Writing S_t^+ and S_t^- for the target and alternative solver steps,

$$x_{t-1}^+ = S_t^+(x_t^+; c^+), \qquad x_{t-1}^- = S_t^-(x_t^-; c^-), \tag{24}$$

$$d_t^{\mathrm{SDG}} = u_\theta(x_t^+, t, c^+) - u_\theta(x_t^-, t, c^-), \tag{25}$$

$$\widetilde{u}_t^+ = u_\theta(x_t^+, t, c^+) + \lambda \mathcal{N}\left(d_t^{\mathrm{SDG}}\right). \tag{26}$$

The alternative state x_t^- is itself the result of all earlier alternative updates. Expanding the state recursively gives $x_t^- = S_{t+1}^- \circ \cdots \circ S_T^-(x_T)$, so d_t^{SDG} mixes the current conditional contrast with the history of a second trajectory. This is appropriate when the goal is to maintain separation from a globally undesirable or physics-violating trajectory, but it is not a local response measurement around the actual target trajectory.

PhysReflex local counterfactual cycle. PhysReflex keeps one persistent target state x_t and creates only temporary probe states. Let $F_{t\to t-m}^+$ be m target-conditioned forward solver steps and let $B_{t-m\to t}^a$ be the corresponding return map under condition $a \in \{+,-\}$. The shared target-shaped future and the two returns are

$$y_t = F_{t\to t-m}^+(x_t), \tag{27}$$

$$r_t^+ = B_{t-m\to t}^+(y_t), \qquad r_t^- = B_{t-m\to t}^-(y_t), \tag{28}$$

$$d_t^{\mathrm{PR}} = r_t^- - r_t^+. \tag{29}$$

The persistent target state is then corrected and advanced once:

$$x_t^{\mathrm{ref}} = x_t + \eta \mathcal{N}\left(d_t^{\mathrm{PR}}\right), \tag{30}$$

$$x_{t-1} = S_t^+(x_t^{\mathrm{ref}}; c^+). \tag{31}$$

Both return maps start from the identical endpoint y_t, use the same interval and numerical solver, and are discarded after d_t^{PR} is computed. Therefore the PhysReflex contrast is local in trajectory space but finite-step in its event response: it measures how the two admissible conditions react after the target condition has shaped the intermediate state.

The four mechanisms differ along three independent axes: (i) whether an unconditional or event-counterfactual condition is used, (ii) whether the compared evaluations share only an anchor or also a target-shaped future, and (iii) whether the alternative branch persists in the final trajectory. CFG uses an unconditional reference and no alternative trajectory; direct difference uses a same-anchor event reference; SDG uses a persistent dual trajectory; PhysReflex uses a shared target-shaped endpoint and temporary return branches. This decomposition is the intended mechanism distinction, independent of the particular solver parameterization.

H Finite-step expansion and higher-order response

For one Euler interval of size h, let f_+ and f_- be the target and alternative fields. The target forward endpoint is

$$y = x + h f_+(x). \tag{32}$$

The two returns are $x^{\mathrm{cf}} = y - h f_-(y)$ and $x^{\mathrm{tar}} = y - h f_+(y)$, giving

$$\Delta_{\mathrm{FCC}} = h[f_+(y) - f_-(y)]. \tag{33}$$

Expanding around x,

$$f_\pm(y) = f_\pm(x) + h J_\pm(x) f_+(x) + O(h^2), \tag{34}$$

$$\Delta_{\mathrm{FCC}} = h[f_+(x) - f_-(x)] \tag{35}$$

$$+ h^2[J_+(x) - J_-(x)] f_+(x) + O(h^3). \tag{36}$$

The first term is the direct conditional difference up to the signed step. The second term is a state-response correction: the target field first moves the state, and the two fields are then re-evaluated after that movement. For $m > 1$, the corresponding expression contains Jacobian products and field evaluations along the two return paths. This expansion establishes a structural distinction, not an improvement guarantee. Whether the higher-order response helps event fidelity is an empirical question.

I IMPLEMENTATION SPECIFICATION

I.1 MODEL AND TENSOR CONVENTIONS

The main generator is Wan2.2-A14B T2V with the released two-expert transformer and VAE. We retain the model's native latent layout and dtype policy: bfloat16 transformer computation and float32 VAE decoding. The main run uses 832×480 spatial resolution, 81 frames, 16 FPS, 40 denoising steps, guidance scales 4.0 and 3.0 for the two experts, and the fixed negative prompt recorded in the experiment manifest. The noise schedule is indexed by decreasing σ; a refinement "forward" step means moving toward lower noise along this schedule, not reversing the video in time.

I.2 PROBE ISOLATION

The probe owns private latent copies and does not call the main scheduler. Forward and backward branches use the same sigma grid, solver, endpoint, CFG settings, and random-state policy. The counterfactual branch is deterministic conditional computation; its output is discarded after Δ_i. The main trajectory advances once after the probe. This isolation prevents probe calls from consuming the main random stream or accidentally advancing the scheduler twice.

