PHYSIS-LANG: SELF-EVOLVING LANGUAGE AS A PHYSICAL REPRESENTATION FOR VIDEO WORLD MODEL

Anonymous authors
Paper under double-blind review

ABSTRACT

Video world models are expected to predict how the physical world evolves, yet they often produce visually plausible videos that violate basic physical principles. Existing approaches commonly assume that natural language is insufficient to represent the physical knowledge required for reliable generation, and therefore introduce additional visual, latent, numerical, or planning-based signals. We revisit this assumption and introduce *Physis-Lang*, a self-evolving framework that treats physical language as a shared and optimizable representation across data curation, model training, and video generation. Physis-Lang represents physical processes through language that describes their relevant entities, causes, interactions, governing principles, temporal evolution, and effects. To improve this representation, we construct *PhysCapBench*, which decomposes physical processes into atomic assertions and evaluates captions using recall and precision. An agentic loop iteratively analyzes assertion-level errors and refines the instruction used to produce physical captions. Physis-Lang further converts model deficiencies into textual descriptions and uses language-guided retrieval to identify visually diverse videos that cover missing physical processes. Experiments on four widely used physical video benchmarks with Wan and Cosmos backbones demonstrate consistent improvements in physical plausibility. Notably, starting from open-source Cosmos3-Nano backbones, our Physis-Lang-enhanced models surpass the leading proprietary Veo 3.1 model.

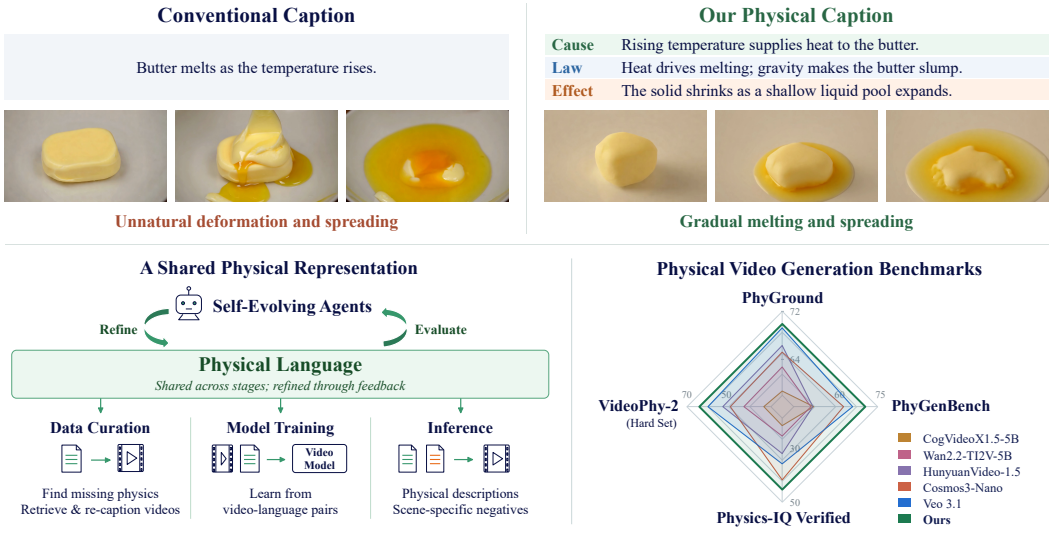


Figure 1: **Overview of Physis-Lang**. Existing approaches treat language mainly as a conditioning interface and introduce physical knowledge through separate visual, latent, numerical, or planning-based signals. Physis-Lang instead treats physical language as a shared and optimizable representation: a self-evolving agent loop evaluates and refines physical descriptions, language-guided retrieval expands the training data toward missing physical processes, and the evolved language supervises training and guides inference toward physically plausible video generation.

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.

1 INTRODUCTION

Video world models are expected not only to synthesize visually appealing videos, but also to predict how the physical world evolves. This capability is fundamental to embodied intelligence, robotics, autonomous systems, and interactive simulation, where a model must anticipate the consequences of actions rather than merely render plausible-looking frames. However, conventional video–text training primarily optimizes visual fidelity and semantic alignment, without explicitly representing the physical knowledge underlying observed dynamics; as a result, current video generation models still frequently violate basic physical principles: objects deform or disappear unexpectedly, collisions produce implausible outcomes, fluids move unnaturally, and causal events occur in the wrong order. Recent benchmarks have further shown that perceptual realism does not necessarily imply physical understanding (Meng et al., 2025; Bansal et al., 2026; Motamed et al., 2026; Zhou et al., 2026a). Improving the *physical plausibility* of video world models therefore remains a central challenge on the path toward general-purpose world simulation.

Existing efforts improve the physical plausibility of video generation by complementing the standard language interface with additional sources of physical information. They enrich video–text data with physics-focused videos or physical annotations (Li et al., 2025; Zhou et al., 2026b), introduce auxiliary signals about geometry, motion, or physical correctness during training (Shen et al., 2026; Yuan et al., 2026b), and provide structured plans, retrieved examples, or external judgments during generation (Hao et al., 2025; Zhuang et al., 2026). Despite their different implementations, these approaches reflect a common design choice: natural language is primarily used to specify semantic content, while more detailed physical information is conveyed through complementary visual, geometric, latent, numerical, or human-designed signals. These complementary signals can provide useful supervision for aspects of physical processes that language alone may not fully capture. Yet an important possibility remains relatively unexplored: *can language itself, when made more explicit and systematically optimized, serve as a unified representation of physical knowledge across the video-generation pipeline?*

Motivated by this question, we hypothesize that the limited role of language in current video models may stem from language not yet being fully exploited as a representation of physical processes, rather than from an intrinsic limitation of language itself. Based on this view, we introduce *Physis-Lang*, a self-evolving agentic framework that constructs, evaluates, and refines physical language as a shared representation, as illustrated in Fig. 1. At the caption level, Physis-Lang goes beyond describing visible actions and makes the relevant physical entities, causal interactions, governing principles, and resulting effects explicit. The self-evolving representation is then reused throughout the pipeline: it guides data curation by identifying missing physical domains and selecting relevant videos, provides supervision for training on video–language pairs, and guides inference through scene-specific physical descriptions. In this way, the three components shown at the bottom of Fig. 1 are connected by the shared physical language at its center, enabling the system to organize, evaluate, and transfer physical knowledge across data curation, training, and inference.

To build this shared representation, we first instantiate it at the caption level through an agentic caption-evolution loop grounded by our *physics-aware critic* and *PhysCapBench*. Given a fixed captioning model and an initial instruction, the model generates candidate physical descriptions for videos in our development set. For each video, we provides human-curated reference assertions organized around *cause*, *physical law*, and *effect*. Using these references, our critic evaluates each candidate description along two dimensions: its coverage of the relevant physical knowledge and the faithfulness of its individual claims to the visual evidence. An evolution agent uses this diagnostic feedback to identify omissions, unsupported claims, and vague causal relations, and then revises the captioning instruction without updating the captioner's weights. Each instruction is then evaluated on our PhysCapBench within the same loop, which continues until benchmark performance saturates. The best-performing instruction is then used to re-caption the training corpus. Thus, our physics-aware critic provides the loop feedback and PhysCapBench provides stopping criterion for language evolution, while Physis-Lang turns that feedback into improved physical descriptions for video-model training and inference.

The self-evolving physical language also enables Physis-Lang to expand the training distribution toward physical domains that the target video model handles poorly. It first summarizes the model's failures into a category-level deficiency profile. These deficient domains are then matched against

physics tags associated with videos in a large gallery. Unlike retrieval based solely on visual similarity, this language-guided matching selects videos according to their physical content rather than their appearance, allowing the data engine to target missing physical domains and collect visually diverse examples relevant to the same deficiency. The selected videos are re-captioned with the self-evolving guidelines and added to the training corpus, directing additional supervision toward the model's weaknesses. Together with positive physical reasoning and scene-specific negative descriptions at inference time, this language-guided data engine connects targeted data curation, model training, and generation through a shared physical-language representation.

Our main contributions are summarized as follows:

- We introduce a new perspective that treats **language as a physical world representation**. We show that language, when explicitly structured and self-evolving, can serve as a representation for carrying, refining, and applying physical knowledge in video world models.
- We develop a **self-evolving agent system** that optimizes physical-language guidelines using a physics-aware critic. We introduce **PhysCapBench** to evaluate physical coverage and claim faithfulness and to guide iterative refinement.
- We build a **language-guided data engine** that uses physics-domain tags and text-level matching to select training videos relevant to model deficiencies, then enriches them with self-evolving physical captions.
- Extensive experiments on four physical video benchmarks demonstrate significant and consistent improvements on both Wan and Cosmos backbones. Our resulting models also outperform Veo 3.1 under the same evaluation protocols, validating language as a practical representation for improving video world models.

2 Related Work

Physics-Aware Video Generation. Recent efforts to enhance the physical plausibility of video generative models can be categorized by the source of the physical priors they exploit. The largest line of work draws physical priors from pretrained vision-language or foundation models, utilizing them to distill implicit physical dynamics into auxiliary branches (Bhowmik et al., 2026; Satish et al., 2026; Shen et al., 2026; Lin et al., 2026b; Song et al., 2025; Kim et al., 2026a; Wang et al., 2026b), retrieve physically grounded reference motions (Peruzzo et al., 2025; Zhu et al., 2025; Cheng et al., 2026), verify or critique generated contents (Hao et al., 2025; Kim et al., 2026b; Tang et al., 2026; Yuan et al., 2026a; Zhao et al., 2026b; Zhang et al., 2025; Xue et al., 2025; Liu et al., 2025), or plan over intermediate representations before rendering (Pathak et al., 2026; Feng et al., 2026; Wang et al., 2026c; Huang et al., 2026b; Shang et al., 2026). A second line treats physical plausibility as an objective to be optimized via reinforcement learning or preference optimization, aligning the generator with reward or preference signals derived from physics-aware judges or contrastive trajectories (Chen et al., 2025a; Wang et al., 2025b; Zhao et al., 2026a; Zhu et al., 2026; Meng et al., 2026; Li et al., 2026b; Wu et al., 2026; Wang et al., 2026a; Meyarian et al., 2026). A third line of works injects explicit graphics- or simulator-derived structures like physical equations, trajectories, or executable simulation code into the generation pipeline (Yuan et al., 2026b; Chen et al., 2025b; Luan & Shi, 2026; Baumann et al., 2026; Stracke et al., 2026; Zhuang et al., 2026; Zhang et al., 2026a; Xu et al., 2026; Bai et al., 2026; Li et al., 2026a). Some works also curate large-scale physics-annotated datasets (Zhou et al., 2026b; Li et al., 2025; 2026c).

Despite these advances, most existing approaches rely on auxiliary mechanisms, such as reward models, intermediate features, or retrieved samples, to incorporate physical knowledge. Simulation-based approaches, on the other hand, are constrained by high computational costs and limited generalizability. In contrast, we posit that language provides a compact, explicit, and transferable representation of physical knowledge. Building on this view, we introduce an agentic self-evolution prompting scheme that iteratively refines the language representation to better capture physical priors. Different from the views of many recent works, we prove that ***language alone*** is very powerful to lead to physically plausible video generation if we utilize it properly.

Agentic Self-Evolution. Self-evolution has emerged as a distinct paradigm for improving agents, which has been applied to a range of tasks, including web navigation (Putta et al., 2024; Fang et al.,

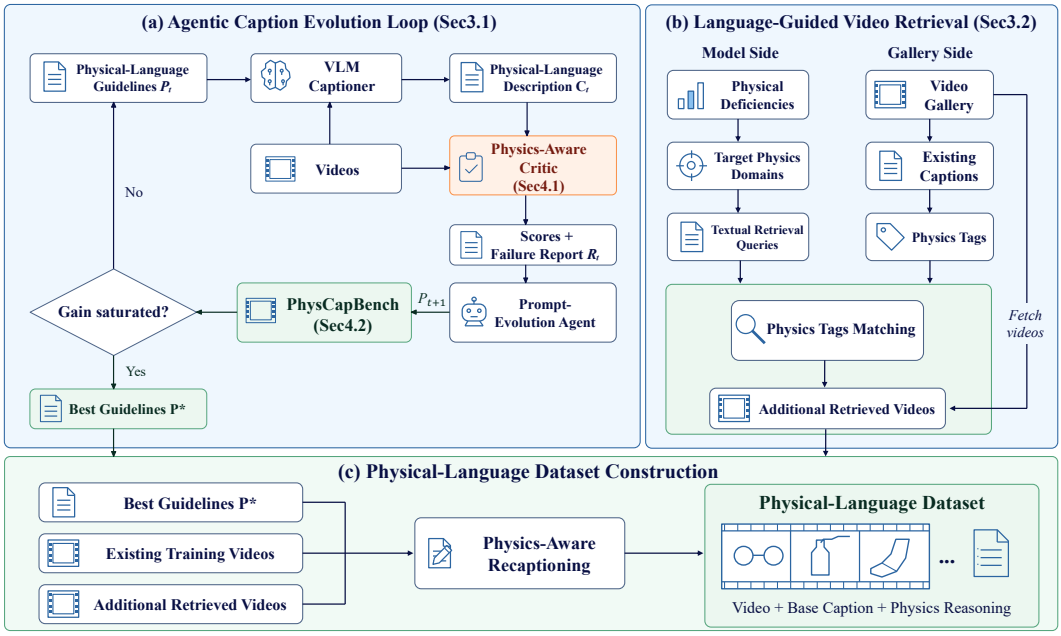


Figure 2: **Overview of our method**. Physical processes are represented through natural language, including a base caption, explicit physics reasoning, and scene-adaptive negative physics descriptions. A PhysCapBench-driven agentic loop evaluates and refines this physical language, which is then used to expand the training data, supervise video-model training, and guide inference toward physically plausible generation.

