VISIONCREATOR-P1: CLOSING THE LOOP FOR LONG-HORIZON PHYSICAL VISUAL GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-horizon physical visual generation requires modeling how a scene should evolve under physical laws, not just how each frame looks. However, open-loop models lack explicit physical feedback: small local violations accumulate across steps and progressively derail multi-step trajectories. To address this, we present **VisionCreator-P1**, a closed-loop physics-informed agent that casts multi-image trajectory synthesis as a process of verified state transitions. At inference, the agent alternates milestone planning, next-state proposal, and transition-level physics reflection, committing a state only after iterative verification against the given query. For training, we address the credit assignment bottleneck that destabilizes long-horizon agentic reinforcement learning by decomposing optimization into short, prefix-conditioned blocks. Each block branches from a shared verified prefix and competes under a physics-aware hierarchical reward, thereby isolating local transition errors from accumulated drift and enabling stable policy learning. Experiments show that VisionCreator-P1 achieves superior performance on physical accuracy and temporal coherence, establishing closed-loop local transition verification as a core paradigm for physically grounded, long-horizon generation.

1 INTRODUCTION

The value of physical visual generation lies in modeling how a scene should evolve, not merely how each frame should appear. This capability is essential for a wide range of real-world applications, including embodied AI (Wang et al., 2024b), interactive simulators (Liu et al., 2024; Bruce et al., 2024), and autonomous driving (Wang et al., 2024a; Hu et al., 2023). Although recent video generators achieve impressive visual fidelity, they continue to struggle with long-horizon physical scenarios (Zheng et al., 2025): fluids evolve inconsistently, rigid bodies teleport, and collisions violate energy constraints. We argue that a fundamental limitation of current video generation is the entanglement of high-level physical reasoning with dense pixel interpolation. Physical laws govern discrete state transitions, yet smooth pixel changes in dense video frames often mask structural physical errors. To achieve physically grounded generation, we therefore focus on a foundational task: **Multi-image Physical Trajectory Generation**. By producing sparse, causally linked key states, we can evaluate transition-level constraints while separating state-level reasoning from pixel-level interpolation, which aligns naturally with modern keyframe-based video synthesis. Here, *long-horizon* refers to the number of causally linked sparse transitions.

Prior efforts to inject physical plausibility address only part of this gap. Data-driven generators absorb broad visual priors from large-scale video corpora (Wang et al., 2025b; Li et al., 2025a), but lack explicit physical verification. Simulator-augmented methods incorporate Newtonian structure via external engines (Liu et al., 2024; Lv et al., 2024), yet they typically reduce open-world objects to primitive shapes and are sensitive to specific parameter choices. In particular, while MLLM-based pipelines have shown considerable success in planning causal event sequences through semantic reasoning (Huang et al., 2025; Wang et al., 2026), their generation remains open-loop—a single early error cascades through all subsequent states, leaving no opportunity for correction.

We term this open-loop failure mode *Global Causal Drift*. Specifically, a long-horizon physical trajectory is not a collection of independent images; it is a chain of causally connected state transitions. Small errors—a misplaced object, an invalid contact—become the initial context for subsequent steps, snowballing into global causal failures. Consequently, a physical generator must verify each transition

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.

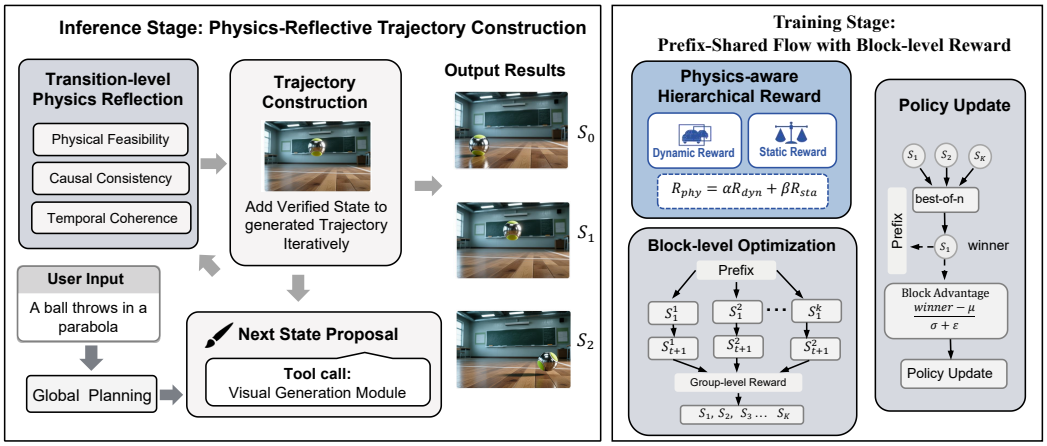


Figure 1: Framework of VISIONCREATOR-P1. VisionCreator-P1 combines PRTC for closed-loop state generation through physics reflection with PFBR for physics-aware long-horizon optimization.

before proceeding. Yet frame-by-frame verification of dense continuous video is computationally prohibitive and largely redundant. In contrast, generating a sparse sequence of multi-image key states makes rigorous, closed-loop reasoning highly feasible.

To this end, we introduce **VisionCreator-P1**, a closed-loop generation agent that models physical phenomena as progressively verified state transitions. Its effectiveness rests on a closed-loop inference pipeline and a tailored reinforcement learning (RL) strategy (Fig. 1). For inference, we propose *Physics-Reflective Trajectory Construction* (PRTC). The agent decomposes a query into an overall goal and intermediate sub-goals; it then predicts and renders the next state, admitting it to the trajectory only after it passes a transition-level physical verification step. For training, we further design *Prefix-Shared Flow with Block-level Reward* (PFBR) to address the credit assignment bottleneck that destabilizes long-horizon agentic RL (Lightman et al., 2023; Zhou et al., 2024; Samarinas et al., 2026). Concretely, instead of relying on full-trajectory feedback, PFBR divides the optimization into short blocks. Candidates branch from a shared, physically verified prefix and compete via a physics-aware hierarchical reward. This block-level evaluation avoids the high variance caused by sparse terminal rewards over long sequences, thereby yielding stable policy learning.

To support this framework, we develop a complementary data suite. We train on **PhysData-M**, a new corpus of 900 complex multi-step physical transitions, mixed with $1,000$ general visual samples from VCR-RL (Lai et al., 2026b) to preserve base generation capabilities. For evaluation, we introduce **PhysBench-M**, a 900-sample benchmark covering three physical scenarios: *Continuum* (fluid dynamics, thermodynamics, waves), *Translation* (rigid-body kinematics), and *Collision* (two-body interactions). Each sample provides precise physical checkpoints for fine-grained, VLM-judged evaluation. Experiments demonstrate strong performance across the evaluated PhysBench-M scenarios while simultaneously improving general visual generation on VCR-Bench.

Our primary contributions are summarized as follows:

- We formulate multi-image physical trajectory generation as a foundational task for physically grounded visual generation, decoupling causal state reasoning from dense pixel interpolation.
- We identify *Global Causal Drift* as a critical failure mode of open-loop generation. To address it, we propose **VisionCreator-P1**, a closed-loop agent that verifies each state transition before proceeding and stabilizes long-horizon agentic RL by decomposing the optimization into short blocks. Experiments show strong performance under the evaluated physical and temporal criteria.
- We release **PhysData-M** and **PhysBench-M**, a training corpus and benchmark covering three physical scenarios with precise checkpoints, to facilitate research on physically grounded generation.