I.3 CACHING AND CALL ACCOUNTING

At each anchor, the target prediction at the original state is cached. The first cycle includes one target forward excursion, one target return, and one counterfactual return, followed by a target re-evaluation for the response gate. If the gate accepts a second cycle, only uncached calls are added to the ledger. The ledger stores raw transformer calls, accepted cycles, wall-clock time, and peak allocated memory for every sample.

J ADAPTIVE GATE AND CALIBRATION

The gate threshold τ is fixed in the run configuration before evaluation. The denominator in Eq. equation 11 uses the mean RMS scale of the two clean predictions, with ε fixed in the configuration file. The reserve M_i is computed from the number of later anchors that still require their first cycle. Consequently, a high-response early anchor cannot consume the budget reserved for mandatory later probes. We report accepted-cycle counts without interpreting q_i as a physical correctness score.

K EVALUATION PROMPTS AND SCORING

For VideoPhy2, the physical-commonsense evaluator receives the benchmark's physical criteria and sampled frames; the semantic evaluator receives the caption and the same frame list. A sample passes PCwhen its physical score meets the benchmark pass threshold, passes SAwhen the semantic/event score meets its threshold, and passes JOINTwhen both hold. Qwen3-VL-32B and Gemini 2.5 Flash Lite are run separately with temperature zero. We retain raw JSON responses, request hashes, frame indices, and video hashes so that a score can be audited without regenerating the clip.

For the PhyWorldBench subset, we use the official Physics prompts and eight PNG frames at original resolution. The subset contains 120 seed-0 scenarios and excludes Anti-Physics. Because Qwen and Gemini replace the original default judge, the resulting numbers are disclosed as substitute-judge subset results. They should not be conflated with the official full-benchmark leaderboard.

L FAILURE CASES AND SCOPE

There are three expected failure modes. First, the target and counterfactual prompts may be too similar in the generator's representation, producing a negligible return difference. Second, the generator may represent neither event, in which case a larger contrast can amplify artifacts instead of recovering the requested motion. Third, a lexical change may unintentionally alter an entity, camera, or

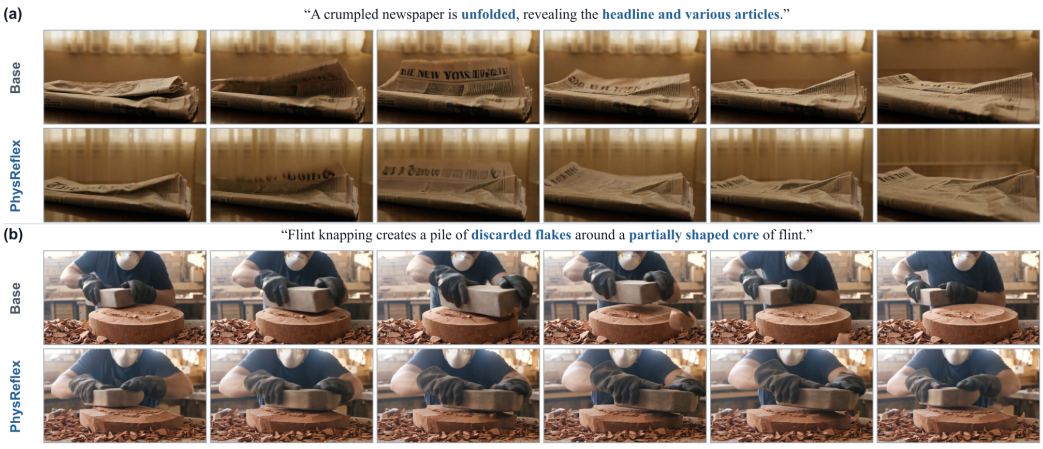

Figure 4: **Representative failure cases.** For each prompt, six temporally ordered frames are shown for Base (top) and PhysReflex (bottom). In (a), PhysReflex unfolds the newspaper but inverts the revealed page content, whereas the base output preserves its orientation. This shows that the correction can amplify an appearance change unrelated to the requested event. In (b), stones appear abruptly in both outputs rather than emerging through a coherent flint-fracture process; PhysReflex provides no improvement when the required dynamics are not represented by the frozen generator.

scene attribute; this violates the matched-pair assumption. We retain such failures in the audit rather than filtering them after seeing the output.

A fourth limitation concerns evaluator sensitivity. A frame-based judge may recognize that a car is braking without reliably measuring the rate of deceleration. For this reason, the paper interprets the VideoPhy2 semantic score as prompt/event adherence under the benchmark protocol, not as a direct physical measurement of acceleration. Future paired-event tests should add trajectory-sensitive human or automated measurements for attributes such as braking rate and bounce count.

M PER-EXAMPLE RELEASE SCHEMA

Each row in the released manifest contains a stable example identifier, target prompt, counterfactual prompt, benchmark category, seed, assigned GPU, and source checksum. Each generated-video record contains the output path, video checksum, dimensions, frame rate, frame count, raw transformer-call count, wall-clock time, and peak memory. Each evaluator record contains the model identifier, prompt hash, sampled frame indices, raw response, parsed score, and usage metadata. This schema supports paired analysis and prevents a changed video from silently being scored under an old response.