2025; Murty et al., 2024; He et al., 2025), mathematical and code reasoning (Huang et al., 2026a; Yuksekgonul et al., 2024; Agrawal et al., 2026; Suzgun et al., 2026; Zhang et al., 2026c; Yin et al., 2025; Ling et al., 2026), general instruction-following and question answering (Wang et al., 2023; Xu et al., 2024), as well as interactive tool use (Zhai et al., 2025; Yang et al., 2026; Zhang et al., 2026b). Across these tasks, an agent's prompts, policy, memory, or tools are iteratively refined. Representative methods involve revising subsequent attempts through verbal self-critique of past failures (Shinn et al., 2023; Asai et al., 2024; Hu et al., 2025), using the model itself as a judge to construct preference data for iterative alignment training (Aksitov et al., 2023; Yuan et al., 2024), or training agents via multi-turn reinforcement learning within an interactive environment (Wang et al., 2025c; Chen et al., 2026; Hu et al., 2026). In contrast, we are the first to apply agentic self-evolution to physics reasoning that ultimately enhances the physical realism of video generation tasks. We establish a physics caption benchmark to evaluate the agents, which allows the upsampled prompts to converge toward a format that is optimally suited for generating physically realistic videos.

3 METHOD

3.1 SELF-EVOLVING AGENTS TO IMPROVE PHYSICS CAPTION

Fig. 2(a) demonstrates the overview of our agentic caption evolution loop. Conventional captions often describe scene semantics without explaining physical causes and consequences. We make these mechanisms explicit in language so that video–text training can provide richer supervision for physical dynamics. To this end, we augment base caption with a natural-language ***physics_reasoning*** field describing physical details, dynamics, and causal relations. The same representation supports training supervision and image/text-conditioned prompt expansion at inference. Because captioning instructions can omit key details or induce unsupported reasoning, we use a self-evolving agent to iteratively refine the instruction while keeping the captioning model fixed.

To enable systematic prompt evolution, we first develop a **physics-aware critic** that evaluates the correctness and completeness of physical descriptions. The detailed protocol is introduced in Sec. 4.1. Based on this critic, we further introduce **PhysCapBench**, a held-out benchmark of 246

physical videos for evaluating physics-aware video captioning. It serves both as a standalone benchmark and as the validation set for prompt evolution.

We then construct an agentic self-evolution loop over a fixed 20-video development set. At iteration t, a fixed captioner uses prompt P_t to generate captions $\mathcal{C}_t$. The critic produces a score and diagnostic report R_t, which summarizes errors and guides an evolution agent to revise the prompt: $P_t \rightarrow \mathcal{C}_t \rightarrow R_t \rightarrow P_{t+1}$. After each iteration, the updated prompt is evaluated on PhysCapBench. We continue the loop while benchmark performance improves and stop once the gain saturates, selecting the best-performing prompt as the final physics-captioning instruction. Detailed implementation is provided in App. B.2.

During training, we apply the final captioner to re-caption the collected videos with physics-rich descriptions. During inference, we construct a corresponding upsampler that expands short text prompt and an input image (optional) into a detailed physics-aware prompt, providing the video generator with more explicit physical guidance and improving the physical realism of the generated video. We further derive a scene-specific ***physics_negative_prompt*** describing likely physically implausible evolutions, which serves as negative conditioning to suppress physical inconsistencies and improve generation realism at inference time.

3.2 Language-Guided Video Retrieval

Beyond improving the linguistic representation of physical processes, we further leverage language as a key modality for targeted data expansion, using it to identify and collect training examples that address the physical deficiencies of pretrained video generative models. Fig. 2(b) illustrates the framework of this data collecting pipeline.

We first build a GPT-5.5-based diagnosis agent to identify physical failures in generated videos and assign each failure to physics categories, such as rigid-body motion, collision and fluid dynamics. Aggregating these results yields a category-level deficiency profile that captures the model's dominant physical weaknesses. We then utilize this profile to retrieve targeted training data from a large video gallery, where each candidate video is tagged with physics-domain labels derived from its caption, enabling retrieval that matches the deficient categories. The retrieved videos are added to the training set and re-captioned with our physics-aware pipeline in Sec. 3.1. In this way, language connects model diagnosis, data indexing, and targeted data acquisition, allowing the training distribution to be adapted to the generator's observed physical weaknesses.

3.3 PhysThinker for Physics Reasoning

The VLM captioner and upsampler introduced in Sec. 3.1 are initially instantiated with proprietary models. While these models provide strong physics reasoning capability, relying on them for large-scale video re-captioning and inference-time upsampling incurs high monetary cost.

To address this issue, we replace the costly GPT-based captioner and upsampler by distilling their physics reasoning capabilities into two efficient 4B vision-language models: **PhysThinker-C** for physics-aware video captioning and **PhysThinker-U** for inference-time prompt upsampling. PhysThinker-C learns from GPT-generated video-to-text annotations, while PhysThinker-U is trained to expand short text or image-text conditions into physics-rich prompts. Together, they provide scalable and low-cost alternatives to commercial models for large-scale annotation and inference. Training details of our PhysThinker are provided in App. B.5.3.

4 Benchmark and Critic

4.1 Critics for Physics Caption

Fig. 3(b) illustrates the overview of our physics-aware critic used in the agentic loop. To systematically evaluate whether a video caption faithfully and comprehensively represents the physical dynamics in a video, inspired by the caption evaluation protocol used by Cosmos 3 (NVIDIA et al., 2026), we adopt **precision** and **recall** on the atomic assertions as critics, and specialize both metrics to physical content. Precision evaluates whether generated physical claims are visually supported by the video, while recall measures coverage of salient physical processes.

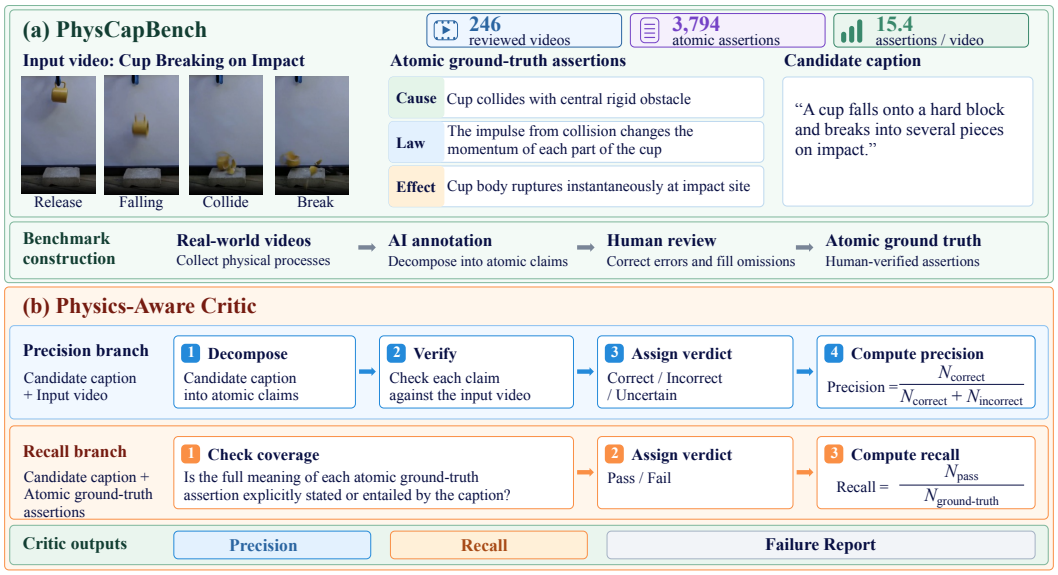


Figure 3: **Overview of our physics-aware critic and physics caption benchmark**. The critic evaluates generated captions from complementary precision and recall perspectives by verifying atomic claims against the input video and measuring the coverage of human-curated physical assertions.

For evaluation of **precision**, the critic first decomposes the complete generated caption into atomic and independently verifiable claims. The critic is then given the full video together with each atomic claim and classifies it as *correct*, *incorrect*, or *uncertain*. We compute micro-averaged precision as $\text{Precision} = N_{\text{correct}}/(N_{\text{correct}} + N_{\text{incorrect}})$. **Recall** instead focuses specifically on physical dynamics. The critic receives the generated caption together with a set of human-curated atomic physical assertions and determines whether each ground-truth assertion is sufficiently covered by the caption. An assertion receives a positive match only when the caption explicitly states or unambiguously entails its complete physical meaning. Recall is therefore computed as $\text{Recall} = N_{\text{pass}}/N_{\text{ground-truth}}$. We finally report the F1 score, defined as the harmonic mean of precision and recall, $\text{F1} = 2 \cdot \text{Precision} \cdot \text{Recall}/(\text{Precision} + \text{Recall})$, as an overall measure of caption quality.

4.2 Physics Caption Benchmark

Based on the critic, we further introduce **PhysCapBench** (Fig. 3(a)), a dedicated benchmark for evaluating whether video captioning models can faithfully and comprehensively describe salient physical phenomena and their underlying dynamics, consisting of 246 physics-rich videos collected from Physics-IQ (Motamed et al., 2026) and YouTube. We construct fine-grained physical annotations for each video using an agent-assisted, human-verified pipeline. AI annotators identify salient physical processes and decompose them into atomic assertions describing causes, relevant physical laws and effects. All candidate annotations are subsequently reviewed by humans, who remove incorrect or ambiguous assertions and supplement missing ones when necessary.

The final PhysCapBench contains **3,794 human-verified physical assertions** over 246 videos, with an average of 15.4 assertions per video. These annotations provide fine-grained supervision for evaluating whether a caption captures the key physical dynamics and reasoning in each video. Detailed implementations are provided in App. B.1, and benchmark samples are listed in App. E.

5 Experiment

5.1 Experimental Details

Dataset and Backbones. At the core of Physis-Lang is the construction of a physics-enriched video–language training dataset comprising 183K real-world videos. It combines 71K high-quality samples filtered from WISA-80K (Wang et al., 2025a) with 112K additional videos selected through

Table 1: Physical fidelity on Physics-IQ Verified.

Model	ST	S	WS	MSE	Final
CogVideoX1.5-5B [66]	17.41	34.49	21.71	16.18	22.45
Wan2.2-TI2V-5B [53]	23.26	33.89	23.18	23.19	25.88
HunyuanVideo-1.5 [52]	22.80	47.10	30.46	26.05	31.60
Veo 3.1	19.01	52.16	35.45	33.32	34.99
Kandinsky-WM [22]	25.69	36.37	17.22	12.69	22.99
Self-Refinement [32]	22.49	37.93	23.25	25.13	27.20
PhiZero [45]	56.79	37.01	27.59	42.29	40.91
Cosmos3-Nano	29.38	51.76	38.92	40.87	40.23
Ours (Cosmos3-Nano)	33.59	55.38	42.45	42.23	**43.41**

Table 2: Physical adherence on PhyGround.

Model	General	Physics	Overall
CogVideoX1.5-5B [66]	59.02	58.36	58.68
Wan2.2-TI2V-5B [53]	62.94	62.46	62.70
HunyuanVideo-1.5 [52]	67.14	65.42	66.28
Veo 3.1	71.92	66.54	69.24
Kandinsky-WM [22]	58.96	57.28	58.12
Self-Refinement [32]	58.54	57.88	58.22
PhiZero [45]	52.27	63.42	57.85
Cosmos3-Nano	65.46	64.90	65.18
Ours (Cosmos3-Nano)	71.54	68.26	**69.90**

Table 3: Physical correctness on PhyGenBench.