2 RELATED WORKS

Physics-Informed Visual Generation. To enhance physical plausibility in video generation, existing approaches generally fall into three paradigms. *The first paradigm* enforces physical consistency

through explicit simulations. Several works (Liu et al., 2024; Lv et al., 2024; Foo et al., 2026; Liu et al., 2026; Romero et al., 2025) approximate complex objects with geometric primitives or integrate simulator constraints to render rigid-body dynamics. However, mapping diverse open-world visuals to strict simulator constraints often restricts their generalization. To handle open-domain physical phenomena, *the second paradigm* injects physical knowledge into data-driven architectures (Wang et al., 2025b; Li et al., 2025a; Zhang et al., 2025b; Yin et al., 2024). For instance, WISA (Wang et al., 2025b) and PhysHPO (Chen et al., 2025) guide models to implicitly internalize physical laws from decomposed principles, while VideoREPA (Zhang et al., 2025b) leverages physical priors from foundation models. *The third paradigm* explores explicit reasoning over causal and temporal structures. Methods such as VChain (Huang et al., 2025), VLIPP (Yang et al., 2025a), and DiffPhy (Zhang et al., 2025a) leverage structured logical reasoning (e.g., Chain-of-Thought) to guide generation around key events and their causal dependencies (Wang et al., 2026; Hao et al., 2025; Xue et al., 2025). By maintaining consistency through structured constraints rather than pure data priors, these approaches effectively enforce robust physical and temporal grounding.

Visual Generation Agent. To overcome the inherent limitations of conventional generative models in handling complex instructions and long-horizon dynamics, research has increasingly pivoted toward MLLM-driven agentic paradigms (Lai et al., 2026b; Wang et al., 2024c; Li et al., 2025b; Lu et al., 2025; Lai et al., 2026a; Qin et al., 2024; Tran et al., 2025; Wang et al., 2025a). For instance, GenArtist (Wang et al., 2024c) employs an MLLM as a centralized agent to decompose complex generative tasks via a tree-structured planning process. Similarly, MCCD (Li et al., 2025b) introduces multi-agent collaboration for precise compositional generation, while Proactive Agents (Lu et al., 2025) explores multi-turn interactions to align generation with human intent under uncertainty. More recently, the field has evolved toward native end-to-end agents: VisionCreator (Lai et al., 2026a) pioneers a framework that unifies understanding, thinking, planning, and creation capabilities for highly autonomous generation trajectories. Building upon this, VisionCreator-R1 (Lai et al., 2026b) introduces reflection mechanisms, empowering the agent to self-correct based on intermediate visual outputs. However, when generating long-horizon sequences governed by physical laws, existing agents predominantly operate in an open-loop manner. Lacking transition-level physical feedback and progressive verification, they inevitably accumulate microscopic local deviations.

Credit Assignment in Agentic RL. Extending Reinforcement Learning (RL) to long-horizon reasoning and agentic tasks is notoriously hindered by the dual challenges of sparse rewards and high variance in credit assignment (Zhang, 2026). While process reward models (PRMs) (Lightman et al., 2023) provide fine-grained supervision to mitigate reward sparsity, conventional step-level advantage estimation over entire sampled trajectories often exacerbates policy gradient variance due to the accumulation of stochasticity in subsequent steps (Samarinas et al., 2026). Recent advances have sought to bound this variance through various mechanisms, such as implicitly mapping multi-turn preferences via advantage-level fusion (Liu et al., 2025) and adapting baseline estimations (Pan et al., 2026; Gong et al., 2026). Adapting Group Relative Policy Optimization (GRPO) (Shao et al., 2024) for agentic RL, our PFBR changes the sampling topology by anchoring candidate continuations to a shared, verified prefix. This shared-prefix construction compares local continuations under a common context; its optimization benefit is evaluated empirically rather than treated as a formal variance-reduction guarantee.

3 METHODOLOGY

3.1 PROBLEM FORMULATION

Given a user query $\mathcal{Q}$ specifying a physical scenario, our objective is to generate a temporally coherent keyframe trajectory $\tau = \{S_0, S_1, \ldots, S_T\}$, where $S_t \in \mathcal{S}$ denotes the keyframe at step t. We formulate this generation process as sampling from an autoregressive conditional policy π_θ:

$$\pi_\theta(\tau \mid \mathcal{Q}) = \prod_{t=1}^{T} \pi_\theta(S_t \mid S_{<t}, \mathcal{Q}). \tag{1}$$

Let $\mathcal{T}^* : \mathcal{S} \to \mathcal{S}$ denote an ideal physical transition operator. For a fixed initial reference state $S_0^\star$, define the ideal trajectory by $S_t^\star = \mathcal{T}^*(S_{t-1}^\star)$. We measure states with a distance $\mathcal{D}$ on a physics-relevant representation (e.g., pose, contact, identity, and material attributes), rather than raw

RGB space. Within each locally valid physical regime, assume $\mathcal{T}^*$ is locally, piecewise Lipschitz: $\mathcal{D}(\mathcal{T}^*(S), \mathcal{T}^*(S')) \leq L\mathcal{D}(S, S')$ for nearby states. Contacts, occlusions, and phase changes may separate such regimes.

Definition (Global Causal Drift). Let $e_t = \mathbb{E}[\mathcal{D}(S_t, S_t^\star)]$ be the deviation from the ideal trajectory, let $e_0 = \mathbb{E}[\mathcal{D}(S_0, S_0^\star)]$, and let $\epsilon_t = \mathbb{E}[\mathcal{D}(S_t, \mathcal{T}^*(S_{t-1}))]$ be the one-step transition error. Under the local bound and the triangle inequality,

$$e_t \leq \epsilon_t + Le_{t-1}, \tag{2}$$

and hence

$$e_T \leq L^T e_0 + \sum_{t=1}^{T} L^{T-t}\epsilon_t. \tag{3}$$

If the ideal trajectory lies on the physical manifold $\mathcal{M}$, i.e., $S_t^\star \in \mathcal{M}$, then e_T upper-bounds the distance to that reference trajectory and therefore provides a conditional bound on distance to $\mathcal{M}$. For $L < 1$, propagated errors contract; for $L = 1$, they accumulate additively; and for $L > 1$, earlier errors are amplified, illustrating Global Causal Drift. The $L > 1$ case is illustrative rather than a requirement for all regimes.

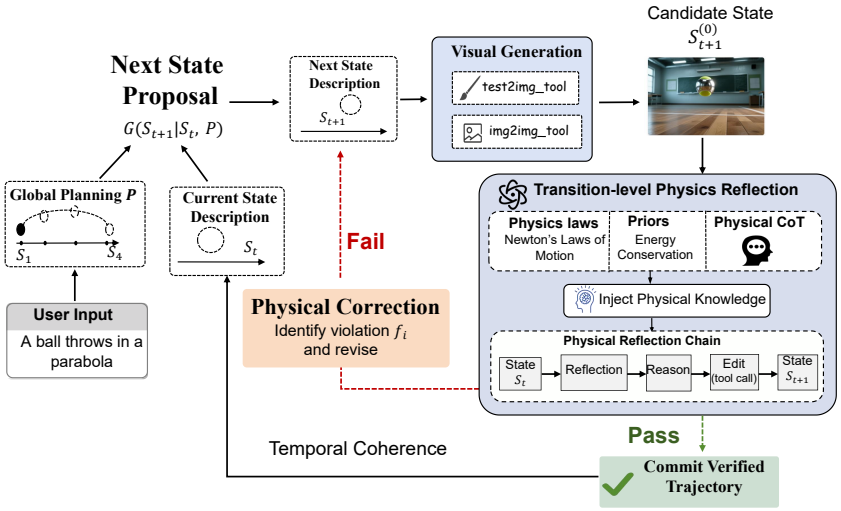


Figure 2: Physics-Reflective Trajectory Construction. (Sec. 3.2). PRTC iteratively plans the global trajectory, proposes the next state, and verifies each transition through physics reflection, revising failed candidates before committing them to the trajectory.

3.2 Physics-Reflective Trajectory Construction

VisionCreator-P1 extends agentic visual generation to long-horizon multi-image physical trajectory construction. Rather than generating the entire sequence in one pass, it constructs the trajectory through a closed-loop process that alternates between global planning, next-state proposal, and physics-informed reflection. Fig. 2 illustrates the inference procedure.

