AGENTIC PHYSICS-GROUNDED VIDEO GENERATION VIA MEMORY-AUGMENTED IN-CONTEXT REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Physical plausibility is essential for realistic video generation. Existing methods impose physical guidance before synthesis via simulations or refined prompts, but this can limit dynamic diversity, miss errors that emerge only after generation, and require repeated regeneration when inconsistencies remain. We therefore study a *generate–diagnose–repair* paradigm that preserves the original generation capability and selectively repairs physical inconsistencies afterward, raising two key questions: what to repaired, and how to repair? To address them, we introduce **PhysRepair**, an agentic video generation framework that combines causal reasoning over the observed physical failure with prior repair experience to enable localized and adaptive correction. Specifically, **Spatiotemporal Causal Constraints** identify the first invalid physical state and its causally affected states, producing a repair specification that guides physically plausible local regeneration while preserving unaffected content. **Experience-augmented In-context Learning** combines the ongoing repair context with relevant experience from previous repairs, allowing the agent to adapt its repair decisions across diverse physical failures and accumulate completed trajectories for future cases. Experiments on Physics-IQ Verified across solid mechanics, fluid dynamics, optics, thermodynamics, and magnetism demonstrate that PhysRepair consistently improves the physical plausibility of generated videos over existing methods.

1 INTRODUCTION

Modern video generators achieve impressive visual quality but still struggle with physical plausibility, often producing implausible motion, incorrect contact or collision behavior, and unrealistic responses to physical forces (Bansal et al., 2024; 2025; Guo et al., 2025). Existing approaches to physically plausible video generation primarily seek to prevent physical failures during synthesis by incorporating physical knowledge through training data, motion priors, simulation, or iteratively refined physical prompts (Liu et al., 2024; Xue et al., 2025; Yang et al., 2025; Nguyen et al., 2026). However, imposing extensive physical guidance before synthesis can constrain the model's ability to generate diverse and complex dynamics, favoring simpler and predefined physical interactions. In contrast, we follow a *generate–diagnose–repair* paradigm that preserves the generator's original synthesis capability and introduces physical guidance only after specific inconsistencies are observed, enabling localized correction of the affected interactions while leaving already-valid content unchanged, as shown in Fig. 1. This formulation raises two coupled challenges: 1) **What to repair.** A spatial mask can localize an object, but cannot determine which temporal states should change, whether another object's trajectory should be preserved, or how a physical failure propagates across objects, space, and time (Yi et al., 2020; Motamed et al., 2026); 2) **How to repair.** Diverse physical failures require different physical evidence, simulation configurations, and generation conditions that cannot be predefined for every case, while overly broad corrections may disrupt already-valid interactions and dynamics and introduce new inconsistencies (Shinn et al., 2023; Xu et al., 2026).

To address these challenges, we cast physical video repair as an **adaptive agentic process**: the agent reasons about the causal influence of an observed physical failure to localize the repair to the affected content, while drawing on prior repair processes of related physical failures to determine how the repair should be performed. This enables targeted and adaptive correction without globally modifying already-valid dynamics or solving each failure from scratch. By iteratively diagnosing

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.

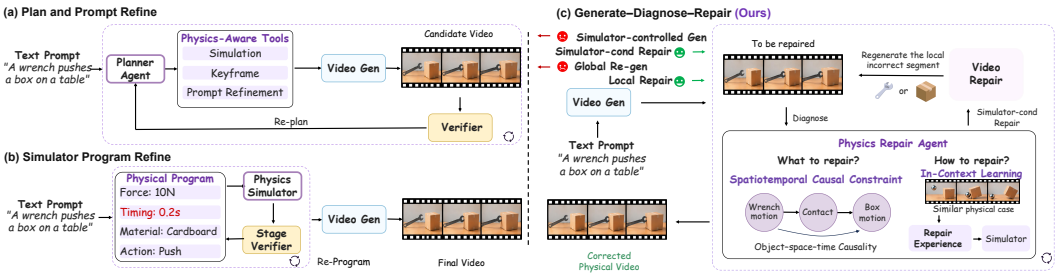


Figure 1: **Comparison of physics-grounded video generation paradigms. (a) Plan and Prompt Refine** iteratively updates generation plans or physics-aware prompts, while **(b) Simulator Program Refine** iteratively updates physical programs or simulator configurations; both impose physical guidance before synthesis and may require repeated regeneration. **(c) PhysRepair** instead follows a *generate–diagnose–repair* paradigm: it diagnoses physical failures from the generated video (what to repair?) and performs localized, causally constrained repair, while leveraging prior repair experience to adapt the intervention and preserve generated already-valid content (how to repair?).

repaired results and refining subsequent interventions, the agent progressively improves the physical plausibility of the generated video.