Model	Force	Light	Heat	Material	Overall
CogVideoX1.5-5B [66]	38.33	57.33	55.56	43.33	48.75
Wan2.2-TI2V-5B [53]	40.83	60.67	43.33	40.83	47.50
HunyuanVideo-1.5 [52]	47.50	56.67	44.44	41.67	48.33
Veo 3.1	65.00	70.67	57.78	65.83	65.63
PhysVid [40]	48.33	62.00	33.33	40.83	47.92
PhyGDPO [8]	43.33	62.67	43.33	41.67	48.96
Self-Refinement [32]	46.67	59.33	44.44	42.50	49.17
Cosmos3-Nano	64.17	64.00	57.78	59.17	61.67
Ours (Cosmos3-Nano)	67.50	72.67	75.56	69.17	**71.04**

Table 4: Physical commonsense on VideoPhy-2.

Model	SA	PC	SA≥4	PC≥4	All	Hard
CogVideoX1.5-5B [66]	3.74	4.10	63.79	74.79	49.41	33.15
Wan2.2-TI2V-5B [53]	3.69	4.39	60.91	85.45	53.98	42.13
HunyuanVideo-1.5 [52]	3.84	4.57	67.85	92.39	62.94	51.69
Veo 3.1	4.39	4.06	84.43	80.54	**68.87**	58.43
PhysVid [40]	3.50	4.27	55.33	82.74	48.22	29.78
PhyGDPO [8]	3.74	4.64	62.27	93.57	59.56	44.94
Self-Refinement [32]	3.60	4.26	56.01	82.91	47.88	28.09
Cosmos3-Nano	3.86	4.49	66.33	90.19	60.41	48.31
Ours (Cosmos3-Nano)	4.09	4.44	75.97	89.00	68.02	**62.36**

our deficiency-guided retrieval pipeline (Sec. 3.2). We use the dataset to fine-tune several existing video-generation backbones without modifying their architectures and training objectives. Specifically, we consider Wan2.1-14B (Wan et al., 2025), for which the T2V-14B and I2V-14B variants are fine-tuned separately, and the Cosmos3 family (NVIDIA et al., 2026), including Edge (4B), Nano (16B), and Super (64B), whose unified backbone supports both T2V and I2V generations. Further implementation details are provided in App. B.

Baselines and Evaluation. Our comparison involves cutting-edge pretrained models, including CogVideoX1.5-5B (Yang et al., 2025), Wan2.2-TI2V-5B (Wan et al., 2025), HunyuanVideo-1.5 (Tencent Hunyuan Foundation Model Team, 2025), and the closed-source Veo 3.1. We also compare with other methods for promoting physical realism, including PhysVid (Pathak et al., 2026), PhyGDPO (Cai et al., 2026), Self-Refinement (Liu et al., 2025), Kandinsky-WM (Kandinsky Lab, 2026), and PhiZero (Shang et al., 2026). The evaluation is conducted on four widely used physical benchmarks: VideoPhy-2 (Bansal et al., 2026) and PhyGenBench (Meng et al., 2025) for Text-to-Video (T2V) generation, and PhyGround (Lin et al., 2026a) and Physics-IQ Verified (Rädsch et al., 2026) for Image-to-Video (I2V) generation. To better distinguish differences in generation quality, we replace the original offline VLM evaluators of VideoPhy-2 and PhyGenBench with GPT-5.5, which provides more discriminative assessments in our comparisons (App. G). We evaluate all models using the same scoring protocol within each benchmark to ensure fair comparisons.

5.2 Comparison with State-of-the-Art Models

Tables 1–4 compare Cosmos3-Nano fine-tuned using our Physis-Lang with cutting-edge generative models and other baselines designed to improve physical realism in video generation. For better comparison, we rescale each benchmark to a 100-point scoring scale. We highlight three key advantages of our method: (i) **Consistent and significant improvement over base model.** Building on pretrained Cosmos3, Physis-Lang further improves its performance across all four evaluated physical video benchmarks. (ii) **State-of-the-art among open-source models.** Our method achieves the strongest performance among all open-source models and methods across the four benchmarks. (iii) **Outperforming the best closed-source model.** Our method surpasses Veo 3.1 on Physics-IQ Verified, PhyGround, and PhyGenBench, and remains competitive on VideoPhy-2, trailing by only 0.85 points on the full set (68.02 vs. 68.87) while outperforming it by 3.93 points on the hard split (62.36 vs. 58.43). Qualitative examples in App. C demonstrate the improved generation quality of our method across the evaluated benchmarks, while App. D further shows our strong performance in driving and robotics scenarios.

Table 5: **Effectiveness on different backbones.** Each cell shows the improvement brought by our method. Mean Δ is the average improvement across the four benchmarks.

Generator	PhyGenBench	Physics-IQ Verified	VideoPhy-2	PhyGround	Mean Δ
Across model families					
Wan2.1-14B	56.67 → 65.83	27.87 → 35.15	57.02 → 65.65	61.52 → 64.64	+7.05
Cosmos3-Nano-16B	61.67 → 71.04	40.23 → 43.41	60.41 → 68.02	65.18 → 69.90	+6.22
Across Cosmos scales					
Cosmos3-Edge-4B	48.96 → 52.50	32.80 → 34.69	32.99 → 40.61	66.66 → 66.56	+3.24
Cosmos3-Nano-16B	61.67 → 71.04	40.23 → 43.41	60.41 → 68.02	65.18 → 69.90	+6.22
Cosmos3-Super-64B	66.04 → 70.21	45.92 → 50.00	60.91 → 72.42	68.67 → 68.98	+5.02

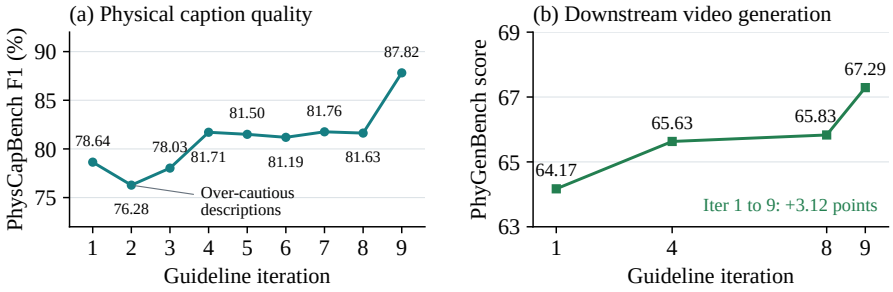


Figure 4: **From caption evolution to video generation.** Left: PhysCapBench F1 at different iterations. Right: physics-aware video generation performance of the same pretrained Cosmos3 model with inference captions produced at different iterations.

5.3 Generalization across Backbones and Scales

Tab. 5 evaluates whether our method generalizes across model families and parameter scales. We draw three conclusions: (i) **Consistent gains across model families.** Our method improves Wan2.1-14B and Cosmos3-Nano-16B by 7.05 and 6.22 points on average, respectively, with gains on all four physical benchmarks. (ii) **Effectiveness across model scales.** Within the Cosmos 3 family, our method yields average gains of 3.24, 6.22, and 5.02 points on Edge-4B, Nano-16B, and Super-64B, respectively. These results demonstrate benefits across a broad range of model sizes. (iii) **Preserved general video quality.** Our fine-tuning strategy does not compromise general video generation quality; please refer to App. H for more details.

5.4 Ablation and Analysis

5.4.1 Agentic Caption Evolution

Fig. 4 evaluates whether caption evolution improves both physical descriptions and downstream video generation. We highlight two observations: (i) **The agentic loop improves physical caption quality.** On PhysCapBench, the F1 score increases from 78.64 at Iteration 1 to 87.82 at Iteration 9, showing the effectiveness of critic-guided guideline refinement. (ii) **The self-evolving guidelines improve video generation.** Keeping the pretrained Cosmos3-Nano fixed, we change only the inference captions produced using guidelines from Iterations 1, 4, 8 and 9. PhyGenBench scores increase from 64.17 to 65.63, 65.83 and 67.29, respectively. App. A provides the complete evolution trajectory. App. F analyses the effect of *physics_reasoning* derived from our agentic loop.

5.4.2 Language-Guided Data Curation

Language-guided data curation (Sec. 3.2) consistently improves physical generation. Adding retrieved videos to the training set improves benchmarks, for an average gain of 3.01 points (Tab. 6). To further examine whether these gains are consistent across our physical tags, we break down the improvements on VideoPhy-2 by physical category (a sample may belong to multiple categories). The gains are positive across all frequently represented categories, such as chemical processes (+8.00), fracture mechanics (+7.45), and cloth deformation (+7.19 points) (Fig. 5).

Table 6: **Effect of retrieved training data.** Benchmark scores before and after data expansion.

Benchmark	WISA	+ Retrieved	Gain
PhyGenBench	68.12	71.04	+2.92
Physics-IQ Verified	40.68	43.41	+2.73
VideoPhy-2	64.63	68.02	+3.39
Mean	57.81	60.82	+3.01

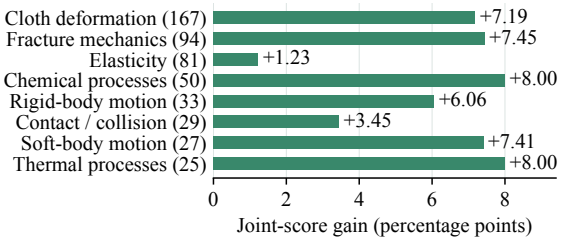


Figure 5: **VideoPhy-2 category gains** after retrieval.

Table 7: **Prompting and SFT on PhyGenBench.** Base: base caption; M: manually designed physics caption; P: ***physics_reasoning***; N: ***physics_negative_prompt***.

ID	SFT data	Prompt	Score	ID	SFT data	Prompt	Score
$\mathbb{A}$	None	Base	61.67	$\mathbb{E}$	WISA	Base + P	65.62
$\mathbb{B}$	None	M	61.86	$\mathbb{F}$	WISA	Base + P + N	68.12
$\mathbb{C}$	None	Base + P	63.33	$\mathbb{G}$	WISA + retrieved	Base + P	66.88
$\mathbb{D}$	None	Base + P + N	**67.29**	$\mathbb{H}$	WISA + retrieved	Base + P + N	**71.04**

5.4.3 Prompting, Fine-Tuning, and Negative Guidance

Tab. 7 highlights three findings: (i) **Physical prompting is effective without fine-tuning.** Our physical reasoning prompt improves the pretrained model from 61.67 to 63.33 ($\mathbb{A} \rightarrow \mathbb{C}$), demonstrating a zero-shot gain. Moreover, our agentic loop further improves over the manually designed prompt (Cai et al., 2026)($\mathbb{B} \rightarrow \mathbb{C}$), validating the effectiveness. (ii) **SFT provides additional gains.** With inference prompts fixed, fine-tuning brings further improvement: 3.55 points ($\mathbb{C} \rightarrow \mathbb{G}$) and 3.75 points ($\mathbb{D} \rightarrow \mathbb{H}$). (iii) **Negative guidance complements SFT.** Adding physics negative prompts improves the fine-tuned model by 2.50 ($\mathbb{E} \rightarrow \mathbb{F}$) and 4.16 points ($\mathbb{G} \rightarrow \mathbb{H}$), respectively.

5.5 Open-Source Deployment with PhysThinker

We train PhysThinker-C and PhysThinker-U for video captioning and inference-time prompt upsampling, respectively. Tab. 8 demonstrates an accuracy–cost trade-off: commercial, captioner-only replacement, and fully local pipelines improve over the pretrained baseline by 7.05, 6.76, and 4.76 points on average, while estimated external annotation costs decrease from \$24.12K to \$0.12K and \$0. Costs cover the re-captioning of training data and the upsampling for benchmarks, whereas accuracy is averaged on four reported benchmarks.

Table 8: **PhysThinker replacement on Wan2.1-14B.** VideoPhy-2 reports All / Hard; Cost denotes API expenditure for training-data captioning and benchmark upsampling (USD).

Captioner	Upsampler	PhyGenBench	Physics-IQ Verified	VideoPhy-2	PhyGround	Mean Δ	Cost (USD)
Pretrained Wan2.1-14B		56.67	27.87	57.02 / 43.82	61.52	–	–
Commercial	Commercial	65.83	35.15	65.65 / 59.55	64.64	+7.05	∼24.12K
PhysThinker-C	Commercial	65.00	34.71	65.99 / 54.49	64.40	+6.76	∼0.12K
PhysThinker-C	PhysThinker-U	63.33	34.13	60.07 / 55.06	64.60	+4.76	0

6 Conclusion

Video world models require physical plausibility beyond visual quality. We introduce *Physis-Lang*, which treats structured physical language as a shared representation across data curation, model training, and inference. PhysCapBench and an agentic self-evolution loop refine physical captions, while language-guided retrieval targets missing physical processes. Experiments with Wan2.1 and Cosmos3 family models across four physical benchmarks of video generation show consistent and significant improvements over the pretrained models. These results show that language, when explicitly structured and self-evolving, can serve not merely as a conditioning interface, but as a practical representation for carrying, refining, and applying physical knowledge in video world models.