Given a textual query $\mathcal{Q}$ describing a physical scenario, the agent recursively constructs a trajectory $\tau = \{S_0, S_1, \ldots, S_T\}$:

$$S_{t+1} = \mathcal{A}_\theta(S_t, P, \mathcal{Q}), \tag{4}$$

where S_t denotes the current state, P is the global plan inferred from $\mathcal{Q}$, and $\mathcal{A}_\theta$ is the agentic transition operator. The operator consists of three stages:

Terminal-guided trajectory planning. Given the query $\mathcal{Q}$, the agent first builds a global plan P that describes the overall motion and key intermediate steps. This enables the agent to stay consistent over time while avoiding a fixed layout for each frame.

Next-state proposal. At each step, the agent predicts the next physical state conditioned on the current state S_t and the global plan P. Instead of directly rendering the next image, it first generates a textual next-state description, which is then decoded by visual generation tools into an initial candidate $S_{t+1}^{(0)}$. This separation makes the physical transition easier to inspect and revise.

Transition-level physics reflection. The agent verifies whether the candidate transition from S_t to $S_{t+1}^{(0)}$ is physically consistent. If a violation f_i is detected, such as an implausible motion trend, contact relation, or temporal ordering, the agent revises the candidate and repeats the verification. Once the candidate passes the check, it is committed as S_{t+1}. This closed-loop verification reduces local physical errors and prevents them from accumulating into globally inconsistent trajectories.

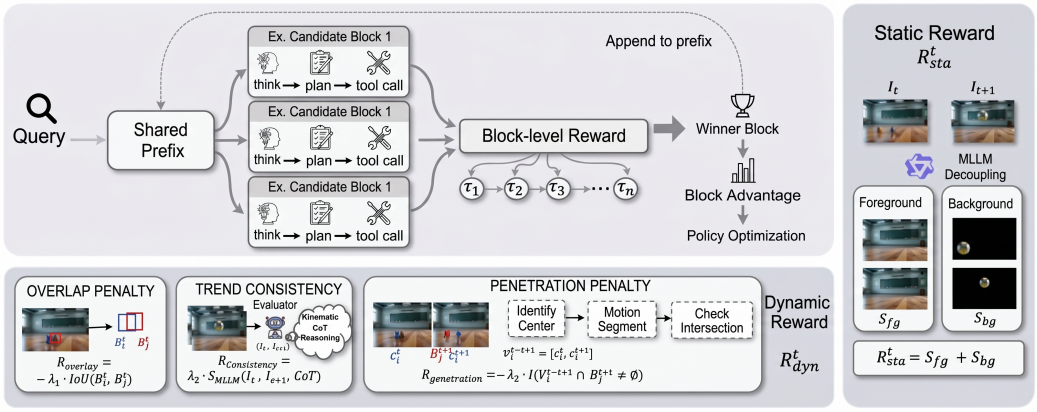


Figure 3: Training with Prefix-Shared Flow and Block-level Reward (Sec. 3.3). PFBR branches candidate continuations from a shared prefix and selects the best block using physics-aware dynamic and static rewards.

3.3 Training via Physics-Reflective Optimization

As illustrated in Fig. 3, training consists of a physics-aware hierarchical reward and the PFBR optimization strategy. The reward provides local feedback on transition dynamics and visual consistency, while PFBR compares short continuations from a shared prefix to provide a localized credit-assignment signal.

Physics-Aware Hierarchical Reward. We decompose the transition reward into dynamic and static terms:

$$R^{(t)} = \alpha R^{(t)}_{\text{dyn}} + \beta R^{(t)}_{\text{stat}}. \tag{5}$$

Dynamic Interaction Reward. Let I_t denote the generated frame, B_i^t the bounding box of rigid object i, and $\mathbf{c}_i^t$ its center. Let $\mathcal{R}$ denote the set of valid rigid–rigid object pairs. We define

$$\begin{aligned} R^{(t)}_{\text{dyn}} = & -\lambda_1 \sum_{(i,j)\in\mathcal{R}\setminus\mathcal{C}_t} \text{IoU}(B_i^t, B_j^t) \\ & -\lambda_2 \sum_{(i,j)\in\mathcal{R}\setminus\mathcal{C}_t} \mathbb{I}\big(\mathcal{V}_i^{t\to t+1} \cap B_j^{t+1} \neq \emptyset\big) \\ & +\lambda_3 S_{\text{MLLM}}(I_t, I_{t+1}; \text{CoT}), \end{aligned} \tag{6}$$

where $\mathcal{V}_i^{t\to t+1} = [\mathbf{c}_i^t, \mathbf{c}_i^{t+1}]$, the λ's are non-negative, and $\mathcal{C}_t$ contains pairs in an annotated or prompt-specified contact window. Geometric penalties are therefore applied only to valid rigid–rigid pairs outside expected contact windows. Continuum–continuum and continuum–rigid interactions use the MLLM consistency term only, since bounding-box penalties are not well-defined for fluids, smoke, heat, or other deformable processes.

For Continuum scenarios, the MLLM term is evaluated using five localized heuristic checkpoints: initial state and stage completeness, physical-law-consistency cues, object or material consistency,

spatial or environmental continuity, and temporal or causal ordering. These checkpoints provide localized evidence but are not substitutes for numerical simulation or exact conservation-law verification.

Static Consistency Reward. As depicted in the bottom-left of Fig. 3, the static reward preserves object identity and background stability. Utilizing an MLLM decoupling strategy, we explicitly separate the visual transition into foreground entities and background canvases:

$$R_{\text{stat}}^{(t)} = \mu_1 \, \mathcal{S}_{\text{fg}}(I_t, I_{t+1}) + \mu_2 \, \mathcal{S}_{\text{bg}}(I_t, I_{t+1}), \tag{7}$$

where $\mathcal{S}_{\text{fg}}$ and $\mathcal{S}_{\text{bg}}$ measure foreground semantic coherence and background invariance based on the MLLM evaluator, respectively. This structural decomposition provides dense learning signals without relying on brittle, hand-crafted annotations.

Prefix-Shared Flow with Block-level Reward (PFBR). Full-trajectory GRPO rollouts are noisy because early errors contaminate subsequent states. PFBR instead samples K candidate response blocks from one shared verified prefix. A block is a complete semantic agent action unit, such as a think–plan–tool-call sequence; it includes the proposed state description, rendering call, and transition-verification output when those actions occur in the current round. Each candidate is sampled from the old policy, evaluated with the hierarchical block reward, and assigned $A_k = (r_k - \bar{r})/(\text{std}(r) + \varepsilon)$ within the K-candidate group. The candidate with the largest A_k is selected as the winner and deterministically appended to the prefix before the next round. The policy ratio ρ in Eq. 8 is computed for the tokens in the sampled block under the updated and old policies; discarded candidates provide the within-group baseline but are not appended. Thus, PFBR is a training-time shared-prefix sampling and optimization procedure, whereas PRTC is the inference-time transition-construction procedure.

The policy is optimized with the clipped objective:

$$\mathcal{J}_{\text{PFBR}}(\theta) = \mathbb{E}\left[\min\left(\rho A, \text{clip}(\rho, 1-\epsilon, 1+\epsilon)A\right) - \beta \, \text{KL}\left(\pi_\theta \| \pi_{\text{ref}}\right)\right], \tag{8}$$

where ρ is the policy-ratio term and A is the within-group normalized advantage. The VLM reward is a heuristic training signal rather than a simulator-level guarantee; geometric checks, multiple Continuum checkpoints, strict data separation, and held-out evaluation are used to reduce evaluator-specific reward hacking.

4 Experiments

4.1 Experimental Setup

Training Data. We trained VisionCreator-P1 on **PhysData-M**, a custom dataset containing 900 samples of complex, multi-step physical interactions. To prevent the model from losing its general image generation skills, we also included 1,000 high-quality visual samples from VCR-RL (Lai et al., 2026b) during the training process.