Concretely, we introduce **PhysRepair**, an agentic video generation system that operationalizes this adaptive repair process through **Spatiotemporal Causal Constraints** and **Experience-augmented In-context Learning**. For Spatiotemporal Causal Constraints, we organize the diagnosed interaction into an object-centric causal representation, locate the first physically invalid state, and trace its downstream effects to construct a causal repair specification. The specification is then used to reconstruct a physically plausible alternative to the incorrect dynamics and guide localized video regeneration while suppressing conflicting source dynamics. Experience-augmented In-context Learning maintains the evolving repair process in short-term working memory and augments it with relevant repair experience retrieved from long-term experience memory. Together, they provide in-context guidance to adapt agentic repair decisions, while completed trajectories are distilled back into the long-term experience memory for future repairs.

We evaluate PhysRepair on Physics-IQ Verified across five physics categories–solid mechanics, fluid dynamics, optics, thermodynamics, and magnetism–to assess physical plausibility. Compared with existing methods, PhysRepair demonstrates improved physical plausibility on Physics-IQ Verified. Ablation studies further examine the effects of Spatiotemporal Causal Constraints and Experience-augmented In-context Learning. Our main contributions are:

- We introduce **PhysRepair**, an agentic *generate–diagnose–refine* formulation that iteratively identifies and corrects physical failures after generation, enabling targeted intervention while preserving already-valid dynamics and visual content.
- We address the complementary challenges of what to repair and how to repair through two components: Spatiotemporal Causal Constraints and Experience-augmented In-context Learning. The former identifies causally affected object–space–time states to localize intervention while preserving unaffected content, whereas the latter retrieves relevant prior repair experience to adapt repair decisions across diverse physical failures.
- We conduct extensive experiments on Physics-IQ Verified across five physics categories–solid mechanics, fluid dynamics, optics, thermodynamics, and magnetism, demonstrating consistent improvements in physical correctness.

2 Related Work

2.1 Physics-aware video generation and physical evaluation

Recent benchmarks reveal that video generators often fail to satisfy physical laws, even when they achieve high photorealism and temporal smoothness. VideoPhy and PhyGenBench evaluate physical common sense across solid, fluid, and multi-domain phenomena (Bansal et al., 2024; Meng

et al., 2024). Physics-IQ tests whether models can predict physically valid futures from observed experimental frames (Motamed et al., 2025). Meanwhile, VideoPhy-2 and T2VPhysBench evaluate fine-grained rule adherence and first-principles conservation laws (Bansal et al., 2025; Guo et al., 2025). Together, these studies highlight the need to explicitly assess contacts, trajectories, and transitions, which directly provides candidate failure categories for our verifier. Beyond evaluation, another line of work injects physical guidance prior to video synthesis. For example, PhysGen renders rigid-body simulations via diffusion priors (Liu et al., 2024). PhyT2V iteratively refines text prompts using physical rules (Xue et al., 2025). Similarly, VLIPP plans coarse trajectories (Yang et al., 2025), and PhysPlan optimizes motion updates away from passive regions via agentic planning (Nguyen et al., 2026). However, imposing extensive physical constraints before synthesis restricts the diversity of generated dynamics, cannot address physical errors that only appear post-generation, and often requires costly global regeneration. In contrast, our *generate–diagnose–repair* paradigm preserves the base model's generative capacity. We introduce physical guidance only after detecting specific inconsistencies, enabling localized corrections while keeping already-valid content intact.

2.2 Agentic verification, refinement, and experience

Large language models (LLMs) and vision-language models (VLMs) are increasingly used for physical planning and output evaluation. Beyond pre-generation planning (Xue et al., 2025; Nguyen et al., 2026), VeriPhy translates prompts into explicit physical rules, collects tool-based evidence, and verifies outcomes before rewriting prompts (Xu et al., 2026). These systems demonstrate the value of verification feedback, but primarily use it to revise prompts for global regeneration rather than guide localized repair. Moving from global regeneration to localized repair requires not only identifying physical failures but also selecting intervention strategies suited to each case. To support such adaptive decision-making, frameworks such as Reflexion and Voyager offer a useful precedent: they reuse feedback and accumulated experience to guide subsequent actions without updating model weights (Shinn et al., 2023; Wang et al., 2023), providing a basis for experience-driven in-context learning (ICL). Building on this principle, our Experience-Augmented In-Context Learning combines current repair feedback with relevant prior repair experience in the agent's context, enabling adaptive repair decisions without solving each failure from scratch.

3 Method

3.1 PhysRepair: Generate–Diagnose–Refine Framework

PhysRepair is an agentic video generation framework that turns video synthesis from a one-pass procedure into an iterative *generate–diagnose–refine* process. Given a generation condition c, a base video generator first produces an initial video

$$V^{(0)} = G_0(c). \tag{1}$$

Rather than treating $V^{(0)}$ as the final output, PhysRepair regards it as an intermediate generation whose realized dynamics can be explicitly inspected. At each iteration k, a Physics Repair Agent diagnoses the current video $V^{(k)}$, constructs a targeted physical correction, executes a localized counterfactual repair, and verifies the resulting video. This feedback-driven process allows physical guidance to be determined from the failure that actually emerges in the generated video, rather than being fully specified before synthesis.