REPRODUCIBILITY STATEMENT

Sec. 3.2 describes deficiency-guided video retrieval, while Sec. 5.1 and App. B.5.1 detail the training-data composition and filtering procedure. Sec. 4.1–4.2 and App. B.1 describe the construction of PhysCapBench, human verification of atomic physical assertions, and the precision, recall, and F1 evaluation protocol, including the evaluator configuration. Sec. 3.1and App. B.2 describe the development set, captioning and critic models, prompt-revision loop, and adaptation to inference-time upsampling; App. A reports the prompt-evolution trajectory and identifies the best-performing iteration. App. B.3 specifies the sliding-window text encoding used for long prompts in Wan2.1. App. B.5.1–B.5.3 provide the training configurations for Cosmos3, Wan2.1, and PhysThinker, including preprocessing, trainable parameters, optimization settings, batch sizes, computational resources, and evaluation checkpoints. App. B.4 specifies the four video-generation benchmarks, their sample counts, scoring criteria, and aggregation procedures, while Sec. 5.1 and App. G document the use of GPT-5.5 as the evaluator for PhyGenBench and VideoPhy-2. App. F describes the attention-extraction and aggregation procedure used in our analysis.

AI USE STATEMENT

Large Language Models (LLMs) were used to aid in the writing and polishing of the manuscript. Specifically, we used an LLM to assist in refining the language, improving readability, and ensuring clarity in various sections of the paper. The model helped with tasks such as sentence rephrasing, grammar checking, and enhancing the overall flow of the text. All research concepts, ideas, and analyses were developed and conducted by the authors. The contributions of the LLM were solely focused on improving the linguistic quality of the paper. The authors take full responsibility for the content of the manuscript, including any text generated or polished by the LLM. We have ensured that the LLM-generated text adheres to ethical guidelines and does not contribute to plagiarism or scientific misconduct.

REFERENCES

Lakshya A. Agrawal, Shangyin Tan, Dilara Soylu, Noah Ziems, Rishi Khare, Krista Opsahl-Ong, Arnav Singhvi, Herumb Shandilya, Michael J. Ryan, Meng Jiang, Christopher Potts, Koushik Sen, Alexandros G. Dimakis, Ion Stoica, Dan Klein, Matei Zaharia, and Omar Khattab. GEPA: Reflective prompt evolution can outperform reinforcement learning. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=RQm2KQTM5r.

Renat Aksitov, Sobhan Miryoosefi, Zonglin Li, Daliang Li, Sheila Babayan, Kavya Kopparapu, Zachary Fisher, Ruiqi Guo, Sushant Prakash, Pranesh Srinivasan, Manzil Zaheer, Felix Yu, and Sanjiv Kumar. Rest meets react: Self-improvement for multi-step reasoning llm agent. *arXiv preprint arXiv:2312.10003*, 2023.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=hSyW5go0v8.

Xiangyu Bai, He Liang, Bishoy Galoaa, Utsav Nandi, Shayda Moezzi, Yuhang He, and Sarah Ostadabbas. MoReGen: Multi-agent motion-reasoning engine for code-based text-to-video synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7632–7642, 2026.

Hritik Bansal, Clark Peng, Yonatan Bitton, Roman Goldenberg, Aditya Grover, and Kai-Wei Chang. VideoPhy-2: A challenging action-centric physical commonsense evaluation in video generation. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=HA8KSQW7SO.

Stefan Andreas Baumann, Jannik Wiese, Tommaso Martorella, Mahdi M. Kalayeh, and Björn Ommer. Envisioning the future, one step at a time. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6823–6836, 2026.

Aritra Bhowmik, Denis Korzhenkov, Cees G. M. Snoek, Amirhossein Habibian, and Mohsen Ghafoorian. MoAlign: Motion-centric representation alignment for video diffusion models. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=OR0ySm4l9h.

Yuanhao Cai, Kunpeng Li, Menglin Jia, Jialiang Wang, Junzhe Sun, Feng Liang, Weifeng Chen, Felix Juefei-Xu, Chu Wang, Ali Thabet, Xiaoliang Dai, Xuan Ju, Alan Yuille, and Ji Hou. PhyGDPO: Physics-aware groupwise direct preference optimization for physically consistent text-to-video generation. In *European Conference on Computer Vision*, pp. 91–109. Springer, 2026.

Guhong Chen, Yingcheng Shi, Yongbin Li, Binhua Li, Xander Xu, Hu Wei, Shiwen Ni, Min Yang, and Jieping Ye. Evotrainer: Co-evolving llm policies and training harnesses for autonomous agentic reinforcement learning. *arXiv preprint arXiv:2606.03108*, 2026.

Harold Haodong Chen, Haojian Huang, Qifeng Chen, Harry Yang, and Ser-Nam Lim. Hierarchical fine-grained preference optimization for physically plausible video generation. *Advances in Neural Information Processing Systems*, 38:148434–148466, 2025a.

Zhifei Chen, Tianshuo Xu, Leyi Wu, Luozhou Wang, Dongyu Yan, Zihan You, Wenting Luo, Guo Zhang, and Yingcong Chen. Stance: Motion coherent video generation via sparse-to-dense anchored encoding. *arXiv preprint arXiv:2510.14588*, 2025b.

Kexu Cheng, Zicheng Liu, Mingju Gao, Chunhe Song, and Hao Tang. PhysRAG: Enhancing physics-awareness in video generation via retrieval-augmented generation. In *European Conference on Computer Vision*, pp. 558–578. Springer, 2026.

Hyung Won Chung, Noah Constant, Xavier Garcia, Adam Roberts, Yi Tay, Sharan Narang, and Orhan Firat. UniMax: Fairer and more effective language sampling for large-scale multilingual pretraining. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=kXwdL1cWOAi.

Tianqing Fang, Hongming Zhang, Zhisong Zhang, Kaixin Ma, Wenhao Yu, Haitao Mi, and Dong Yu. WebEvolver: Enhancing web agent self-improvement with co-evolving world model. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 8959–8975, 2025.

Yuxiang Feng, Juncheng Wang, Chao Xu, Yijie Qian, Huihan Wang, Wenlong Hou, Yang Liu, Baigui Sun, Yong Liu, and Shujun Wang. Newton: Agentic planning for physically grounded video generation. *arXiv preprint arXiv:2605.18396*, 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.

Hongliang He, Wenlin Yao, Kaixin Ma, Wenhao Yu, Hongming Zhang, Tianqing Fang, Zhenzhong Lan, and Dong Yu. OpenWebVoyager: Building multimodal web agents via iterative real-world exploration, feedback and optimization. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 27545–27564, 2025.

Minda Hu, Tianqing Fang, Jianshu Zhang, Jun-Yu Ma, Zhisong Zhang, Jingyan Zhou, Hongming Zhang, Haitao Mi, Dong Yu, and Irwin King. WebCoT: Enhancing web agent reasoning by reconstructing chain-of-thought in reflection, branching, and rollback. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 5155–5173, 2025.

Yihao Hu, Zhihao Wen, Xiujin Liu, Pan Wang, Xin Zhang, and Wei Wu. Seal: Synergistic co-evolution of agents and learning environments. *arXiv preprint arXiv:2605.24426*, 2026.

Chengsong Huang, Wenhao Yu, Xiaoyang Wang, Hongming Zhang, Zongxia Li, Ruosen Li, Jiaxin Huang, Haitao Mi, and Dong Yu. R-Zero: Self-evolving reasoning LLM from zero data. In *International Conference on Learning Representations*, 2026a. URL https://openreview.net/forum?id=96apU6YzSO.

Ziqi Huang, Ning Yu, Gordon Chen, Haonan Qiu, Paul Debevec, and Ziwei Liu. VChain: Chain-of-visual-thought for reasoning in video generation. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 226–250, 2026b.

Kandinsky Lab. Kandinsky wm 1.0: A family of models for physical ai. `https://github.com/kandinskylab/kandinsky-wm`, 2026. Image-to-video models for autonomous driving, robotics, and general physics.

Manjin Kim, Suha Kwak, and Minsu Cho. Tempered self-similarity alignment for physically plausible video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 5148–5158, 2026a.

Subin Kim, Sangwoo Mo, Mamshad Nayeem Rizve, Yiran Xu, Difan Liu, Jinwoo Shin, and Tobias Hinz. Rethinking prompt design for inference-time scaling in text-to-visual generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22090–22099, 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 *International Conference on Machine Learning*, pp. 35685–35709. PMLR, 2025.

Haodong Li, Tianfei Ren, Xiaoxiao Ma, Chunmei Qing, Zhen Fang, Sipeng He, Ziyu Guo, Haoyu Wu, Juanxi Tian, Yihang Zou, Ruichuan An, Dongzhi Jiang, Boxue Yang, Ji Xie, Xu Huang, Wenhao Yan, Jialv Zou, Zhengrong Yue, Yaxin Luo, Xiaotong Li, Yuzhu Wang, Junyan Ye, Jinjing Zhao, Zehui Chen, Lin Chen, Renye Yan, Feng Zhao, and Pheng-Ann Heng. Videococo: Code-as-cot for physically-consistent video generation via an agentic dual-engine system. *arXiv preprint arXiv:2607.27380*, 2026a.

Jiaxing Li, Jiepeng Wang, Junyao Gao, Yang Liu, Eric Li, Bo An, and Hao-Xiang Guo. DynamicsBoost: Dynamic plausible video generation via annotation-free continuation preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20024–20033, 2026b.

Yanxun Li, Hao Wen, Bingze Song, Jiashu Zhu, Aiming Hao, Chubin Chen, Jintao Chen, Jiahong Wu, Xiangxiang Chu, and Miao Wang. Learning explicit physical parameter control and benchmarking for video generation. *arXiv preprint arXiv:2607.18924*, 2026c.

Juyi Lin, Arash Akbari, Yumei He, Lin Zhao, Haichao Zhang, Arman Akbari, Xingchen Xu, Zoe Y. Lu, Enfu Nan, Hokin Deng, Edmund Yeh, Sarah Ostadabbas, Yun Fu, Jennifer Dy, Pu Zhao, and Yanzhi Wang. Phyground: Benchmarking physical reasoning in generative world models. *arXiv preprint arXiv:2605.10806*, 2026a.

Shubo Lin, Xuanyang Zhang, Wei Cheng, Weiming Hu, Gang Yu, and Jin Gao. Mmphysvideo: Physically plausible video generation through joint rgb-perception modeling. *arXiv preprint arXiv:2604.02817*, 2026b.

Chen Ling, Pei Chen, Albert Guan, Jiaming Qu, Shayan Ali Akbar, Madhu Gopinathan, and Erwin Cornejo. Pace: Two-timescale self-evolution for small language model agents. *arXiv preprint arXiv:2605.23019*, 2026.

Yang Liu, Xilin Zhao, Peisong Wen, Siran Dai, and Qingming Huang. Bootstrapping physics-grounded video generation through vlm-guided iterative self-refinement. *arXiv preprint arXiv:2511.20280*, 2025.

Xueyu Luan and Chenwei Shi. Ph-dreamer: A physics-driven world model via port-hamiltonian generative dynamics. *arXiv preprint arXiv:2605.18303*, 2026.

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

Siwei Meng, Yawei Luo, Shu Zhang, and Ping Liu. When physical preferences meet semantic constraints: Physical and semantic direct preference optimization for text-to-video generation. *arXiv preprint arXiv:2607.16947*, 2026. Accepted to ACM Multimedia 2026.

Abolfazl Meyarian, Amin Karimi Monsefi, Rajiv Ramnath, and Ser-Nam Lim. Direct: Disentangled regularization of contrastive trajectories for physics-refined video generation. *arXiv preprint arXiv:2603.25931*, 2026.

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

Shikhar Murty, Christopher D. Manning, Peter Shaw, Mandar Joshi, and Kenton Lee. BAGEL: Bootstrapping agents by guiding exploration with language. In *International Conference on Machine Learning*, pp. 36894–36910. PMLR, 2024.