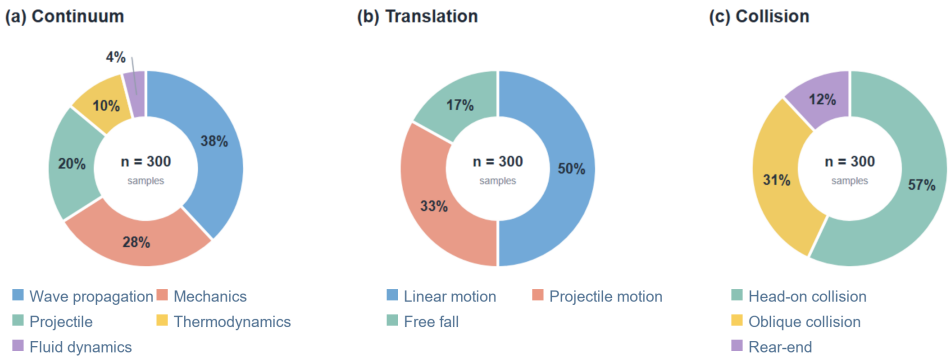


Figure 4: Distribution of PhysBench-M across Continuum, Translation, and Collision categories.

Benchmarks and Evaluation. We evaluate multimodal agents on two complementary benchmarks that probe physical dynamics generation and general visual reasoning.

Table 1: Agent and tool baselines evaluated with Qwen3-VL-235B on PhysBench-M and VCR-Bench.

Model	Type	PhysBench-M			VCR-Bench		
		Continuum	Translation	Collision	Single-Img	Multi-Img	Img2Img
Qwen-Image	Tool	0.142	0.126	0.083	0.795	0.561	0.632
Qwen3-VL-32B-Reflection	Agent	0.243	0.117	0.094	0.842	0.715	0.746
Gemini-2.5-Pro	Agent	0.310	0.224	0.110	0.844	0.783	0.737
VisionCreator-R1	Agent	0.330	0.246	0.122	0.858	0.832	0.750
VisionCreator-P1 (ours)	Agent	**0.391**	**0.326**	**0.313**	**0.875**	**0.855**	**0.776**

Table 2: Video generation baselines under the aligned keyframe protocol on PhysBench-M, evaluated with the held-out Gemini-2.5-Pro judge.

Method	Type	Continuum	Translation	Collision
CogVideoX-5B	Video	0.369	0.379	0.163
Wan2.2-TI2V-5B	Video	0.484	0.641	0.314
VisionCreator-P1 (ours)	Agent	**0.694**	**0.691**	**0.375**

- **PhysBench-M** is a curated multi-image generation benchmark containing 900 samples, evenly distributed across three categories: *Continuum*, which evaluates the evolution of continuous processes; *Translation*, which evaluates single-rigid-body motion and identity preservation; and *Collision*, which evaluates two-body approach–contact–separation dynamics. The benchmark uses trajectory-level evaluation with predefined instructions and checkpoints at critical timesteps. Details of PhysBench-M are provided in Appendix B. Primary scores are obtained with Qwen3-VL-235B (Bai et al., 2025), and all outputs are additionally re-evaluated by the held-out Gemini-2.5-Pro evaluator; the corresponding held-out video-baseline results are reported in Tab. 2.
- **VCR-Bench** (Lai et al., 2026b) measures general visual generation and reasoning capabilities across single-image, multi-image, and image-to-image settings, with multiple evaluation checkpoints targeting semantic coherence, identity preservation, and instruction following.

Baselines. We compare our approach with representative agentic, tool-based, and open-loop video generation baselines:

- **Qwen-Image** (Wu et al., 2025) is the core underlying image generation and editing tool employed in our framework, evaluated directly to serve as a foundational, non-agentic baseline for comparison.
- **Qwen3-VL-32B-Reflection** (Bai et al., 2025) is a Qwen3-VL-32B-based agent equipped with self-reflection mechanism, serving as a highly competitive reasoning baseline for closed-loop tasks.
- **Gemini-2.5-Pro** (Comanici et al., 2025) operates as a powerful proprietary standalone framework, featuring exceptionally strong native multimodal reasoning and planning capabilities.
- **VisionCreator-R1** is a previous state-of-the-art visual generation agent built on Qwen3-VL-32B.
- **CogVideoX-5B** (Yang et al., 2025b) and **Wan2.2-TI2V-5B** (Wan et al., 2025) are representative open-loop video generation baselines. We evaluate them by extracting keyframes at the same normalized timestamps used by PhysBench-M; the common inference and keyframe protocol is detailed in Appendix B.

Implementation Details. To ensure a fair comparison, our proposed VisionCreator-P1 is also instantiated from Qwen3-VL-32B (Bai et al., 2025). All agent models share the same image-generation tool (Qwen-Image (Wu et al., 2025)) and system prompt. The video baselines receive the same PhysBench-M prompts and are evaluated with the fixed keyframe extraction protocol described in the combined results table. Full hyperparameters, reward coefficients, training costs and block/step-level configurations are detailed in the Appendix.

4.2 Main Results

We evaluate *VisionCreator-P1* on PhysBench-M and VCR-Bench for physical dynamics and general visual generation, respectively. The corresponding quantitative results are detailed in Tables 1 and 2.

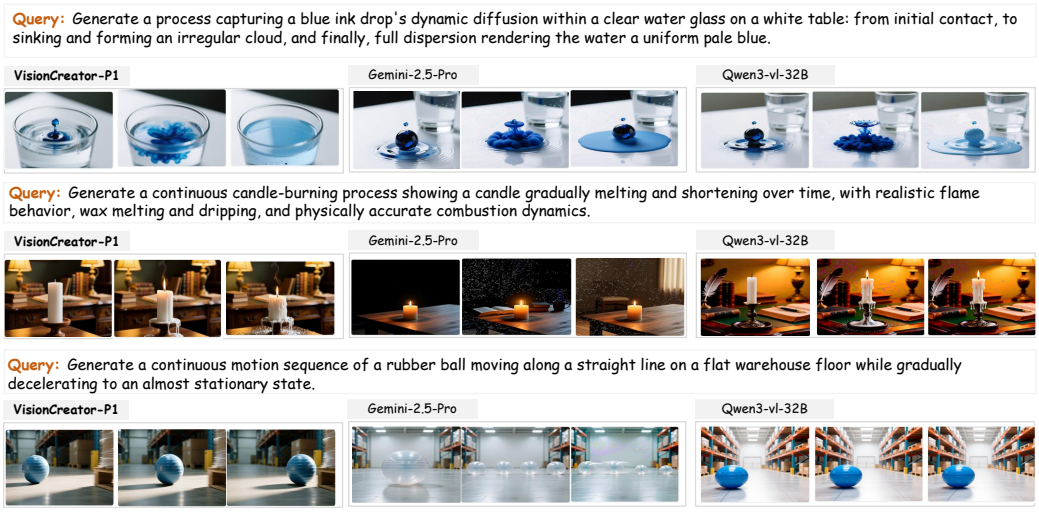


Figure 5: Qualitative examples generated by VisionCreator-P1 on PhysBench-M.