Evidence-grounded physical diagnosis. Given the current generation $V^{(k)}$, the agent first performs a coarse global scan using sparsely sampled frames to identify suspicious physical events and approximately localize where and when a potential violation occurs. It then increases the temporal resolution around the localized event window and selectively invokes perception tools for tracking, segmentation, motion, depth, and contact cues. The resulting fine-grained evidence is used to determine the affected entities, failure type, and precise temporal relation, yielding a structured diagnosis $D^{(k)}$. This evidence-grounded diagnosis provides the basis for subsequent repair while avoiding unnecessary modification of already valid content.

Counterfactual physics-guided repair. Based on $D^{(k)}$, PhysRepair constructs a localized repair specification $R^{(k)}$ that determines the interaction to be corrected and the content to be preserved.

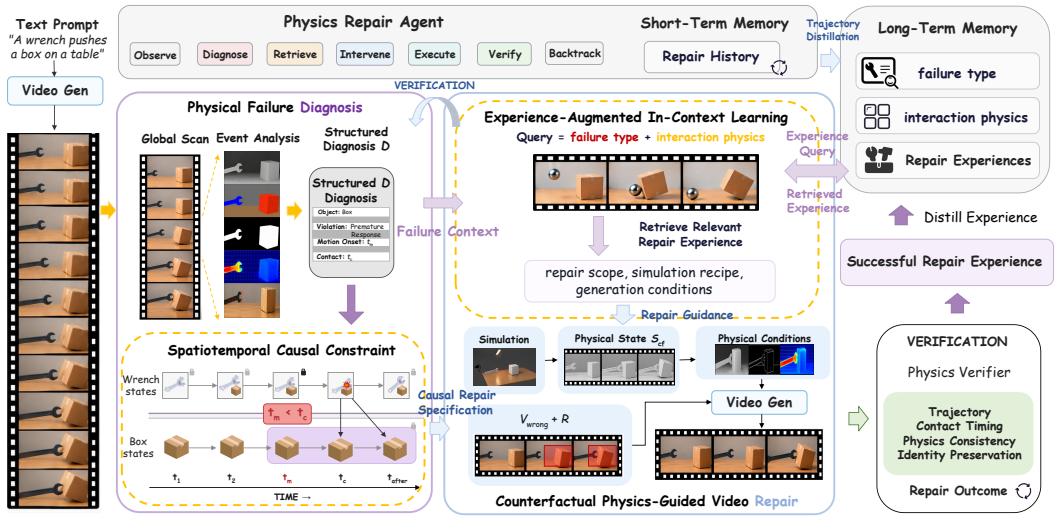


Figure 2: **Overview of PhysRepair.** Starting from an initially generated video, a Physics Repair Agent diagnoses the realized physical failure and performs localized counterfactual repair. Spatiotemporal causal constraints determine *what may change and what must remain fixed*, while experience-augmented in-context learning retrieves relevant repair experience to adapt *how to repair*. The resulting counterfactual physical state guides video regeneration, followed by physics-and-preservation verification and experience distillation for subsequent repair.

We defer the construction of $R^{(k)}$ to Sec. 3.2. Given $R^{(k)}$, the framework reconstructs only the local physical scene required to resolve the diagnosed interaction. An executable physics simulator $\mathrm{Sim}(\cdot)$ then evolves this local scene to obtain a physically admissible counterfactual state

$$S_{\mathrm{cf}}^{(k)} = \mathrm{Sim}\left(R^{(k)}\right), \tag{2}$$

which describes the corrected physical evolution rather than the final photorealistic appearance.

PhysRepair transfers this physical correction back to video generation through two parallel branches. First, source dynamics that conflict with the diagnosed correction are suppressed within the editable region, producing a decontaminated source $V_{\mathrm{clean}}^{(k)}$ while preserving valid appearance and unaffected dynamics. In parallel, $S_{\mathrm{cf}}^{(k)}$ is converted into generation-compatible physical conditions $C_{\mathrm{phys}}^{(k)}$, including trajectory, depth, motion flow, and contact or mask cues. The resulting candidate repair is

$$\widetilde{V}^{(k+1)} = G_\theta\left(V_{\mathrm{clean}}^{(k)}, C_{\mathrm{phys}}^{(k)}, R^{(k)}\right). \tag{3}$$

Verification and iterative refinement. A physics verifier evaluates whether $\widetilde{V}^{(k+1)}$ follows the counterfactual physical target while preserving non-target content. A successful repair is accepted; otherwise, the verification outcome is returned to the Physics Repair Agent for another targeted refinement.