NVIDIA, Aditi, Niket Agarwal, Arslan Ali, Jon Allen, Martin Antolini, Adeline Aubame, Alisson Azzolini, Junjie Bai, Maciej Bala, Yogesh Balaji, Josh Bapst, Aarti Basant, Mukesh Beladiya, Mohammad Qazim Bhat, Zaid Pervaiz Bhat, Dan Blick, Vanni Brighella, Han Cai, Tiffany Cai, Eric Cameracci, Jiaxin Cao, Yulong Cao, Mark Carlson, Carlos Casanova, Ting-Yun Chang, Yan Chang, Yu-Wei Chao, Prithvijit Chattopadhyay, Roshan Chaudhari, Chieh-Yun Chen, Junyu Chen, Ke Chen, Qizhi Chen, Wenkai Chen, Xiaotong Chen, Yu Chen, An-Chieh Cheng, Click Cheng, Xiu Chia, Jeana Choi, Chaeyeon Chung, Wenyan Cong, Yin Cui, Magdalena Dadela, Nalin Dadhich, Wenliang Dai, Joyjit Daw, Alperen Degirmenci, Rodrigo Vieira Del Monte, Robert Denomme, Sameer Dharur, Marco Di Lucca, Ke Ding, Wenhao Ding, Yifan Ding, Yuzhu Dong, Nicole Drumheller, Yilun Du, Aigul Dzhumamuratova, Aleksandr Efitorov, Hamid Eghbalzadeh, Naomi Eigbe, Imad El Hanafi, Hassan Eslami, Benedikt Falk, Jiaojiao Fan, Jim Fan, Amol Fasale, Sergiy Fefilatyev, Liang Feng, Francesco Ferroni, Sanja Fidler, Xiao Fu, Vikram Fugro, Prashant Gaikwad, TJ Galda, Katelyn Gao, Yihuai Gao, Wenhang Ge, Sreyan Ghosh, Arushi Goel, Vivek Goel, Akash Gokul, Rama Govindaraju, Jinwei Gu, Miguel Guerrero, Elfie Guo, Aryaman Gupta, Siddharth Gururani, Hugo Hadfield, Song Han, Ankur Handa, Zekun Hao, Mohammad Harrim, Ali Hassani, Nathan Hayes-Roth, Yufan He, Chris Helvig, Cyrus Hogg, Madison Huang, Michael Huang, Sophia Huang, Yufan Huang, Jacob Huffman, DeLesley Hutchins, Suneel Indupuru, Boris Ivanovic, Arihant Jain, Joel Jang, Ryan Ji, Yanan Jian, Dongfu Jiang, Jingyi Jin, Atharva Joshi, Nikhilesh Joshi, Pranjali Joshi, Andy Ju, Jaehun Jung, Weiwei Kang, Scott Kassekert, Jan Kautz, Ashna Khetan, Julia Kiczka, Slawek Kierat, Gwanghyun Kim, Kuno Kim, Sunny Kim, Kezhi Kong, Xin Kong, Zhifeng Kong, Tomasz Kornuta, Egor Krivov, Hui Kuang, Saurav Kumar, Chia-Wen Kuo, George Kurian, Wojciech Kutak, JF Lafleche, Himangshu Lahkar, Omar Laymoun, Jayjun Lee, Sanggil Lee, Gabriele Leone, Boyi Li, Freya Li, Jiajun Li, Jinfeng Li, Ling Li, Pengcheng Li, Shangru Li, Tingle Li, Xiaolong Li, Xuan Li, Zhaoshuo Li, Zhiqi Li, Hao Liang, Maosheng Liao, Chen-Hsuan Lin, Tsung-Yi Lin, Ming-Yu Liu, Sifei Liu, Zihan Liu, Hai Loc Lu, Xiangyu Lu, Alice Luo, Ruipu Luo, Wenjie Luo, Jiangran Lyu, Martin Ding Ma, Nic Ma, Qianli Ma, Dawid Majchrowski, Louis Marcoux, Miguel Martin, Qing Miao, Ashkan Mirzaei, Shreyas Misra, Kaichun Mo, Durra Mohsin, Hyejin Moon, Pawel Morkisz, Saeid Motiian, Kirill Motkov, Seungjun Nah, Yashraj Narang, Deepak Narayanan, Thabang Ngazimbi, Julian Ouyang, Shubham Pachori, David Page, Yatian Pang, Sehwi Park, Mahesh Patekar, Mostofa Patwary, Marco Pavone, Trung Pham, Wei Ping, Soha Pouya, Shrimai Prabhumoye, Varun Praveen, Delin Qu, Hesam Rabeti, Morteza Ramezanali, Marilyn Reeb, Xuanchi Ren, Kristen Rumley, Wojciech Rymer, Jun Saito, Yeongho Seol, John Shao, Piyush Shekdar, Tianwei Shen, Humphrey Shi, Min Shi, Stella Shi, Kevin Shih, Mohammad Shoeybi, Mateusz Sieniawski, Shuran Song, Alexander Sotelo, Amir Sotoodeh, Sunil Srinivasa, Vignesh Srinivasakumar, Bartosz Stefaniak, Rahul Heinrich Steiger, Shangkun Sun, Jiaxiang Tang, Shitao Tang, Yangyang Tang, Yue Tang, Tolou Tavakkoli, Kayley Ting, Krzysztof Tomala, Wei-Cheng Tseng, Jibin Varghese, Sergei Vasilev, Thomas Volk, Raju Wagwani, Roger Waleffe, Andrew Z. Wang, Boxiang Wang, Haoxiang Wang, Qiao Wang, Shihao Wang, Shijie Wang, Ting-Chun Wang, Yan Wang, Yu Wang, Rohit Watve, David Wehr, Fangyin Wei, Xinshuo Weng, Jay Zhangjie Wu, Kedi Wu, Hongchi Xia, Summer Xiao, Tianjun Xiao, Kevin Xie, Daguang Xu, Jiashu Xu, Mengyao Xu, Ruqing Xu, Xingqian Xu, Yao Xu, Dinghao Yang, Dong Yang, Hans Yang, Xiaodong Yang, Xuning Yang, Yichu Yang, Yurong You, Zhiding Yu, Hao Yuan, Simon Yuen, Xiaohui Zeng, Pengcuo Zeren, Cindy Zha, Haotian Zhang, Jenny Zhang, Jing Zhang,

Liangkai Zhang, Paris Zhang, Shun Zhang, Xuanmeng Zhang, Zhizheng Zhang, Ann Zhao, Yilin Zhao, Yuliya Zhautouskaya, Charles Zhou, Fengzhe Zhou, Shilin Zhu, Yuke Zhu, Dima Zhylko, and Artur Zolkowski. Cosmos 3: Omnimodal world models for physical ai. *arXiv preprint arXiv:2606.02800*, 2026.

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

Elia Peruzzo, Dejia Xu, Xingqian Xu, Humphrey Shi, and Nicu Sebe. RagMe: Retrieval augmented video generation for enhanced motion realism. In *Proceedings of the 2025 International Conference on Multimedia Retrieval*, pp. 1081–1090, 2025.

Pranav Putta, Edmund Mills, Naman Garg, Sumeet Motwani, Chelsea Finn, Divyansh Garg, and Rafael Rafailov. Agent q: Advanced reasoning and learning for autonomous ai agents. *arXiv preprint arXiv:2408.07199*, 2024.

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

Siddarth Nilol Kundur Satish, Devesh Jaiswal, Hongyu Chen, and Abhishek Bakshi. Physvideogenerator: Towards physically aware video generation via latent physics guidance. *arXiv preprint arXiv:2601.03665*, 2026.

Shuyao Shang, Yuqi Wang, Ruopeng Gao, Xu Chen, Tieniu Tan, Lue Fan, and Zhaoxiang Zhang. Phizero: A world model built around physical language. *arXiv preprint arXiv:2607.28624*, 2026.

Ying Shen, Jerry Xiong, Tianjiao Yu, and Ismini Lourentzou. PHANTOM: Physics-infused video generation via joint modeling of visual and latent physical dynamics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11185–11194, 2026.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36, 2023.

Selena Song, Ziming Xu, Zijun Zhang, Kun Zhou, Jiaxian Guo, Lianhui Qin, and Biwei Huang. Learning plug-and-play memory for guiding video diffusion models. *arXiv preprint arXiv:2511.19229*, 2025.

Nick Stracke, Kolja Bauer, Stefan Andreas Baumann, Miguel Ángel Bautista, Josh Susskind, and Björn Ommer. Learning long-term motion embeddings for efficient kinematics generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 42581–42591, 2026.

Mirac Suzgun, Mert Yuksekgonul, Federico Bianchi, Dan Jurafsky, and James Zou. Dynamic cheatsheet: Test-time learning with adaptive memory. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7080–7106, 2026.

Chujun Tang, Lei Zhong, and Fangqiang Ding. Seeking physics in diffusion noise. *arXiv preprint arXiv:2603.14294*, 2026.

Tencent Hunyuan Foundation Model Team. Hunyuanvideo 1.5 technical report. *arXiv preprint arXiv:2511.18870*, 2025.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Jing Wang, Ao Ma, Ke Cao, Jun Zheng, Zhanjie Zhang, Jiasong Feng, Shanyuan Liu, Yuhang Ma, Bo Cheng, Dawei Leng, Yuhui Yin, and Xiaodan Liang. Wisa: World simulator assistant for physics-aware text-to-video generation, 2025a. URL https://arxiv.org/abs/2502.08153.

Peiyao Wang, Weining Wang, and Qi Li. Physcorr: Dual-reward dpo for physics-constrained text-to-video generation with automated preference selection. *arXiv preprint arXiv:2511.03997*, 2025b.

Yifan Wang, Yanyu Li, Gordon Guocheng Qian, Sergey Tulyakov, Yun Fu, and Anil Kag. Diffusion-drf: Free, rich, and differentiable reward for video diffusion fine-tuning. *arXiv preprint arXiv:2601.04153*, 2026a.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13484–13508, 2023.

Zihan Wang, Kangrui Wang, Qineng Wang, Pingyue Zhang, Linjie Li, Zhengyuan Yang, Xing Jin, Kefan Yu, Minh Nhat Nguyen, Licheng Liu, Eli Gottlieb, Yiping Lu, Kyunghyun Cho, Jiajun Wu, Li Fei-Fei, Lijuan Wang, Yejin Choi, and Manling Li. Ragen: Understanding self-evolution in llm agents via multi-turn reinforcement learning. *arXiv preprint arXiv:2504.20073*, 2025c.

Zijun Wang, Panwen Hu, Jing Wang, Terry Jingchen Zhang, Yuhao Cheng, Long Chen, Yiqiang Yan, Zutao Jiang, Hanhui Li, and Xiaodan Liang. ProPhy: Progressive physical alignment for dynamic world simulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14492–14501, 2026b.

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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 38122–38131, 2026c.

Shang Wu, Chenwei Xu, Zhuofan Xia, Weijian Li, Lie Lu, Pranav Maneriker, Fan Du, Manling Li, and Han Liu. Phyprompt: Rl-based prompt refinement for physically plausible text-to-video generation. *arXiv preprint arXiv:2603.03505*, 2026.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, Qingwei Lin, and Daxin Jiang. WizardLM: Empowering large pre-trained language models to follow complex instructions. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=CfXh93NDgH.

Tianshuo Xu, Zhifei Chen, Leyi Wu, Hao Lu, and Ying-cong Chen. Motion forcing: A decoupled framework for robust video generation in motion dynamics. *arXiv preprint arXiv:2603.10408*, 2026.

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)*, pp. 18826–18836, 2025.

Zherui Yang, Fan Liu, Yansong Ning, and Hao Liu. EvoDS: Self-evolving autonomous data science agent with skill learning and context management. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, pp. 6128–6139, 2026.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. CogVideoX: Text-to-video diffusion models with an expert transformer. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=LQzN6TRFg9.

Xunjian Yin, Xinyi Wang, Liangming Pan, Li Lin, Xiaojun Wan, and William Yang Wang. Gödel agent: A self-referential agent framework for recursively self-improvement. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 27890–27913, 2025.

Jianhao Yuan, Xiaofeng Zhang, Felix Friedrich, Nicolas Beltran-Velez, Melissa Hall, Reyhane Askari-Hemmat, Xiaochuang Han, Nicolas Ballas, Michal Drozdzal, and Adriana Romero-Soriano. Inference-time physics alignment of video generative models with latent world models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16118–16129, 2026a.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason E. Weston. Self-rewarding language models. In *International Conference on Machine Learning*, pp. 57905–57923. PMLR, 2024.

Yu Yuan, Xijun Wang, Tharindu Wickremasinghe, Zeeshan Nadir, Bole Ma, and Stanley H. Chan. NewtonGen: Physics-consistent and controllable text-to-video generation via neural Newtonian dynamics. In *International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=rJ6N6sunaU.

Mert Yuksekgonul, Federico Bianchi, Joseph Boen, Sheng Liu, Zhi Huang, Carlos Guestrin, and James Zou. Textgrad: Automatic "differentiation" via text. *arXiv preprint arXiv:2406.07496*, 2024.