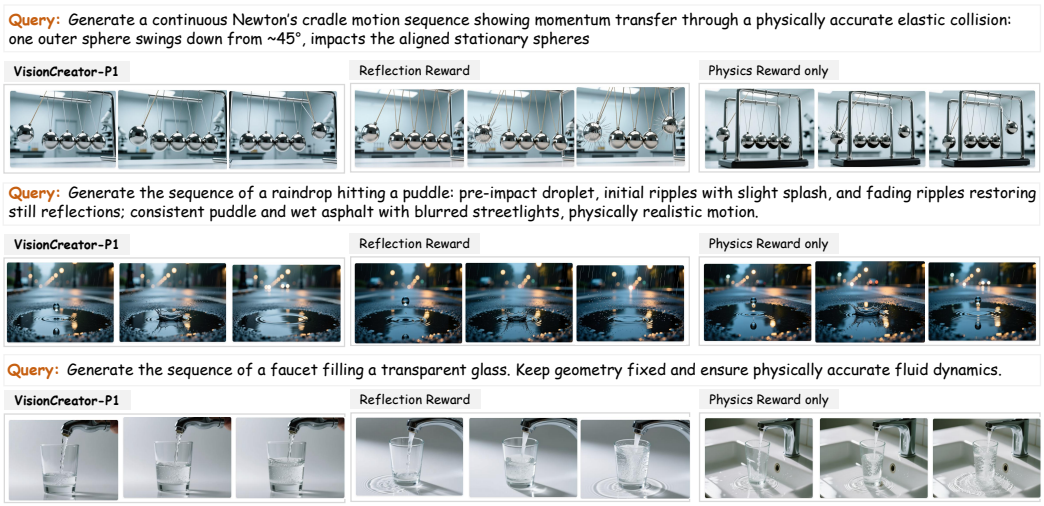


Figure 6: Qualitative ablation study comparing our VisionCreator-P1 against variants trained with reflection reward and physics reward only. VisionCreator-P1 shows stronger checkpoint-level visual and temporal consistency under the evaluated criteria. All prompts are sourced from PhysBench-M.

Video and evaluator-robustness analysis. VisionCreator-P1 obtains higher scores than both evaluated frozen video baselines across all categories under the aligned keyframe protocol (Tab. 2), with the largest advantage on Collision (improving over CogVideoX-5B and Wan2.2-TI2V-5B by 0.212 and 0.061). Re-evaluating 900 PhysBench-M outputs with Gemini-2.5-Pro shows that despite absolute calibration differences from Qwen3-VL-235B (Tables 1 and 2), the systems re-evaluated by Gemini-2.5-Pro preserve the reported category-level ordering and identify Collision as the hardest setting. We thus treat these VLM scores as benchmark-specific consistency evidence rather than physical validity guarantees.

Generalization to General Visual Generation. Introducing physical constraints yields consistent positive transfer without degrading general visual generation. On VCR-Bench (Tab. 1), *VisionCreator-P1* improves across all task types, maximizing gains on *Multi-image* generation (0.855, outperforming Qwen3-VL-32B-Reflection and *VisionCreator-R1* by 0.140 and 0.023). This suggests physics-guided training enhances both dynamic simulation and cross-image spatial alignment in general visual tasks.

4.3 Ablation Studies

We ablate several design choices across PhysBench-M and VCR-Bench (Tab. 3). We compare three training variants: (i) **PFSR** performs *step-level* rollouts with dense step rewards; (ii) **PFBR** performs *block-level* rollouts and uses block-level reward for winner selection; and (iii) **JASC** improves PFSR by anchoring the step reward and advantage to a judge VLM, preventing premature termination and suppressing mode-collapse. The findings are summarized as follows:

Table 3: Ablation study on PhysBench-M and VCR-Bench. We use our full model (**physics reward + PFBR**) as the baseline and ablate its core components: the training strategy (PFBR) and the reward type (physics vs. reflection). **Bolded** numbers indicate the best performance in each column.

Method	PhysBench-M			VCR-Bench		
	Continuum	Translation	Collision	Single-Img	Multi-Img	Img2Img
Ablation on Training Strategy						
replace PFBR with PFSR	0.324	0.300	0.223	0.873	0.841	0.759
replace PFBR with PFSR + JASC	0.359	0.288	0.241	**0.881**	0.821	0.758
Ablation on Reward Type						
w/o PFBR (physics reward only)	0.360	0.294	0.253	0.863	0.852	0.771
w/o PFBR (reflection reward only)	0.317	0.245	0.125	0.843	0.838	0.757
Ours (physics reward + PFBR)	**0.391**	**0.326**	**0.313**	0.875	**0.855**	**0.776**

Quantitative Comparisons. (1) *Necessity of Block-level Competition (PFBR).* We observe that removing PFBR and relying solely on the physics reward leads to a substantial decline in performance across both datasets, particularly in complex physical interactions (e.g., Collision drops from 0.313 to 0.253 on PhysBench-M). Furthermore, substituting PFBR with step-level competition strategies (PFSR or JASC) fails to recover the performance in multi-step generation tasks. While JASC provides denser advantage signals and slightly improves the Single-image score on VCR-Bench (up to 0.881), both step-level variants severely degrade multi-step physics reasoning and Collision performance on PhysBench-M (dropping to 0.223 for PFSR). This supports the usefulness of block-level granularity under our training configuration and aligns much better with the semantic structure of complex, multi-step physical rollouts than step-level credit assignment. (2) *Impact of Reward Types.* To evaluate the contribution of the reward signal, we replace our physics-aware reward with a standard reflection-based reward. This ablation results in consistent and dramatic performance degradation across all sub-tasks. Notably, the Continuum score on PhysBench-M falls from 0.360 to 0.317, and the Single-image score on VCR-Bench drops from 0.863 to 0.843. These results strongly validate that physics-aware rewards provide more task-specific learning signals in this setting compared to generic reflection rewards.

Qualitative Comparisons. As shown in Fig. 6, visual results align with our quantitative findings. VisionCreator-P1 faithfully preserves physical laws with strong temporal and geometric consistency, correctly keeping the middle spheres static in Newton's cradle and preserving the rigid geometry of the glass during fluid overflow. In contrast, reflection reward alone produces physically implausible artifacts such as bent strings or chaotic ripples, while physics reward only improves basic dynamics without achieving fine-grained structural and semantic coherence, underscoring the necessity of our integrated framework.

5 Conclusion

We study long-horizon physical visual generation as causal state evolution over sparse keyframes. VisionCreator-P1 addresses Global Causal Drift through closed-loop transition verification and PFBR-based block-level optimization. We introduce PhysData-M and PhysBench-M, and observe improved physical-trajectory and general visual-generation scores, including under the held-out Gemini-2.5-Pro evaluator. These results support local transition verification as a useful design principle, while our claims remain limited to sparse keyframe consistency and heuristic VLM evaluation; they do not establish dense-video or simulator-level physical validity. Further limitations and future directions are discussed in Appendix C.

AI USE STATEMENT

In this work, generative AI tools were utilized exclusively for assisting with initial language proofreading and LaTeX template formatting. All core conceptual ideas, theoretical derivations, algorithm designs, code implementations, and experimental formulations were independently conceived, verified, and authored by the researchers. The authors take full responsibility for the entirety of the content, claims, and formulations presented in this paper.

ETHICS STATEMENT

This research focuses on foundational algorithms. All datasets used are standard academic benchmarks. The proposed method does not generate harmful content, involve sensitive personal data, or introduce societal bias.

REPRODUCIBILITY STATEMENT

To ensure reproducibility, comprehensive implementation details are provided in Sec. 4.1.

REFERENCES

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

Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, Yusuf Aytar, Sarah Maria Elisabeth Bechtle, Feryal Behbahani, Stephanie C.Y. Chan, Nicolas Heess, Lucy Gonzalez, Simon Osindero, Sherjil Ozair, Scott Reed, Jingwei Zhang, Konrad Zolna, Jeff Clune, Nando De Freitas, Satinder Singh, and Tim Rocktäschel. Genie: Generative interactive environments. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 4603–4623. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/bruce24a.html.

Harold Haodong Chen, Haojian Huang, Qifeng Chen, Harry Yang, and Ser Nam Lim. Hierarchical fine-grained preference optimization for physically plausible video generation. In D. Belgrave, C. Zhang, H. Lin, L. Montoya, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, pp. 133919–133951. Curran Associates, Inc., 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/c25fe7eb6d6cc6931b93236f5599aa97-Paper-Conference.pdf.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Lin Geng Foo, Mark He Huang, Alexandros Lattas, Stylianos Moschoglou, Thabo Beeler, and Christian Theobalt. Physical simulator in-the-loop video generation. *arXiv preprint arXiv:2603.06408*, 2026.