This loop defines the execution backbone of PhysRepair. Two decisions remain to instantiate targeted and adaptive repair: *what should be modified* after a physical failure is diagnosed, and *how the repair process should adapt* across different failures and iterative attempts. We address these decisions with Spatiotemporal Causal Constraints and Experience-Augmented In-Context Learning, respectively.

3.2 Spatiotemporal Causal Constraints

The framework above relies on a repair specification $R^{(k)}$ to localize each intervention, but a structured diagnosis $D^{(k)}$ alone does not determine which states should be modified and which should remain fixed. We therefore derive the repair boundary from the causal structure of the diagnosed

interaction. Specifically, we organize the relevant object states into an object-centric spatiotemporal graph

$$G^{(k)} = (\mathcal{V}^{(k)}, \mathcal{E}^{(k)}_{\text{temp}}, \mathcal{E}^{(k)}_{\text{int}}). \tag{4}$$

Here, $\mathcal{E}_{\text{temp}}$ connects successive states of the same object, while $\mathcal{E}_{\text{int}}$ represents cross-object physical relations such as contact, collision, and support. The graph therefore captures both object-wise temporal evolution and physical interactions between objects.

Using $D^{(k)}$, we locate the first physically invalid state within the diagnosed interaction and trace its causally affected descendants through the graph. Only these affected states are marked as editable, while unrelated object states, trajectories, and scene content are preserved. The editable support is consequently localized jointly over object, space, and time rather than by a spatial mask or temporal interval alone. This yields the causal repair specification

$$R^{(k)} = \left(\mathcal{I}^{(k)}, \mathcal{E}^{(k)}, \mathcal{P}^{(k)}\right), \tag{5}$$

where $\mathcal{I}^{(k)}$ denotes the intervention target, $\mathcal{E}^{(k)}$ the editable object–space–time region, and $\mathcal{P}^{(k)}$ the states and visual content to be preserved. $R^{(k)}$ serves both as the specification of the local interaction for counterfactual simulation and as the constraint for suppressing contaminated source dynamics during video reconditioning.

3.3 Experience-Augmented In-Context Learning

The causal repair specification determines where to intervene, but not how the repair strategy should adapt across diverse physical failures. Different failures may require different evidence, tool selections, simulation configurations, generation conditions, and recovery strategies. We therefore augment the Physics Repair Agent with in-context repair experience at both the current-sample and cross-sample levels.

For the current video, a short-term working memory $\mathcal{M}_s$ maintains the evolving repair context, including observed evidence, current hypotheses, repair history, and verification feedback. A long-term experience memory $\mathcal{M}_l$ stores successful repairs, failed attempts, and reusable repair skills accumulated across previous videos. Retrieved experience is provided in context to the agent together with the current working state, rather than being directly injected into the video generator.

At each decision step t, the agent retrieves experience relevant to the current failure context q_t and selects its next action according to

$$a_t = \pi(o_t, \mathcal{M}_s, \text{Retrieve}(\mathcal{M}_l, q_t)), \tag{6}$$

where o_t denotes the current observation and a_t may correspond to analysis, intervention, execution, verification, or backtracking. Retrieved experience can therefore guide both the choice of repair operations and the recovery from unsuccessful attempts.

Verification feedback updates $\mathcal{M}_s$ for subsequent refinement, while completed repair trajectories and their outcomes are distilled into $\mathcal{M}_l$ as cross-sample experience. This retrieval–feedback– distillation process enables the agent to adapt its repair decisions using both the current repair history and experience from related physical failures.

4 Experiments

We evaluate whether PhysRepair improves physical plausibility while preserving content outside the diagnosed repair scope. We further study the contribution of its causal constraint and experience-augmented repair strategy.

4.1 Experimental Setup

Benchmark and metrics. We evaluate on **Physics-IQ Verified** (Motamed et al., 2025; Rädsch et al., 2026), comprising 198 take-1 instances from 66 real-world physical experiments observed from three viewpoints. The benchmark spans solid mechanics, fluid dynamics, optics, thermodynamics, and magnetism. We use the verified physical descriptions and switch frames for generation,

Table 1: **Quantitative results on Physics-IQ.** SP, ST, and WS denote Spatial IoU, Spatiotemporal IoU, and Weighted Spatial IoU respectively.