Yunpeng Zhai, Shuchang Tao, Cheng Chen, Anni Zou, Ziqian Chen, Qingxu Fu, Shinji Mai, Li Yu, Jiaji Deng, Zouying Cao, Zhaoyang Liu, Bolin Ding, and Jingren Zhou. Agentevolver: Towards efficient self-evolving agent system. *arXiv preprint arXiv:2511.10395*, 2025.

Guofeng Zhang, Angtian Wang, Jacob Zhiyuan Fang, Liming Jiang, Haotian Yang, Bo Liu, Yiding Yang, Guang Chen, Longyin Wen, Alan Yuille, and Chongyang Ma. TGT: Text-grounded trajectories for locally controlled video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22028–22037, 2026a.

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

Qifan Zhang, Dongyang Ma, Tianqing Fang, Jia Li, Jing Tang, Nuo Chen, Haitao Mi, and Yan Wang. Training llm agents for spontaneous, reward-free self-evolution via world knowledge exploration. *arXiv preprint arXiv:2604.18131*, 2026b.

Qizheng Zhang, Changran Hu, Shubhangi Upasani, Boyuan Ma, Fenglu Hong, Vamsidhar Kamanuru, Jay Rainton, Chen Wu, Mengmeng Ji, Hanchen Li, Urmish Thakker, James Zou, and Kunle Olukotun. Agentic context engineering: Evolving contexts for self-improving language models. In *International Conference on Learning Representations*, 2026c. URL https://openreview.net/forum?id=eC4ygDs02R.

Pu Zhao, Juyi Lin, Timothy Rupprecht, Arash Akbari, Chence Yang, Rahul Chowdhury, Elaheh Motamedi, Arman Akbari, Yumei He, Chen Wang, Geng Yuan, Weiwei Chen, and Yanzhi Wang. Phyworld: Physics-faithful world model for video generation. *arXiv preprint arXiv:2605.19242*, 2026a.

Zengqun Zhao, Ziquan Liu, Yu Cao, Shaogang Gong, Zhensong Zhang, Jifei Song, Jiankang Deng, and Ioannis Patras. LatSearch: Latent reward-guided search for faster inference-time scaling in video diffusion. In *European Conference on Computer Vision*, pp. 170–189. Springer, 2026b.

Fengzhe Zhou, Jiannan Huang, Jialuo Li, Deva Ramanan, and Humphrey Shi. PAI-Bench: A comprehensive benchmark for physical AI. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21522–21536, 2026a.

Siyuan Zhou, Hejun Wang, Hu Cheng, Jinxi Li, Dongsheng Wang, Junwei Jiang, Yixiao Jin, Jiayue Huang, Shiwei Mao, Shangjia Liu, Yafei Yang, Hongkang Song, Shenxing Wei, Zihui Zhang, Bing Wang, Zhihua Wang, Chuhang Zou, and Bo Yang. PhysInOne: Visual physics learning and reasoning in one suite. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 33131–33142, 2026b.

Chenhui Zhu, Yilu Wu, Shuai Wang, Gangshan Wu, and Limin Wang. MotionRAG: Motion retrieval-augmented image-to-video generation. *Advances in Neural Information Processing Systems*, 38:192087–192110, 2025.

Jingyuan Zhu, Biaolong Chen, Le Zhang, Aixi Zhang, Hao Jiang, and Pipei Huang. Diffusion-apo: Trajectory-aware direct preference alignment for video diffusion transformers. *arXiv preprint arXiv:2605.07503*, 2026.

Sihan Zhuang, Xinyuan Chen, Tianfan Xue, and Yaohui Wang. Causalmotion: Structured physical reasoning as keyframe and trajectory guidance for training-free video generation. *arXiv preprint arXiv:2606.14317*, 2026.

APPENDIX

A PROMPT SELF-EVOLUTION ANALYSIS

As mentioned in Sec. 5.4.1, to better understand how the captioning prompt evolves during iterations, here we visualize the F1-score curve on **PhysCapBench** in Fig. 6. The annotations summarize five observed representative modifications to the prompt design.

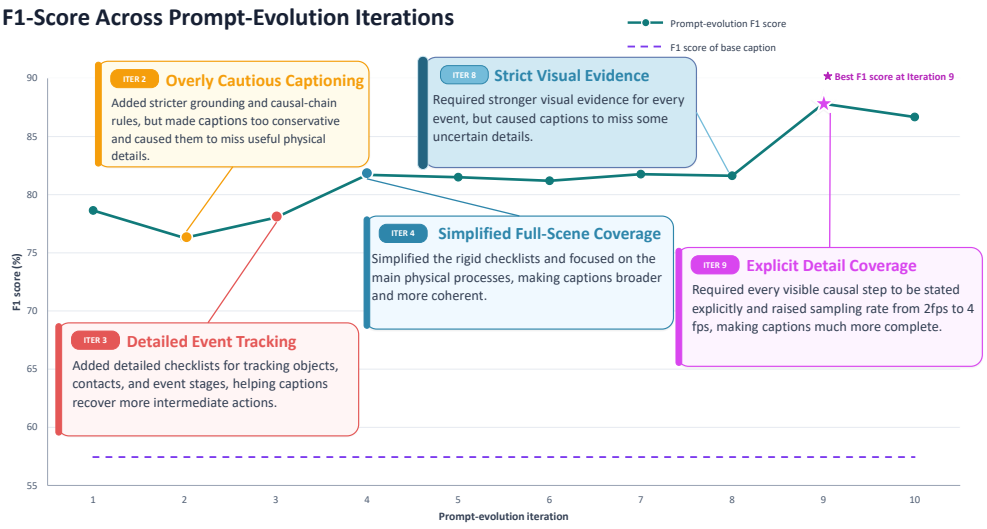


Figure 6: **F1-score trajectory across prompt self-evolution iterations on PhysCapBench**. The dashed purple line denotes the F1 score of base caption. The five callouts summarize the representative prompt modifications introduced at Iterations 2, 3, 4, 8 and 9. The star identifies the best-performing iteration. The annotations describe the prompt-design changes, whereas all plotted F1 scores are obtained from PhysCapBench.

B Additional Implementation Details

B.1 PhysCapBench and Evaluation Protocol

As illustrated in Sec. 4.2, our PhysCapBench contains 246 videos and includes 3,794 human-verified physical-dynamics assertions with a mean of 15.42 assertions per video. All captions are evaluated by Gemini-3-Flash, using temperature 0.1, a 65,536-token JSON-output limit, and batches of at most 30 claims. Precision decomposes the caption into atomic claims, verifies each against the video, and computes $N_{\text{correct}}/(N_{\text{correct}} + N_{\text{incorrect}})$. Recall denotes the proportion of the 3,794 reference assertions supported by the generated caption.

B.2 More Details on the Agentic Loop

Here we elaborate more implementation details for our agentic loop described in Sec. 3.1. We optimized the physics captions through a generate–critic–revise loop on a fixed development set of 20 videos, which is randomly sampled from our training set, with 273 human-verified atomic physics assertions in total. At each iteration, a fixed GPT-5.5 captioner processed temporally ordered frames and jointly generated *physics_reasoning* and *physics_negative_prompt* in a single call, while only *physics_reasoning* was scored. We use Gemini-3.1-Pro to measure assertion-level recall and video-grounded claim precision, and the evolution agent used both aggregate scores and claim-level failure rationales to revise the prompt.

To construct the upsampling prompt, we preserved the physics-focused requirements of the final caption prompt while adapting its input interface. Instead of describing the visual content from scratch, the upsampler expands the original short benchmark caption, together with visual inputs when required, into the same structured *physics_reasoning* and *physics_negative_prompt* fields.

B.3 Long Context Support on Wan

In Sec. 5.1, one of the backbones going through our fine-tuning is Wan2.1 (Wan et al., 2025). The text encoder UMT5-XXL (Chung et al., 2023) of Wan2.1 imposes a maximum context length of 512 tokens, which can be restrictive for our physics-aware prompts that contain both detailed descriptions and reasoning. To accommodate arbitrarily long captions without modifying the pretrained text encoder, we adopt a sliding-window strategy. Specifically, an over-length prompt is partitioned into overlapping windows, each within the 512-token limit, and the resulting text representations are aggregated to provide conditioning for the video generation model. In practice, the overlap between two consecutive context windows is set to 192 tokens, and we take the average of the embeddings from the two windows over the overlapping region, as shown in Fig. 7.

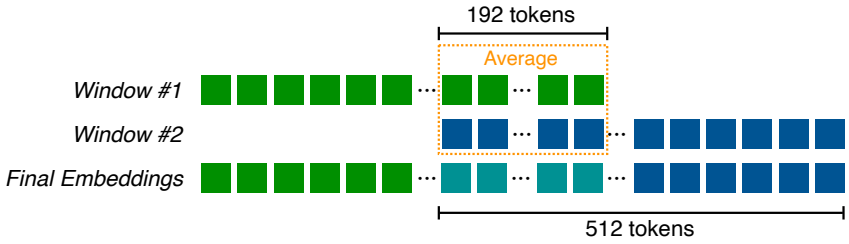


Figure 7: **Sliding-window mechanism for text encoding in Wan2.1.**

B.4 More Details on Physics Video Generation Benchmarks

In Sec. 5, we evaluate models on four benchmarks: PhyGenBench and VideoPhy-2 for Text-to-Video generation, and Physics-IQ Verified and PhyGround for Image-to-Video generation. In this section, we briefly describe the evaluation protocol and the physical capabilities assessed by each benchmark.

PhyGenBench. The 160 prompts of PhyGenBench can be divided into four categories of physical phenomena: *force*, *light*, *heat*, and *material*. We calculate the video-level score with the official

scripts, but replacing the evaluator with GPT-5.5. Each video is graded on a 0–3 scale, where 0 indicates a complete fantastical physical behavior and 3 indicates strong physical realism. The final score is obtained by averaging across all 160 samples.

VideoPhy-2. VideoPhy-2 evaluates generated videos from two aspects: *semantic consistency* (SA) and *physical correctness* (PC), each is graded on a 0–5 scale. Its test set contains 591 samples covering common physical phenomena in everyday scenarios, with 178 challenging samples selected to form a Hard subset. We replace its offline VLM evaluator with GPT-5.5 (App. G). The joint accuracy measures the fraction of videos that simultaneously satisfy both SA ≥ 4 and PC ≥ 4.

Physics-IQ Verified. Physics-IQ (verified version) evaluates generated videos across multiple physical concepts using four complementary metrics: *spatial consistency* (S), *spatial-temporal consistency* (ST), *weighted-spatial consistency* (WS), and *mean squared error* (MSE). It contains 66 physical scenarios, with three perspectives evaluated per scenario (198 samples in total). It provides ground-truth videos for reference. The final score is computed by aggregating the scores across these metrics and averaging over all evaluated samples.

PhyGround. PhyGround evaluates each generated video using *semantic adherence* (General) and *physical temporal validity* (Physics). Each criterion is scored on a 1–5 scale, and averaged across all 250 samples.

B.5 Training Details

Here we offer training details for Cosmos3, Wan2.1 and our PhysThinker in Sec. 5.

B.5.1 Cosmos 3 family

Cosmos3-Nano. We fine-tuned Cosmos3-Nano using the dataset of about 183K video–caption pairs (Sec. 5.1), including 71K WISA-80K samples retained after filtering for motion and aesthetic quality and removing multi-shot or metadata-incomplete videos, plus 112K videos selected through our deficiency-guided retrieval data pipeline. Each video was paired with a caption combining the base visual description and the generated physics-reasoning annotation. Videos were processed at a spatial resolution of (256×256), with a maximum caption length of 4,096 tokens and a maximum packed sequence length of 45,056 tokens. During training, the conditioning format was sampled as 70% text-to-video, 20% image-to-video using the first frame, and 10% video-to-video using the first five frames; classifier-free guidance dropout was set to 0.1. We applied LoRA only to the query, key, value, and output projections of the attention layers, using rank 16 and scaling factor 32. Training used bfloat16 precision, fully sharded data parallelism, and full activation checkpointing across 16 GPUs on two nodes. Each GPU processed one packed sequence at a time, with four gradient-accumulation steps, giving a nominal global batch size of 64 packed sequences per optimizer update. We optimized the model with fused AdamW using a learning rate of (5×10^{-4}), $(\beta_1 = 0.9)$, $(\beta_2 = 0.95)$, $(\epsilon = 10^{-6})$, no weight decay, and gradient clipping at 0.1. The checkpoint at iteration 1,340 was used for the reported benchmark evaluation.