Shijin Gong, Kai Ye, Jin Zhu, Xinyu Zhang, Hongyi Zhou, and Chengchun Shi. Kernelized advantage estimation: From nonparametric statistics to llm reasoning. *arXiv preprint arXiv:2604.28005*, 2026.

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.

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.

Ziqi Huang, Ning Yu, Gordon Chen, Haonan Qiu, Paul Debevec, and Ziwei Liu. Vchain: Chain-of-visual-thought for reasoning in video generation. *arXiv preprint arXiv:2510.05094*, 2025.

Jinxiang Lai, Zexin Lu, Jiajun He, Rongwei Quan, Wenzhe Zhao, Qinyu Yang, Qi Chen, Qin Lin, Chuyue Li, Tao Gao, et al. Visioncreator: A native visual-generation agentic model with understanding, thinking, planning and creation. *arXiv preprint arXiv:2603.02681*, 2026a.

Jinxiang Lai, Wenzhe Zhao, Zexin Lu, Hualei Zhang, Qinyu Yang, Rongwei Quan, Zhimin Li, Shuai Shao, Song Guo, and Qinglin Lu. Visioncreator-r1: A reflection-enhanced native visual-generation agentic model. *arXiv preprint arXiv:2603.08812*, 2026b.

Chenyu Li, Oscar Michel, Xichen Pan, Sainan Liu, Mike Roberts, and Saining Xie. PISA experiments: Exploring physics post-training for video diffusion models by watching stuff drop. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 35685–35709. PMLR, 13–19 Jul 2025a. URL https://proceedings.mlr.press/v267/li25bu.html.

Mingcheng Li, Xiaolu Hou, Ziyang Liu, Dingkang Yang, Ziyun Qian, Jiawei Chen, Jinjie Wei, Yue Jiang, Qingyao Xu, and Lihua Zhang. Mccd: Multi-agent collaboration-based compositional diffusion for complex text-to-image generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 13263–13272, 2025b.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The twelfth international conference on learning representations*, 2023.

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, 2024.

Wei Liu, Ziyu Chen, Zizhang Li, Yue Wang, Hong-Xing Yu, and Jiajun Wu. Realwonder: Real-time physical action-conditioned video generation. *arXiv preprint arXiv:2603.05449*, 2026.

Xiaoqian Liu, Ke Wang, Yuchuan Wu, Fei Huang, Yongbin Li, Junge Zhang, and Jianbin Jiao. Online process reward leanring for agentic reinforcement learning. *arXiv e-prints*, pp. arXiv–2509, 2025.

Yaxi Lu, Shenzhi Yang, Cheng Qian, Guirong Chen, Qinyu Luo, Yesai Wu, Huadong Wang, Xin Cong, Zhong Zhang, Yankai Lin, Weiwen Liu, Yasheng Wang, Zhiyuan Liu, Fangming Liu, and Maosong Sun. Proactive agent: Shifting llm agents from reactive responses to active assistance. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 47431–47457, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/75c37811e830bf029584b1c6fac17726-Paper-Conference.pdf.

Jiaxi Lv, Yi Huang, Mingfu Yan, Jiancheng Huang, Jianzhuang Liu, Yifan Liu, Yafei Wen, Xiaoxin Chen, and Shifeng Chen. Gpt4motion: Scripting physical motions in text-to-video generation via blender-oriented gpt planning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1430–1440, 2024.

Chengjun Pan, Shichun Liu, Jiahang Lin, Dingwei Zhu, Jiazheng Zhang, Shihan Dou, Songyang Gao, Zhenhua Han, Binghai Wang, Rui Zheng, et al. Evpo: Explained variance policy optimization for adaptive critic utilization in llm post-training. *arXiv preprint arXiv:2604.19485*, 2026.

Jie Qin, Jie Wu, Weifeng Chen, Yuxi Ren, Huixia Li, Hefeng Wu, Xuefeng Xiao, Rui Wang, and Shilei Wen. Diffusiongpt: Llm-driven text-to-image generation system. *arXiv preprint arXiv:2401.10061*, 2024.

David Romero, Ariana Bermudez, Hao Li, Fabio Pizzati, and Ivan Laptev. Learning to generate object interactions with physics-guided video diffusion. *arXiv e-prints*, pp. arXiv–2510, 2025.

Chris Samarinas, Haw-Shiuan Chang, and Hamed Zamani. Truncated step-level sampling with process rewards for retrieval-augmented reasoning. *arXiv preprint arXiv:2602.23440*, 2026.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Khanh-Tung Tran, Dung Dao, Minh-Duong Nguyen, Quoc-Viet Pham, Barry O'Sullivan, and Hoang D Nguyen. Multi-agent collaboration mechanisms: A survey of llms. *arXiv preprint arXiv:2501.06322*, 2025.

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.

Jifang Wang, Xue Yang, Longyue Wang, Zhenran Xu, Yiyu Wang, Yaowei Wang, Weihua Luo, Kaifu Zhang, Baotian Hu, and Min Zhang. A unified agentic framework for evaluating conditional image generation. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12626–12646, Vienna, Austria, July 2025a. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.620. URL https://aclanthology.org/2025.acl-long.620/.

Jing Wang, Ao Ma, Ke Cao, Jun Zheng, Jiasong Feng, Zhanjie Zhang, Wanyuan Pang, and Xiaodan Liang. WISA: World simulator assistant for physics-aware text-to-video generation. In *Advances in Neural Information Processing Systems*, 2025b. URL https://openreview.net/forum?id=4jWuS5hye1. NeurIPS 2025 Spotlight.

Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, Jiagang Zhu, and Jiwen Lu. Drivedreamer: Towards real-world-drive world models for autonomous driving. In *European conference on computer vision*, pp. 55–72. Springer, 2024a.

Yufei Wang, Zhou Xian, Feng Chen, Tsun-Hsuan Wang, Yian Wang, Katerina Fragkiadaki, Zackory Erickson, David Held, and Chuang Gan. RoboGen: Towards unleashing infinite data for automated robot learning via generative simulation. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 51936–51983. PMLR, 21–27 Jul 2024b. URL https://proceedings.mlr.press/v235/wang24cc.html.

Zhenyu Wang, Aoxue Li, Zhenguo Li, and Xihui Liu. Genartist: Multimodal llm as an agent for unified image generation and editing. *Advances in Neural Information Processing Systems*, 37: 128374–128395, 2024c.

Zixuan Wang, Yixin Hu, Haolan Wang, Feng Chen, Yan Liu, Wen Li, and Yinjie Lei. Chain of event-centric causal thought for physically plausible video generation. *arXiv preprint arXiv:2603.09094*, 2026.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

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 (CVPR)*, 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 (ICCV)*, pp. 12360–12370, October 2025a.

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, 2025b.

Tianwei Yin, Qiang Zhang, Richard Zhang, William T Freeman, Fredo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast causal video generators. *arXiv e-prints*, pp. arXiv–2412, 2024.

Chenchen Zhang. From reasoning to agentic: Credit assignment in reinforcement learning for large language models. *arXiv preprint arXiv:2604.09459*, 2026.

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*, 2025a.

Xiangdong Zhang, Jiaqi Liao, Shaofeng Zhang, Fanqing Meng, Xiangpeng Wan, Junchi Yan, and Yu Cheng. Videorepa: Learning physics for video generation through relational alignment with foundation models. In D. Belgrave, C. Zhang, H. Lin, L. Montoya, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, pp. 122647–122676. Curran Associates, Inc., 2025b. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/b1d4973f5d21708abef3cd6f17d842c8-Paper-Conference.pdf.

Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Lulu Gu, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, et al. Vbench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. *arXiv preprint arXiv:2503.21755*, 2025.

Yifei Zhou, Andrea Zanette, Jiayi Pan, Sergey Levine, and Aviral Kumar. ArCHer: Training language model agents via hierarchical multi-turn RL. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 62178–62209. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/zhou24t.html.