Method	SP↑	ST↑	WS↑	MSE score↑	Physics-IQ↑
Video generation models					
MAGI-1 4.5B (Sand AI, 2025)	14.3	13.3	7.0	–	18.8
Cosmos-Predict2-2B (NVIDIA, 2025)	–	–	–	–	27.7
Wan2.2-TI2V-5B (Wan Team, 2025)	35.9	26.4	23.0	25.5	27.7
CogVideoX-I2V-5B (Yang et al., 2024)	37.8	35.5	21.8	32.0	31.8
Wan2.2-I2V-14B (Wan Team, 2025)	51.1	20.5	28.5	28.9	32.2
HunyuanVideo-1.5-I2V (Wu et al., 2025)	47.1	26.9	29.7	30.0	33.4
MiniMax H3 FL2VA	58.9	22.8	40.5	37.0	39.8
Cosmos3-Super-Image2Video (NVIDIA, 2026)	57.4	32.5	43.0	37.8	42.7
Physics-aware methods					
VLIPP (Yang et al., 2025)	–	–	–	–	34.6
PhyCo (Narayanan et al., 2026)	–	–	–	–	36.3
PhysRepair (Ours)	58.2	34.7	44.2	37.4	43.6

with model-specific prompt formatting following the benchmark protocol when available. Reference continuations and evaluator masks are never exposed to PhysRepair or its experience memory and are accessed only during evaluation. We follow the official evaluation protocol and report the aggregate **Physics-IQ Verified** score. The evaluator additionally measures Spatial IoU, Spatiotemporal IoU, Weighted Spatial IoU, and RGB MSE, which jointly characterize the spatial and temporal agreement between generated videos and physically valid reference dynamics.

Baselines. We compare against six general-purpose video generators: **Wan2.2** (Wan Team, 2025), **CogVideoX-I2V-5B** (Yang et al., 2024), **Cosmos-Predict2-2B** (NVIDIA, 2025), **Cosmos3-Super-Image2Video** (NVIDIA, 2026), **MAGI-1 4.5B** (Sand AI, 2025), and **HunyuanVideo-1.5-I2V** (Wu et al., 2025), as well as three physics-aware approaches: **VLIPP** (Yang et al., 2025) and **PhyCo** (Narayanan et al., 2026). Every method receives the same switch frame and prompt, and failed samples are retained in the evaluation.

Implementation details. PhysRepair generates and repairs videos using the frozen MiniMax H3 Ref2VA backbone. GPT performs structured diagnosis, causal planning, verification, and experience distillation. Perception and registration combine SAM 2 with motion, tracking, and camera-registration tools, while the diagnosed local interaction is reconstructed and simulated in simulator. The resulting counterfactual state is converted into generation-compatible structural and motion conditions and injected through the H3 control adapter. The generator uses 20 denoising steps, and the agent performs at most 20 repair rounds, terminating once the physics criteria are satisfied.

4.2 Main Results

Quantitative comparison on Physics-IQ. Table 1 compares PhysRepair with general-purpose video generators and physics-aware methods. PhysRepair achieves the highest overall Physics-IQ score of 43.6 among the reported methods, surpassing the strongest video-generation baseline, Cosmos3-Super-Image2Video, by 0.9 points. Compared with MiniMax H3 FL2VA, PhysRepair improves Physics-IQ from 39.8 to 43.6, demonstrating the benefit of targeted physical repair over the initial generation. The improvement is concentrated in the spatiotemporal metrics. In particular, PhysRepair increases ST from 22.8 to 34.7 (+11.9) and WS from 40.5 to 44.2 (+3.7) over MiniMax H3 FL2VA. This pattern is consistent with the design of PhysRepair: rather than globally modifying already-valid visual content, it localizes causally affected states and corrects the corresponding interaction dynamics. The larger gains in ST and WS therefore indicate improved spatiotemporal agreement with physically valid dynamics, while the stable SP and MSE scores suggest that these corrections are achieved without substantially perturbing the original spatial appearance.