Cosmos3-Super. We fine-tuned Cosmos3-Super on the same training set and caption format described above. Videos were resized and center-cropped using aspect-ratio-dependent buckets at the 256-resolution tier (e.g., 256×256 for square videos and 320×192 for landscape 16:9 videos), with a maximum caption length of 4,096 tokens and a maximum packed sequence length of 45,056 tokens. The conditioning format was sampled as 70% text-to-video, 20% image-to-video using the first frame, and 10% video-to-video using the first two latent frames, corresponding to five video frames under temporal compression by a factor of four. Classifier-free guidance dropout was set to 0.1. We applied LoRA only to the query, key, value, and output projections of the generation expert's attention layers, using rank 16 and scaling factor 32. Training used bfloat16 precision, fully sharded data parallelism, and full activation checkpointing across 16 GPUs on four nodes, with context-parallel degree 4 and data-parallel degree 4. Each data-parallel replica processed one packed sequence per microstep, distributed across its four context-parallel ranks. With two gradient-accumulation steps, this yielded a nominal global batch size of eight packed sequences per optimizer update. We used fused AdamW with a learning rate of 5×10^{-4}, $\beta_1 = 0.9$, $\beta_2 = 0.95$, $\epsilon = 10^{-6}$, zero weight decay, and gradient clipping at 0.1. The checkpoint at iteration 1000 was used for evaluation.

Cosmos3-Edge. We fine-tuned Cosmos3-Edge from its pretrained initialization on the same training set and caption format described above. Videos were processed using aspect-ratio-dependent buckets at the 480-resolution tier (e.g., 832×480 for landscape 16:9 videos), with at most 121 frames per video and a maximum caption length of 8,192 tokens. Training used sample-count-based packing with at most two samples per packed batch and a model token capacity of 90,112. The conditioning format was sampled as 70% text-to-video and 30% image-to-video using the first frame; video-to-video conditioning was not used. Classifier-free guidance dropout was set to 0.1. We applied LoRA only to the query, key, value, and output projections of the generation expert's attention layers, using rank 16 and scaling factor 32. The main training run used bfloat16 precision, fully sharded data parallelism, and full activation checkpointing across 32 GPUs on eight nodes, with context-parallel degree 1, data-parallel degree 32, and no gradient accumulation. This corresponds to a nominal global batch size of 64 samples per optimizer update when every packed batch contains two samples. We used fused AdamW with a learning rate of 5×10^{-4}, $\beta_1 = 0.9$, $\beta_2 = 0.95$, $\epsilon = 10^{-6}$, zero weight decay, and gradient clipping at 0.1. Checkpoints were evaluated at 500-update intervals; the checkpoint at iteration 2,000 was used for the reported benchmark evaluation.

B.5.2 WAN2.1

We fine-tuned both Wan2.1-T2V-14B and Wan2.1-I2V-14B using the identical dataset described in B.5.1. Each caption contains both the base visual description and the generated physics-reasoning annotation. During training, classifier-free guidance dropout was set to 0.1. We applied LoRA only to the query, key, value, and output projections of the generation expert's attention layers, using rank 256 and scaling factor 512, while the patch embedder and output modules are fully fine-tuned. Training used bfloat16 precision, fully sharded data parallelism, and full activation checkpointing across 16 NVIDIA H100 GPUs. The global batch size is set to 32. We optimized the model with fused AdamW using a learning rate of (5×10^{-4}) with 1000 warmup steps, $(\beta_1 = 0.9)$, $(\beta_2 = 0.95)$, $(\epsilon = 10^{-6})$, no weight decay, and gradient clipping at 0.1. The checkpoint at iteration 8,000 was used for the reported benchmark evaluation.

B.5.3 PHYSTHINKER

PhysThinker consists of two Qwen3-VL-4B-Instruct models fine-tuned with our training data, differing only in the conditioning modality. The captioner is trained on videos paired with GPT-generated physics captions, while the upsampler is trained on short text prompts, with or without a conditioning image, paired with GPT-expanded physics-rich prompts. We applied LoRA with rank 8, scaling factor 32, and dropout 0.05 to all linear projections of the language model, while the entire vision stack (ViT and aligner) is fully fine-tuned. Training used bfloat16 precision, FlashAttention-2, DeepSpeed ZeRO-3 with optimizer and parameter offload to host memory, and full activation checkpointing including the vision tower. Examples are packed to a 57,344-token budget, with video sampled at 4 fps up to 64 frames. The global batch size is 256 packed sequences. We optimized with fused AdamW at a learning rate of (1×10^{-4}) for the language-model adapters and (2×10^{-5}) for the vision tower and aligner, using a cosine schedule with 5% linear warmup, $(\beta_1 = 0.9)$, $(\beta_2 = 0.95)$, $(\epsilon = 10^{-8})$, weight decay 0.1, and gradient clipping at 1.0. Both models were trained for 3 epochs.

The video-conditioned reasoner takes a video alone and was trained across 128 A100 GPUs. The checkpoint at step 650 was used for the reported benchmark evaluation. The prompt upsampler expands a short generation prompt, and is trained as a two-mode mixture: text-only, and text together with the clip's first frame. Each corpus contributes to both modes, trained across 32 A100 GPUs. The final checkpoint at step 258 was used for the reported benchmark evaluation.

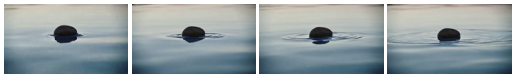
C MORE QUALITATIVE COMPARISONS

As a supplement to Sec. 5.2, here we show the qualitative examples across the evaluated benchmarks.

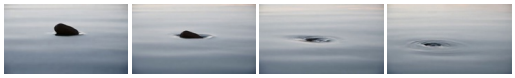
PhyGenBench: stone sinking

Official prompt: A stone is gently placed on the surface of a pool filled with water.

Baseline



Ours



0.00 s 1.42 s 2.83 s 7.83 s

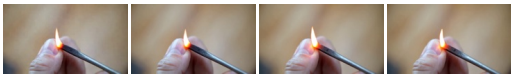
physics_reasoning: "The stone **weight remains larger than the buoyant force** because a typical stone is denser than water. **After the support releases the stone, gravity accelerates the unsupported stone downward** into the pool. The stone **penetrates through the water surface** instead of remaining on top of the surface."

physics_negative_prompt: "The stone **remains floating motionless on top of the pool surface after release**."

PhyGenBench: potassium flame

Official prompt: A piece of potassium is ignited, emitting a vivid and unique flame as it burns steadily.

Baseline



Ours



0.00 s 2.12 s 4.29 s 7.83 s

physics_reasoning: "Ignition **supplies heat** to the exposed potassium surface. ... A vivid **violet-lilac flame forms** at the potassium surface and remains attached to the burning region. The flame color comes from **excited potassium atoms and ions emitting characteristic visible wavelengths as the excited electrons lose energy**."

physics_negative_prompt: "The flame burns as a **steady orange candle-shaped plume with no potassium-colored violet-lilac emission**."

Physics-IQ Verified: domino propagation

Official prompt (excerpt): The platform rotates clockwise and the wooden stick hits the first block as it rotates.

Baseline



Ours



0.00 s 4.29 s 4.67 s 5.00 s

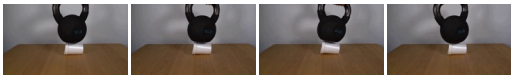
physics_reasoning: "The falling purple block **contacts the right side of the upright blue block**. The purple-blue collision **transfers impulse and angular momentum** from the purple block to the blue block. The blue block **begins rotating leftward about its lower edge** while table friction limits bottom sliding."

physics_negative_prompt: "The colored blocks **topple in separate directions instead of transferring leftward collision impulses along the row**."

Physics-IQ Verified: load-induced cup collapse

Official prompt: A 30lb kettlebell is slowly lowered onto a white styrofoam cup placed on a wooden table on its side. Static shot with no camera movement.

Baseline



Ours



0.00 s 0.92 s 1.83 s 5.00 s

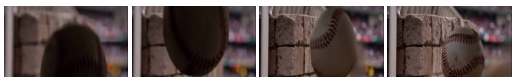
physics_reasoning: "The kettlebell **applies a large downward contact force** to a small region of the styrofoam cup wall. ... The cup material **deforms because the compressive stress exceeds the stiffness and strength** of the thin styrofoam wall. ... The kettlebell center of mass **moves downward as the collapsing cup loses height**."

physics_negative_prompt: "The white cup **collapses downward** while the black kettlebell **remains suspended in place without following the lost support**."

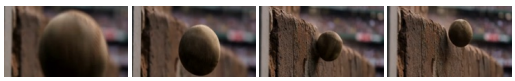
VideoPhy-2: baseball-wall contact

Official prompt: A baseball is thrown with great force; it hits a brick wall and leaves a visible mark.

Baseline



Ours



0.00 s 0.92 s 3.62 s 4.54 s

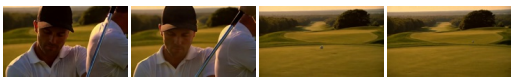
physics_reasoning: "At impact, the baseball **contacts the brick wall** over a small area. The brick wall exerts a **large normal force on the baseball opposite the baseball's incoming velocity**. The baseball exerts an equal and opposite force on the brick surface through **Newton's third law**. The short collision time produces a **large impulse that sharply reduces the baseball's forward momentum**."

physics_negative_prompt: "The baseball **passes through the brick wall without slowing or transferring momentum**."

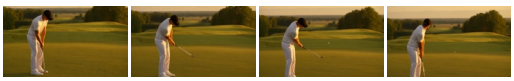
VideoPhy-2: golf chip

Official prompt: A golfer chips the ball onto the green, the ball rolling smoothly towards the hole.

Baseline



Ours



0.00 s 1.04 s 2.29 s 5.00 s

physics_reasoning: "The clubhead **contacts the ball during the swing**. The clubhead applies a **short, strong impulse** to the ball at impact. **Newton's third law** makes the ball apply an equal and opposite impulse back on the clubhead. The impulse **changes the ball from rest to forward and upward motion**."

physics_negative_prompt: "The golf ball **launches into the air before the clubhead contacts it**. The ball flies in a **straight horizontal line without downward curvature from gravity**."

Figure 8: **Qualitative comparisons on different physics benchmarks.** Each case compares frames generated by the Cosmos3-Nano base model using base caption (Baseline) with those generated by our trained model using base caption augmented with our *physics_reasoning* and *physics_negative_prompt* (Ours) at identical timestamps. Official prompts and selected physics reasoning and negative-prompt excerpts are shown, with bold phrases highlighting physical constraints. Negative excerpts describe violations to avoid. All frames are uncropped.

D DEMOS IN DRIVING AND ROBOTICS SCENARIOS

As stated in Sec. 5.2, here we offer 6 demos to illustrate the strong performance of our method in driving and robotics scenarios.



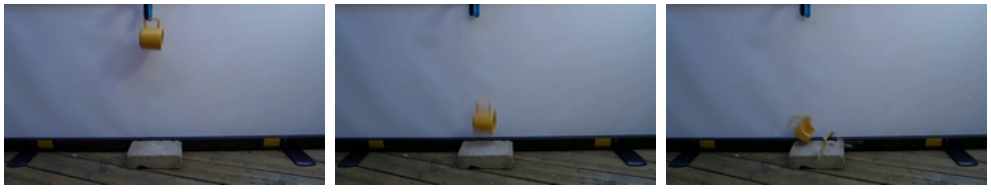
Figure 9: **Qualitative results on driving and robotics.** All sequences are generated by our trained model using base caption augmented with our *physics_reasoning* and *physics_negative_prompt*. Each example includes four uncropped video frames with timestamps, a caption excerpt, and selected physics reasoning and negative-prompt excerpts.

E MORE EXAMPLES OF PHYSCAPBENCH

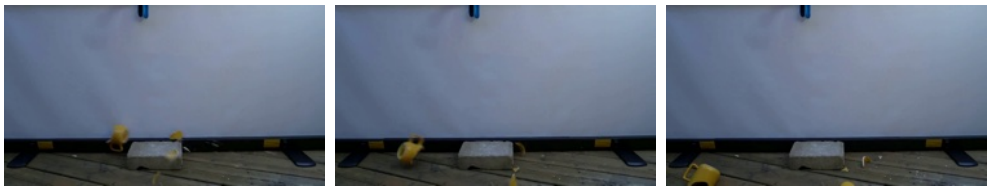
We provide two examples from PhysCapBench (Sec. 4.2): a falling cup that fractures upon impact and a Newton's cradle exhibiting coupled collisions. Each example includes six chronological video frames and its complete set of 20 human-curated annotations, categorized as *Cause*, *Law*, or *Effect*. Together, these examples illustrate how the benchmark captures initiating interactions, governing physical principles, and observable outcomes beyond coarse event descriptions.