A Conditional Stability Analysis of Global Causal Drift

This appendix gives a conditional error-propagation analysis for Eq. 3; it is not a proof that PRTC projects generated states onto the physical manifold. Let $S_t^\star = \mathcal{T}^*(S_{t-1}^\star)$ be an ideal trajectory with $S_t^\star \in \mathcal{M}$, and define $e_t = \mathbb{E}[\mathcal{D}(S_t, S_t^\star)]$ and $e_0 = \mathbb{E}[\mathcal{D}(S_0, S_0^\star)]$. Let $\epsilon_t = \mathbb{E}[\mathcal{D}(S_t, \mathcal{T}^*(S_{t-1}))]$. Within one locally valid physical regime, the piecewise Lipschitz assumption and the triangle inequality give

$$e_t \le \epsilon_t + L e_{t-1}. \tag{9}$$

Unrolling this recurrence yields

$$e_T \le L^T e_0 + \sum_{t=1}^{T} L^{T-t} \epsilon_t. \tag{10}$$

The initial-state term is necessary unless $S_0 = S_0^\star$. If $S_t^\star \in \mathcal{M}$, the distance to this reference trajectory upper-bounds the distance to the manifold. When $L < 1$, the propagated contribution is contractive; when $L = 1$, it is additive; and when $L > 1$, it is expansive. For a uniform error $\epsilon_t = \bar{\epsilon}$ and $L \neq 1$, the sum is

$$\sum_{t=1}^{T} L^{T-t} \bar{\epsilon} = \bar{\epsilon} \frac{L^T - 1}{L - 1}, \tag{11}$$

while for $L = 1$ it equals $T\bar{\epsilon}$.

PRTC can reject or revise proposals that violate its specified heuristic checks, which may reduce the effective local errors ϵ_t. However, accepting a milestone does not imply projection onto $\mathcal{M}$ and does not reset e_t to zero. If verified milestones divide a trajectory into shorter segments, the same recurrence can be analyzed within each segment, but replacing T by T/K requires an additional reset or bounded-boundary-error assumption that is not established by our verifier. We therefore do not claim a formal $\mathcal{O}(L^T)$ to $\mathcal{O}(L^{T/K})$ guarantee; the analysis only explains why reducing local errors and effective segment length may reduce modeled error amplification under the stated assumptions.

B Extended Implementation Details

This appendix provides the complete experimental settings referenced in the main text.

VLM Roles and Reward Computation. During training, Qwen3-VL-32B computes the semantic components of the physics-aware reward, including motion-trend consistency, foreground identity preservation, and background stability. The same model grounds prompt-specified rigid objects for the geometric reward. Qwen3-VL-235B is used only as the primary evaluator for the main benchmark scores, while held-out Gemini-2.5-Pro is used for independent re-evaluation. Neither evaluation model is used as the training reward model. All VLM scores are heuristic measurements of specified visual criteria, not formal physical verification.

Rigid-object grounding and contact masking. For each frame, Qwen3-VL-32B receives the prompt-specified object names and returns one axis-aligned box per visible rigid object as normalized $(x_{\min}, y_{\min}, x_{\max}, y_{\max}) \in [0,1]^4$ coordinates. The grounding prompt is: *"Locate each named rigid object. Return one normalized axis-aligned box per visible object, preserve object names, and return unresolved when the object is mostly occluded, severely deformed, or not identifiable. Do not infer a box from the background."* Identities are linked across adjacent keyframes by the returned name together with appearance and attribute consistency; unmatched or unresolved detections are masked.

The set $\mathcal{R}$ contains pairs of identifiable rigid objects whose interaction is relevant to the prompt. For each transition, let $\mathcal{C}_t \subseteq \mathcal{R}$ be the set of pairs in prompt-specified or verifier-identified contact windows, including the contact frame and immediately adjacent approach/rebound frames when appropriate. Geometric penalties in Eq. 6 are disabled for $(i,j) \in \mathcal{C}_t$, so intended contact and rebound are not treated as interpenetration. These terms are 2D image-plane heuristic proxies for unexplained overlap or crossing, not 3D penetration measurements. For non-contact pairs, box overlap or center-line crossing is penalized only when it is not explained by the expected contact state. Partial occlusion, severe deformation, perspective-induced 2D overlap, and unresolved identity are masked rather than assigned guessed boxes. Approach, contact, and separation are distinguished by

Table 4: Video-baseline inference settings. Values follow the official reference configurations; the seed is fixed for reproducibility.

Parameter	CogVideoX-5B	Wan2.2-TI2V-5B
Resolution	720×480	1280×704
Frames	49	121
Frame rate	8 fps	24 fps
Sampling steps	50	50
Guidance scale	6.0	5.0
Sampling shift	–	5.0
Precision	BF16	official model dtype
Seed	42	42
Prompt extension	disabled	disabled

the prompt checkpoint and temporal order; tunneling is penalized only for an unexplained crossing outside the contact window.

PFBR implementation details. Each candidate block is a complete semantic response unit, such as a think–plan–tool-call sequence; it may include a state description, rendering call, and verification output. At each round, K candidates are sampled from the old policy conditioned on the same verified prefix. If candidate k receives block reward r_k, its within-group advantage is $A_k = (r_k - \bar{r})/(\mathrm{std}(r) + \varepsilon)$. The largest-advantage candidate is appended to the prefix; discarded candidates provide the within-group baseline but are not appended. The policy ratio is computed over tokens in the sampled block as $\rho_k = \pi_\theta(a_k \mid x_k)/\pi_{\mathrm{old}}(a_k \mid x_k)$, and the clipped objective is applied to those block tokens. Rendering and verification calls are environment actions used to obtain the block reward. PFBR is therefore a training-time sampling and optimization procedure, distinct from PRTC at inference time.

Physics Reward Design. The dynamic interaction reward is defined as

$$\mathcal{R}_{\text{dyn}} = 1 - \big(\lambda_1 \cdot \text{IoU_penalty} + \lambda_2 \cdot \text{traj_penalty}\big) + \lambda_3 \cdot \text{motion_trend_score},$$

with $\lambda_1 = 0.3$, $\lambda_2 = 0.5$, and $\lambda_3 = 0.4$. The static consistency reward is

$$\mathcal{R}_{\text{stat}} = \mu_1 \cdot \text{foreground_coherence} + \mu_2 \cdot \text{background_invariance},$$

with $\mu_1 = 0.5$ and $\mu_2 = 0.5$. The two branches are fused with $\alpha = \beta = 0.5$.

PFBR Configuration. Block-level reinforcement uses $k = 4$ candidate blocks, argmax winner selection, a maximum of 16 steps per trajectory, Z-score denominator $\varepsilon = 0.1$, advantage clipping $[-3.0, +3.0]$, and cumulative reward clipping $[-5.0, +5.0]$. PFSR follows the same configuration at the step level. The KL penalty coefficient is 0.01, and the rollout group size is 4.

Training Infrastructure. The RL training stage uses 24 NVIDIA H20 GPUs with 96 GB memory each and typically takes approximately 3–5 days. Image-generation rollouts dominate the training cost.

Video baseline protocol. CogVideoX-5B and Wan2.2-TI2V-5B are evaluated as frozen open-loop video generators; no task-specific fine-tuning, physics-aware post-processing, or trajectory-level correction is applied. Both models receive the same PhysBench-M queries as the agent baselines. Tab. 4 gives the quantitative configurations used for reproducible inference.

For each PhysBench-M prompt, the prompt specifies K keyframes and, when applicable, normalized timestamps $u_1, \ldots, u_K \in [0, 1]$. We extract exactly K frames from each generated video. If timestamps are specified, u_k is mapped to the nearest generated frame using $f_k = \mathrm{round}(u_k(F-1))$, where F is the generated frame count. If only K is specified, we use K uniformly spaced timestamps including the first and last frame. The extracted keyframes are evaluated with the same category-specific protocol and held-out Gemini-2.5-Pro rubric used for Tab. 2. This measures consistency at prompt-aligned keyframes, not validity at unobserved intermediate frames.