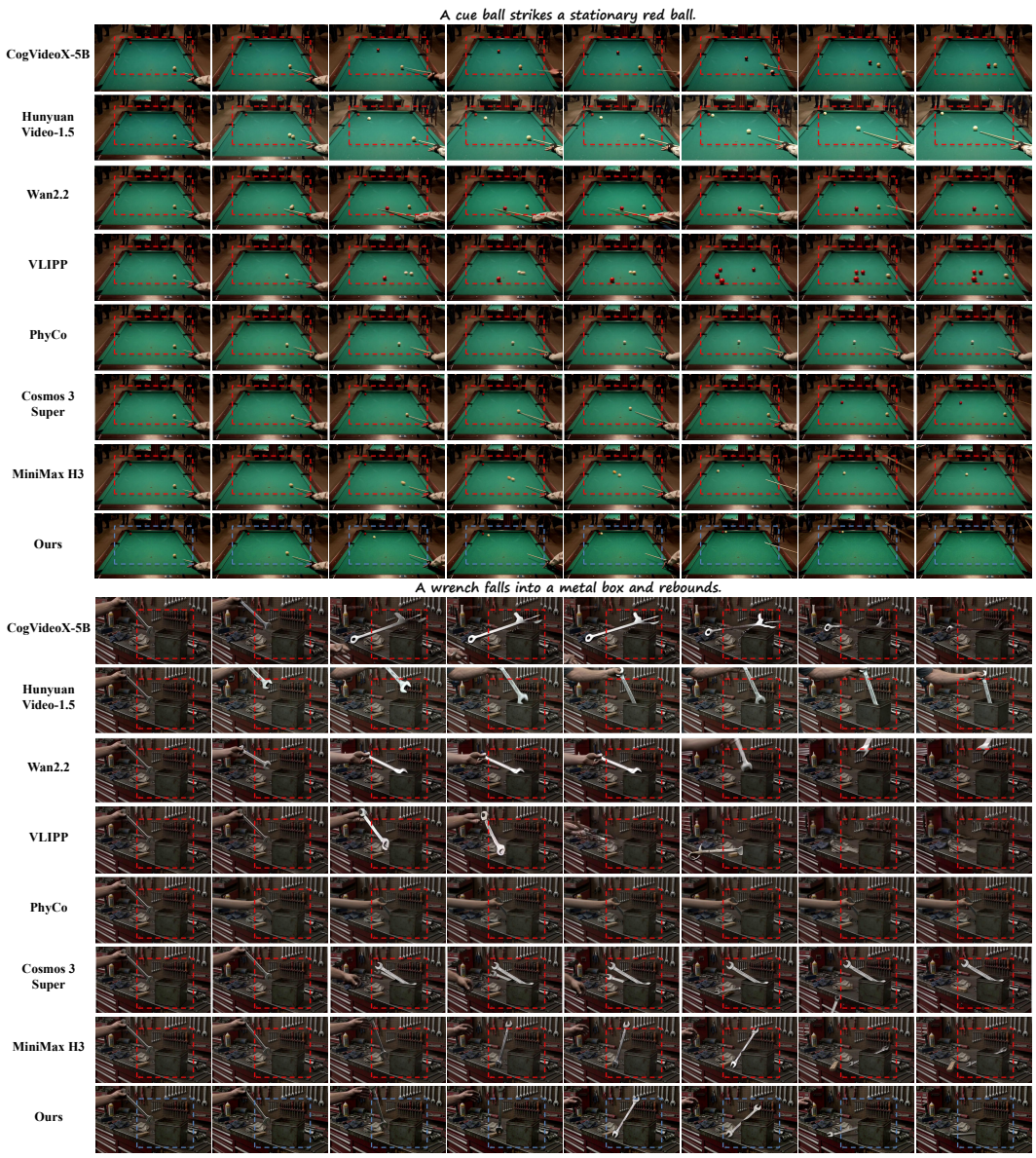


Figure 3: **Qualitative comparison on multi-object spatiotemporal interactions.** Top: cue-ball collision; bottom: wrench–box interaction. Baselines exhibit incorrect contact timing, inconsistent motion, or object artifacts, whereas PhysRepair preserves object identity and coherent contact–response dynamics.

Qualitative comparison on multi-object spatiotemporal interactions. Figure 3 further compares different methods on two contact-driven interactions that require maintaining physical consistency across multiple interacting objects and over time. In the cue-ball example, the generated sequence must preserve the identities of both balls, establish the actual collision, and trigger the red ball's response only after contact. Existing methods frequently break this interaction through premature or missing responses (VLIPP), physically implausible trajectories (Cosmos3-Super), or object duplication (MiniMax H3). In contrast, PhysRepair preserves both object identities and restores the correct contact–response sequence: the cue ball first contacts the red ball, which then moves along a physically plausible trajectory. This result illustrates how our Spatiotemporal Causal Constraints support targeted correction of causally affected states while preserving valid object states, enabling coherent multi-object dynamics without unnecessary changes to unaffected content.



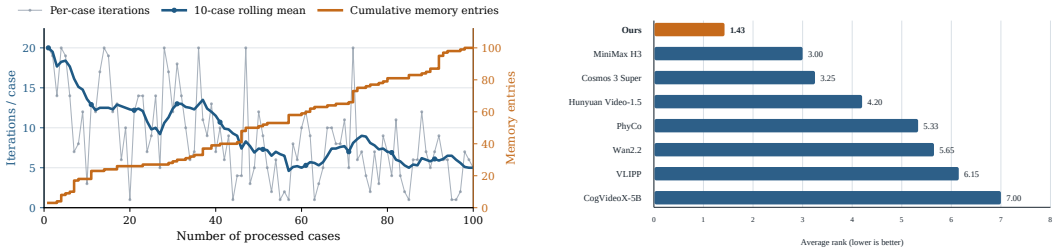
Figure 4: **Qualitative ablation of Spatiotemporal Causal Constraints.** Without causal constraints, repair may correct the immediate target while leaving causally affected states unresolved or disrupting valid interactions. Full PhysRepair jointly constrains editable and preserved states across the interaction, enabling complete and localized physical repair.