Cup drop and brittle fracture

8.00 s | 20 assertions



2.83 s Falling | 3.03 s Before impact | 3.10 s Fracture



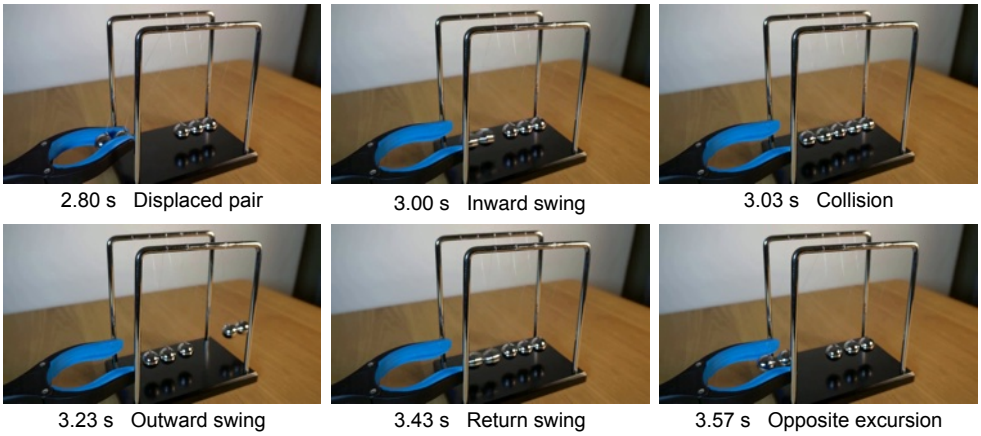
3.20 s Fragment dispersal | 3.40 s Rebound | 4.40 s Debris at rest

ID	Type	Annotation
01	Effect	Cup body ruptures instantaneously at impact site
02	Cause	The impact force from the obstacle triggers cup body rupture
03	Effect	Cup body breaks into fragments of different sizes
04	Effect	Cup body fragments scatter in multiple directions from impact point
05	Cause	Cup collides with central rigid obstacle
06	Law	The impulse from collision changes the momentum of each part of the cup
07	Effect	The main cup body rebounds to the left side of the frame from the impact point
08	Effect	Small fragments fly away to the right of the obstacle from the impact point
09	Law	The obstacle applies a reverse contact force to the cup body
10	Law	Impact energy converts to material fracture energy
11	Cause	Gravity drives the cup downward
12	Effect	Cup falls downward from top of frame toward central obstacle
13	Law	Cup's falling speed increases with time
14	Law	Cup's gravitational potential energy converts to falling kinetic energy
15	Effect	Cup body separates from obstacle after impact
16	Law	Aerial fragments fall along a parabolic path under gravity
17	Effect	After landing the main cup body slides to the left side of the frame
18	Law	The ground applies a supporting force to the debris
19	Law	Ground friction reduces the fragments' sliding speed
20	Effect	Cup body fragments eventually scatter on both sides of obstacle

Figure 10: **Cup drop, impact, and fracture.** Six chronological frames accompany the reference physical assertions in their original wording and rank order. Phase labels summarize the displayed frames; assertion IDs are not temporal alignments.

Newton's cradle: coupled collisions

8.00 s | 20 assertions



2.80 s Displaced pair | 3.00 s Inward swing | 3.03 s Collision

3.23 s Outward swing | 3.43 s Return swing | 3.57 s Opposite excursion

ID	Type	Annotation
01	Cause	The outer steel ball swinging inward collides with an adjacent steel ball
02	Law	The impulse from collision propagates along the steel ball chain
03	Effect	The opposite outer sphere swings outward after collision
04	Effect	Incident sphere significantly decelerates after collision
05	Law	Contact steel ball experiences an equal and opposite force
06	Law	The middle steel ball array transmits most of the incident momentum
07	Effect	The middle steel ball maintains a small overall displacement
08	Law	Tension in suspension line constrains outer ball to follow arc motion
09	Cause	The outer ball displaced from equilibrium possesses gravitational potential energy
10	Cause	Gravity tangential component drives the outer sphere to swing back
11	Effect	The outer ball swinging downward accelerates toward the lowest point
12	Effect	The outer ball reaches a high speed near the lowest point
13	Law	The outer ball swinging upward decelerates with increasing height
14	Effect	The outer ball changes direction at the lateral high point
15	Effect	The outer balls at both ends periodically alternate approaching the steel ball array
16	Law	Dissipation causes the swing amplitude to gradually decrease
17	Effect	The two balls on the far left swing together to the right
18	Effect	The two balls on the far right swing together to the right
19	Effect	The number of balls lifted on each side is identical
20	Law	Momentum is conserved in all ball collisions

Figure 11: **Newton's cradle and momentum transfer.** Six chronological frames accompany the reference physical assertions in their original wording and rank order. Phase labels summarize the displayed frames; assertion IDs are not temporal alignments.

F Analysis on Video Generation Attention Maps

As in Sec. 5.4.1, to measure the effect of *physics_reasoning* derived by self-evolution iterations in Physis-Lang, we use Cosmos3-Nano as the backbone and compare two configurations: inference with base caption only using the pretrained model, and inference with base caption + *physics_reasoning* using the model after SFT on the data we re-captioned with *physics_reasoning* field.

Cosmos3 processes text tokens and video latent tokens in a unified attention sequence. To measure the importance of each token, we compute the attention from sampled visual queries to all text tokens. Specifically, we capture the normalized query and key features after mRoPE and compute the text-conditional attention

$$A(w,q) = \frac{1}{H}\sum_{h=1}^{H} \mathrm{Softmax}_{w'}\left(\frac{Q_h(q)K_h(w')^\top}{\sqrt{d_h}}\right)_w, \tag{1}$$

where w is the given text token, and q is the video-latent query. H denotes the number of attention heads. The score of a token is obtained by averaging $A(w,q)$ over the uniformly sampled visual tokens, denoising steps, and all transformer layers. For better visualization, we divide each word score by the mean score of all word occurrences in the same prompt. Therefore, the larger the scaled score of a text token is, the greater the role that corresponding word plays in video generation.

Similarly, to obtain the focal regions of an RGB frame with respect to a key sentence S, we average the attention assigned to all text tokens within the span of S:

$$M_S(q) = \frac{1}{|\mathcal{T}(S)|}\sum_{w\in\mathcal{T}(S)} A(w,q), \tag{2}$$

where q is a visual latent token encoded from the RGB frame, and $\mathcal{T}(S)$ denotes the corresponding token set. Each sampled visual query retains its latent coordinate (t,y,x), allowing the values $M_S(q)$ to be arranged into a spatial map.

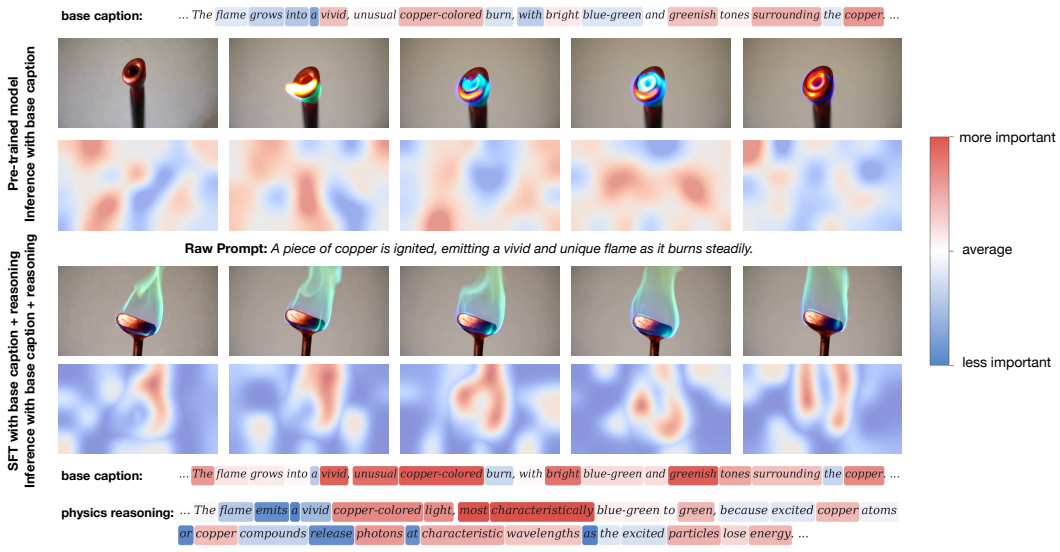


Figure 12: **Visualization for the attention scores of visual and text tokens.** When adding `physics_reasoning` into the SFT and inference process, the generative model can better capture the keywords for the core physical process, and generate videos with physical realism.

As shown in Fig. 12, when incorporating *physics_reasoning* into the SFT and inference process, the model can better locate the keywords which determine the physical process, like "copper-colored" and "greenish" in the base caption, and "photons", "wavelengths" and "energy" in the physics reasoning. For the RGB frames, the region depicting the flame color reaction is also more prominent on the attention map.

Moreover, to examine the important role of our fine-tuning in enabling the model to effectively utilize *physics_reasoning*, we separately feed base caption + *physics_reasoning* into the pretrained model and the fine-tuned model. Tab. 9 lists the relative increase in the attention scores of keywords in *physics_reasoning* after training.

Table 9: **Relative increase of attention scores of physical keywords after fine-tuning.**

Caption	Δ (%)
...Regions directly under fingertips ***COMPRESS*** *first and most strongly. ...*	19.2
... ***HEAT TRANSFER*** *lowers the average molecular kinetic energy of the juice, ...*	52.2
... ***GRAVITY*** *still acts downward on all gas and particle mass, ...*	22.5
...The change in light speed at the lens surfaces ***BENDS*** *the rays. ...*	48.7

G VALIDATION OF GPT-5.5 AS A BETTER EVALUATOR

In Sec. 5.1, we replace the offline VLM evaluators for PhyGenBench and VideoPhy-2 with GPT-5.5. In this section, we illustrate the rationale for this replacement using VideoPhy-2 as an example.

As shown in Fig. 13 and Tab. 10, distinct video generative models often receive nearly identical scores from the pretrained VLM, leading to results that do not faithfully reflect their physical performance. In contrast, GPT-5.5 produces assessments that align with human judgment in most cases.

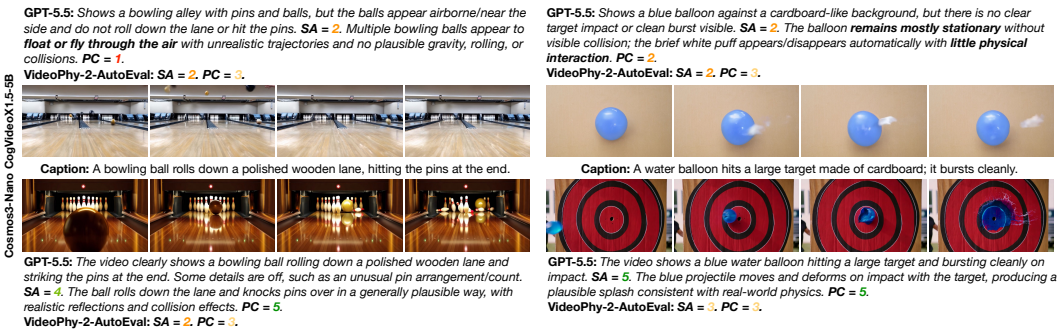


Figure 13: **Different evaluators for videos generated by CogVideoX1.5-5B and Cosmos3-Nano.** While the auto-evaluator of VideoPhy-2 tends to assign similar mid-range scores to videos with substantially different levels of physical realism, GPT-5.5 provides more discriminative assessments, better reflecting the differences in physical realism across videos.

Table 10: **Comparison of VideoPhy-2 scores under different evaluators.**

Generative Model	Pretrained VLM	GPT-5.5
CogVideoX1.5-5B	23.01	49.41
Wan2.1-T2V-14B	23.52	57.02
Cosmos3-Nano	22.50	60.41
Veo-3.1	23.69	68.87

H General Generative Capability of Physis-Lang

In Sec. 5.3, we demonstrate that Physis-Lang explicitly optimizes captions and improves the physical realism of generated videos. A natural question is whether such physics-aware captions could compromise the model's general capability of video generation. To investigate further, we evaluate the pretrained and Physis-Lang-enhanced models on VBench for the I2V task. As shown in Tab. 11, Physis-Lang maintains comparable performance to the pretrained models, suggesting that the richer and more structured captions introduced by Physis-Lang can provide useful information beyond explicitly physical scenarios.

Table 11: **Comparison of different models on VBench.**

Model	VBench-I2V
Wan2.1-14B (Pretrained)	86.86
Wan2.1-14B (Physis-Lang)	87.51
Cosmos3-Nano (Pretrained)	88.32
Cosmos3-Nano (Physis-Lang)	88.69