Benchmark construction. PhysBench-M uses a three-level taxonomy: domain (Continuum, Translation, or Collision), scenario type, and difficulty factor. Representative scenario types include melting, pouring, spreading, free fall, projectile motion, sliding, rolling, head-on, oblique, grazing, rear-end,

Table 5: Blinded human evaluation margins over five checkpoint scores. Positive values favor VisionCreator-P1 over VisionCreator-R1.

Comparison	Continuum	Collision	Translation
P1 vs. R1	$+21.5\%$	$+46.1\%$	$+18.7\%$

Table 6: Inference cost measured by mean tool calls per trajectory. Qwen3-VL-32B is reported over the categories available in the rebuttal evaluation.

Method	Evaluated categories	Mean calls	Success
VisionCreator-P1	All evaluated categories	4.97	100%
Gemini-2.5-Pro	All evaluated categories	6.27	99.6%
Qwen3-VL-32B	Available categories	9.79	100%

elastic, and partially inelastic collisions. Difficulty factors include entity count, occlusion, viewpoint change, material transformation, and multi-stage causality. Each record contains a query, ordered keyframe checkpoints, expected state changes, invariants, and failure criteria.

Candidate prompt–checkpoint pairs are generated under fixed taxonomy-conditioned templates and filtered automatically. We remove records with missing fields, contradictory constraints, under-specified or ambiguous descriptions, duplicate prompts or rendered states, unreadable images, missing checkpoints, and overlap with PhysData-M. Expert annotators independently verify physical correctness, visual realizability, determinacy, identity preservation, temporal order, causal transition, material behavior, spatial support, and prompt–checkpoint consistency; disagreements are adjudicated against the structured schema. The final benchmark contains 900 samples, evenly split across the three domains.

Continuum heuristic checkpoints. Continuum samples use five binary heuristic checkpoints: (i) prompt compliance and stage completeness, (ii) physical-law-consistency cues, (iii) object/attribute consistency, (iv) spatial/environmental continuity, and (v) temporal/causal ordering. These checkpoints localize visual violations but are not substitutes for numerical simulation or exact conservation-law verification.

Human evaluation. We conduct a blinded pairwise comparison between VisionCreator-P1 and VisionCreator-R1. Annotators do not see model names, and presentation order is randomized. For each sample, the five predefined checkpoints are independently marked as satisfied or unsatisfied for both trajectories. The per-sample score is $h = (1/5)\sum_{j=1}^{5} \mathbb{I}[c_j = 1] \in [0, 1]$. The reported human preference margin is $100 \times (\bar{h}_{\mathrm{P1}} - \bar{h}_{\mathrm{R1}})$, in absolute percentage points rather than relative improvement; positive values favor VisionCreator-P1. The margins are $+21.5\%$ for Continuum, $+18.7\%$ for Translation, and $+46.1\%$ for Collision.

We do not directly estimate gradient variance or provide a formal variance-reduction guarantee; PFBR is therefore interpreted as an empirically motivated optimization heuristic.

Inference cost. Each rendered candidate triggers one transition-verification pass, including the final successful verification. We report total image-tool calls per trajectory; all compared agents use the same system prompt and image-generation tool.

VisionCreator-P1 uses 20.7% fewer tool calls per trajectory than Gemini-2.5-Pro. On the categories available for Qwen3-VL-32B, it uses 4.95 versus 9.79 calls, a 49.4% reduction. These are end-to-end tool-call comparisons rather than wall-clock latency measurements.

C LIMITATIONS

We discuss the key limitations of VisionCreator-P1 to provide a transparent assessment of its scope and boundaries.

Single base model. VisionCreator-P1 is instantiated exclusively on Qwen3-VL-32B. While this provides a strong and consistent baseline for fair comparison, the approach's performance may

vary with different backbone architectures, model scales, or multimodal capabilities. Ablation on alternative base models (e.g., larger or differently trained VLMs) is left for future work.

Potential misuse risks. Like any generative model capable of producing realistic imagery, VisionCreator-P1 carries inherent risks of misuse for creating deceptive or fabricated visual content. Responsible deployment would require safeguards such as usage guidelines, content provenance tracking, and watermarking mechanisms. These are not yet implemented in the current framework.

D Broader Impacts

We discuss the broader societal implications of VisionCreator-P1, addressing both positive applications and potential negative consequences.

Positive societal impacts. VisionCreator-P1 enables physically grounded multi-image trajectory generation that can benefit a range of downstream applications. In *embodied AI* and *robotics simulation*, generating realistic, physically plausible scene evolution can reduce the cost and risk of real-world robot training, enabling safe virtual testing of manipulation, navigation, and interaction policies. In *autonomous driving*, the ability to generate consistent multi-image trajectories under physical laws can serve as a data augmentation or scenario simulation tool, improving the robustness of perception and planning models. In *interactive physics simulators and educational tools*, visually coherent physical animations (e.g., fluid dynamics, collisions, projectile motion) can enhance learning experiences and lower the barrier to understanding complex physical concepts. Furthermore, by improving the temporal and physical coherence of generated visual content, VisionCreator-P1 contributes to the broader goal of trustworthy generative AI, where generated imagery adheres to domain-specific constraints rather than hallucinating visually pleasing but physically implausible outputs.

Negative societal impacts. The capability to generate physically consistent visual content also carries potential misuse risks. Specifically, VisionCreator-P1 could be exploited to create convincing but fabricated physical scenarios (e.g., fake accident simulations, misleading scientific visualizations, or synthetic media used for disinformation). Unlike text or image manipulation alone, physically coherent multi-image sequences are harder to distinguish from real footage, potentially amplifying the harm of synthetic media. Additionally, since VisionCreator-P1 builds upon an image generation model as its tool, any existing vulnerabilities or biases present in the underlying generator may propagate into the generated physical trajectories. We emphasize that responsible use of this technology requires adherence to content authenticity standards, and future deployment should consider watermarking, provenance tracking, and usage guidelines. Importantly, VisionCreator-P1 does not unlock fundamentally new generative capabilities beyond what the underlying image generator already provides; rather, it adds physical reasoning and temporal consistency, so the risk profile is largely inherited from the base image generation system.

E Responsible Release Safeguards

We outline the planned safeguards for the responsible release of VisionCreator-P1 artifacts (code, training data, and model weights). These measures are designed to mitigate dual-use risks associated with physically coherent image sequence generation, as discussed in the Broader Impacts section (§ D).

Usage Guidelines. Upon release, all artifacts will be accompanied by a clear *Acknowledgment and Usage Agreement* that explicitly prohibits use cases involving synthetic media intended to deceive, manipulate, or spread disinformation. Users will be required to agree to these terms before accessing the artifacts.

Content Provenance Tracking. We will integrate and recommend industry-standard content provenance metadata (following the C2PA standard) for all generated outputs. This includes embedding information about the generation model, prompt, and timestamp into the output files to enable downstream detection of synthetic content.

Watermarking Mechanisms. The release will include guidance and tooling for applying invisible statistical watermarks to generated images, following established techniques in the literature (e.g.,

adversarial perturbation-based watermarking or steganographic embedding). This enables post-hoc attribution of generated content without visibly degrading image quality.

Access Controls. The model weights and PhysData-M dataset will be released under a restricted license with commercial-use limitations. Access will be managed through a formal request-and-review process, ensuring that applicants agree to the usage terms and demonstrating legitimate research or application intent.

Unsafe Content Exclusion. PhysData-M was constructed from structured physical scenario definitions authored by the research team, without scraping external sources. All generated samples underwent automated sanity checks (e.g., filtering for geometric anomalies, spatial coherence violations, or visually implausible configurations) before inclusion. No unsafe, personally identifiable, or copyrighted content was introduced.

Disclaimer. VisionCreator-P1 builds upon an open-source image generation model (Qwen-Image) as its generation tool. The risk profile of VisionCreator-P1 is largely inherited from this base system. Our contributions are in physical reasoning and trajectory-level consistency, not in unlocking new generative capabilities beyond what the base model provides.