The wrench–box example requires a longer interaction sequence: the wrench falls toward the box, makes contact, and subsequently rebounds. Competing methods often realize only part of this process, e.g., failing to establish effective contact (Wan2.2) or introducing substantial geometry and trajectory inconsistencies (VLIPP). More specifically, MiniMax H3 exhibits a spatiotemporal causal misalignment in which the target box begins moving before the actual collision, together with visible wrench–box interpenetration, whereas Cosmos 3 Super leaves the wrench stalled at the box instead of completing the expected rebound. In contrast, PhysRepair preserves the wrench and box identities while restoring a coherent fall–contact–rebound sequence. This result illustrates how causally constrained, simulation-guided repair can address both incorrect contact timing and subsequent motion inconsistencies while preserving unaffected content. Together, these cases highlight the benefit of identifying causally affected states and guiding their localized regeneration with physically plausible counterfactual dynamics, rather than correcting individual frames or objects in isolation.

User Study. We conduct a user study in which participants rank videos generated by the compared methods according to their perceived physical consistency. As shown in Fig. 5b, PhysRepair achieves the best average rank among all compared methods, indicating that its improvements are also reflected in human judgments of physical plausibility. These results complement the quantitative evaluation and support the effectiveness of our *generate–diagnose–repair* paradigm in producing perceptually more convincing physical dynamics.

4.3 Ablation and Analysis

Spatiotemporal Causal Constraints. We qualitatively examine how Spatiotemporal Causal Constraints affect localized physical repair. As shown in Fig. 4, without the causal constraint, the agent can often correct the most salient error on the directly involved object, but may leave causally affected states unresolved or alter valid interactions that should remain fixed. In the goalkeeper example, MiniMax H3 produces a contradictory outcome in which the shot is both saved and scored. Without the causal constraint, the duplicated ball is removed, but a residual ball artifact in the goal and continued net motion still imply a scoring event. The full model additionally corrects these downstream states, yielding a clean save. In the ball-collision example, the source video exhibits an implausible collision between the soccer ball and the blue ball. Removing the causal constraint resolves the collision but disrupts the preceding foot–ball contact, causing the ball to move without a valid kick. In contrast, the full model preserves the valid kick while correcting the subsequent collision dynamics. These cases show that causal constraints are important for both propagating corrections to affected states and preserving valid states outside the repair scope, enabling complete yet localized physical repair.



(a) **Long-term memory growth and repair efficiency over the synthetic-case sequence.** The gray curve shows per-case repair iterations, the blue curve shows their ten-case rolling mean, and the orange staircase shows cumulative long-term memory entries.

(b) **User study on perceived physical consistency.** PhysRepair achieves the best average rank among eight compared methods, indicating that users consistently perceive its generated videos as more physically plausible.

Figure 5: **Analysis of PhysRepair.** Memory efficiency and user-perceived physical consistency of PhysRepair.

Experience-Augmented In-Context Learning. We examine how accumulated cross-sample experience affects iterative repair efficiency. As shown in Fig. 5a, the moving average of repair rounds decreases as long-term experience memory grows, despite variation across individual cases. Difficult cases can still exhaust the 20-round budget, but such cases become less frequent as experience accumulates. The results suggest that experience reuse can reduce repeated exploration and improve repair efficiency without updating model weights. This trend is consistent with the intended role of our Experience-Augmented In-Context Learning: combining current repair feedback with relevant prior experience to guide intervention selection and recovery from unsuccessful attempts.

5 Limitations and Broader Impact

Limitations. PhysRepair currently relies on a capable multimodal language model as the repair agent for physical diagnosis, tool selection, causal reasoning, verification, and experience reuse. Consequently, its performance is partly bounded by the agent model's physical reasoning and tool-use capabilities, and weaker agent backbones may lead to less reliable repair decisions. This dependence also introduces additional inference cost compared with one-pass video generation, particularly when multiple repair rounds are required.

Broader Impact. PhysRepair provides a general mechanism for improving the physical plausibility of generated videos without retraining the underlying video generator, which may benefit applications that require more reliable physical dynamics, such as simulation, embodied AI, and synthetic data generation. At the same time, increasingly realistic physical video generation and repair may make synthetic content more difficult to distinguish from real recordings, potentially increasing risks of deceptive or misleading media. We therefore view provenance mechanisms and responsible deployment practices as important complements to further improvements in physically grounded video generation.

6 Conclusion

We presented PhysRepair, an agentic framework for improving the physical plausibility of generated videos through an iterative generate–diagnose–repair process. Rather than imposing physical guidance entirely before synthesis, PhysRepair diagnoses the realized failure, uses Spatiotemporal Causal Constraints to localize the causally affected states. Under this constraint, the Physics Repair Agent derives a physically admissible counterfactual evolution for the affected interaction, which guides localized repair while preserving unaffected content. Experience-augmented In-context Learning further adapts repair decisions using current-sample feedback and relevant experience from previous repairs. Experiments on Physics-IQ Verified across five physics categories demonstrate consistent improvements in physical plausibility, suggesting post-hoc adaptive repair as a complementary direction for physically grounded video generation.

AI USE STATEMENT

In this work, we used generative AI tools to aid and polish writing, assist with literature retrieval, provide feedback on experiments, assist with method implementation, translation, and support the interpretation of experimental results. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Hritik Bansal, Zongyu Lin, Tianyi Xie, Zeshun Zong, Michal Yarom, Chenfanfu Jiang, Yizhou Sun, Kai-Wei Chang, and Aditya Grover. Videophy: Evaluating physical commonsense for video generation. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2406.03520.

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. *arXiv preprint arXiv:2503.06800*, 2025. URL https://arxiv.org/abs/2503.06800.

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.

Shaowei Liu, Zhongzheng Ren, Saurabh Gupta, and Shenlong Wang. Physgen: Rigid-body physics-grounded image-to-video generation. In *European Conference on Computer Vision*, 2024. URL https://arxiv.org/abs/2409.18964.

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

Saman Motamed, Laura Culp, Kevin Swersky, Priyank Jaini, and Robert Geirhos. Do generative video models learn physical principles from watching videos? *arXiv preprint arXiv:2501.09038*, 2025. URL https://arxiv.org/abs/2501.09038.

Saman Motamed, William Harvey, Benjamin Klein, Luc Van Gool, Zhuoning Yuan, and Ta-Ying Cheng. Void: Video object and interaction deletion. *arXiv preprint arXiv:2604.02296*, 2026. URL https://arxiv.org/abs/2604.02296.

Sriram Narayanan, Ziyu Jiang, Srinivasa G. Narasimhan, and Manmohan Chandraker. PhyCo: Learning controllable physical priors for generative motion. *arXiv preprint arXiv:2604.28169*, 2026. URL https://arxiv.org/abs/2604.28169.

Minh-Loi Nguyen, Xuan-Vu Le, Thanh-Toan Do, Tam V. Nguyen, Minh-Triet Tran, and Trung-Nghia Le. Physics-aware video generation via agentic planning and graph-guided optimization. *arXiv preprint arXiv:2609.13006*, 2026. URL https://arxiv.org/abs/2609.13006.

NVIDIA. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025. URL https://arxiv.org/abs/2501.03575.

NVIDIA. Cosmos 3: Omnimodal world models for physical ai. *arXiv preprint arXiv:2606.02800*, 2026. URL https://arxiv.org/abs/2606.02800.

Tim Rädsch, Yuki M. Asano, Hilde Kuehne, Stefan Bauer, Priyank Jaini, Robert Geirhos, and Carsten T. Lüth. Physics-iq verified. *arXiv preprint arXiv:2606.18943*, 2026. URL https://arxiv.org/abs/2606.18943.

Sand AI. MAGI-1: Autoregressive video generation at scale. *arXiv preprint arXiv:2505.13211*, 2025. URL https://arxiv.org/abs/2505.13211.

Noah Shinn, Federico Cassano, Ashwin Berman, Ashay Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2303.11366.

Wan Team. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. URL https://arxiv.org/abs/2503.20314.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*, 2023. URL https://arxiv.org/abs/2305.16291.

Bing Wu, Chang Zou, Changlin Li, Duojun Huang, Fang Yang, Hao Tan, Jack Peng, Jianbing Wu, Jiangfeng Xiong, Jie Jiang, et al. Hunyuanvideo 1.5 technical report. *arXiv preprint arXiv:2511.18870*, 2025. URL https://arxiv.org/abs/2511.18870.

Wenzhuo Xu, Yuchen Zhu, Chongjian Ge, Xuan Shen, Jing Shi, Jason Kuen, Yongxin Chen, Molei Tao, Christopher McComb, Noelia Grande Gutiérrez, and Jiuxiang Gu. Veriphy: Agentic physical reasoning for world model evaluation and refinement. *arXiv preprint arXiv:2609.03153*, 2026. URL https://arxiv.org/abs/2609.03153.

Qiyao Xue, Xiangyu Yin, Boyuan Yang, and Wei Gao. Phyt2v: Llm-guided iterative self-refinement for physics-grounded text-to-video generation. *arXiv preprint arXiv:2412.00596*, 2025. URL https://arxiv.org/abs/2412.00596.

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 *International Conference on Computer Vision*, 2025. URL https://arxiv.org/abs/2503.23368.

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. *arXiv preprint arXiv:2408.06072*, 2024. URL https://arxiv.org/abs/2408.06072.

Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B. Tenenbaum. Clevrer: Collision events for video representation and reasoning. In *International Conference on Learning Representations*, 2020. URL https://arxiv.org/abs/1910.01442.