PHYRPR: TRAINING-FREE PHYSICS-CONSTRAINED VIDEO GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent diffusion-based video generation models can synthesize visually plausible videos, yet they often struggle to satisfy physical constraints. A key reason is that most existing approaches remain single-stage: they entangle high-level physical understanding with low-level visual synthesis, making it hard to generate content that require explicit physical reasoning. To address this limitation, we propose a training-free three-stage pipeline, *PhyRPR*: *PhyReason–PhyPlan–PhyRefine*, which decouples physical understanding from visual synthesis. Specifically, *PhyReason* uses a large multimodal model for physical state reasoning and an image generator for keyframe synthesis; *PhyPlan* deterministically synthesizes a controllable coarse motion scaffold; and *PhyRefine* injects this scaffold into diffusion sampling via a latent fusion strategy to refine appearance while preserving the planned dynamics. This staged design enables explicit physical control during generation. Extensive experiments under physics constraints show that our method consistently improves physical plausibility and motion controllability.

1 INTRODUCTION

Diffusion-based video generation models have made progress, synthesizing high-fidelity, visually compelling content. However, despite their visual realism, video diffusion models remain largely correlation-driven: they exploit patterns in large-scale training data, rather than explicitly enforcing physical constraints. As a result, they often fail in scenarios with clear physical constraints.

In image generation, a similar challenge arises: how to infer the deeper physical implications of a prompt through reasoning. Prior methods address this via prompt enhancement or joint training that couples VLM with diffusion models. While effective for images in some cases, directly extending these strategies to video for physically consistent generation is often insufficient. Prompt enhancement typically targets surface-level appearance and is too imprecise to encode temporal dynamics. Training-based video approaches are extremely expensive, and without dedicated physics-aware annotations, models struggle to acquire generalizable physical commonsense through implicit learning. Consequently, many video generators still rely on a single-stage denoising process where physical reasoning is learned implicitly and entangled with visual synthesis. This entanglement makes generation hard to control when explicit kinematic constraints or grounded interactions are required.

To address this problem, we propose a training-free framework that decouples physical understanding from visual synthesis via a three-stage pipeline *PhyRPR*: *PhyReason–PhyPlan–PhyRefine*. The key idea is to separate *what should happen* from *how it is rendered* by exposing intermediate, physically meaningful representations. In *PhyReason*, we leverage a large multimodal model to infer the underlying physical implications of the prompt and extract a sequence of physically key states. These states are visualized as semantically consistent keyframes, together with object-centric masks that provide explicit handles for subsequent control. In *PhyPlan*, we translate the discrete key states into continuous motion trajectories and synthesize a coarse motion scaffold that explicitly encodes object dynamics and interactions. In *PhyRefine*, we integrate this scaffold into diffusion sampling through motion-aware noise-consistent latent fusion, so that the video model can refine details while staying aligned with the planned dynamics. Overall, this design provides controllable generation process that better satisfies kinematic constraints and physics-grounded interactions.

By combining LMM-based physical priors, deterministic planning tools, and the strong rendering capability of video diffusion models, our method forms a physically interpretable, training-free

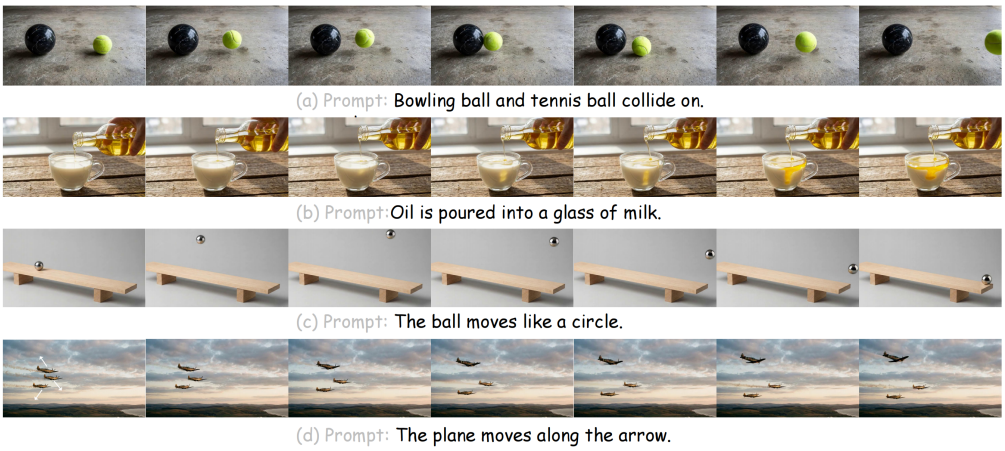

Figure 1: Samples produced by our method. (a–b) require physical priors, while (c–d) emphasize motion constraints.*PhyRPR* (*PhyReason*→*PhyPlan*→*PhyRefine*) decouples physical reasoning from rendering to better satisfy physical constraints while maintaining high-fidelity video quality.

video generation framework. It enables physically plausible generation as well as precise motion control. Extensive experiments demonstrate that our approach achieves clear improvements over existing methods in physical consistency, trajectory controllability, and overall visual quality.

2 Related Work

Physics video generation. Recent video generation models (e.g., WanXWan et al. (2025), CosmosAli et al. (2025)) achieve high-fidelity synthesis, but often lack physical understanding. Physics-centric benchmarks such as PhyGenBenchMeng et al. (2024) and VideoScience-BenchHu et al. (2025) show failures in physical consistency. Prior efforts improve plausibility via dynamics modeling Yuan et al. (2025), physical alignment Wang et al. (2025), verifiable rewards Le et al. (2025), or fine-tuning (VChain Huang et al. (2025)), but require large datasets or test-time training.

Unified Multimodal for Understanding and Generation. Large Multimodal Models (e.g., GPT-4Achiam et al. (2023), GeminiTeam et al. (2023)) motivate unifying understanding and generation, as explored in Show-oXie et al. (2024), BLIP3-oChen et al. (2025), and extended to videos by UniVideo Wei et al. (2025). However, training such unified video models is costly. We propose a training-free paradigm, *PhyRPR*, which orchestrates off-the-shelf LMM reasoning, deterministic coarse generation, and diffusion refinement to enable physics-constrained motion synthesis.

3 Method

3.1 *PhyReason*: Visually Grounded Physical Reasoning

To bridge textual physical understanding and video generation, *PhyReason* obtains visually grounded key physical states via (i) a large multimodal model (LMM) for reasoning and (ii) a high-quality single-frame image generator for synthesis. Given a user prompt $\mathcal{P}$, the LMM generates a physically constrained state-transition prompt sequence $\{p_i\}_{i=1}^{L}$, focusing on key kinematic moments (pivotal transitions and representative states) rather than uniformly sampling timestamps.

We further incorporate visual feedback for grounding. We first synthesize the initial keyframe $I_1 = \mathcal{G}_{\text{T2I}}(p_1)$; for each milestone $i = 2, \ldots, L$, the LMM observes the previous keyframe I_{i-1} and outputs a refined editing instruction e_i, which is executed by an instruction-guided editing model to produce $I_i = \mathcal{G}_{\text{edit}}(I_{i-1}, e_i)$. Here, $\mathcal{G}_{\text{T2I}}$ denotes the text-to-image generator and $\mathcal{G}_{\text{edit}}$ denotes the instruction-guided image editing model. Finally, we parse keyframes in an object-centric manner using an open-vocabulary segmentation model, extracting a binary mask $M_{i,k}$ for each dynamic entity o_k. Overall, *PhyReason* produces physically self-consistent keyframes and object states via grounded reasoning, serving as guidance for downstream planning and refinement.

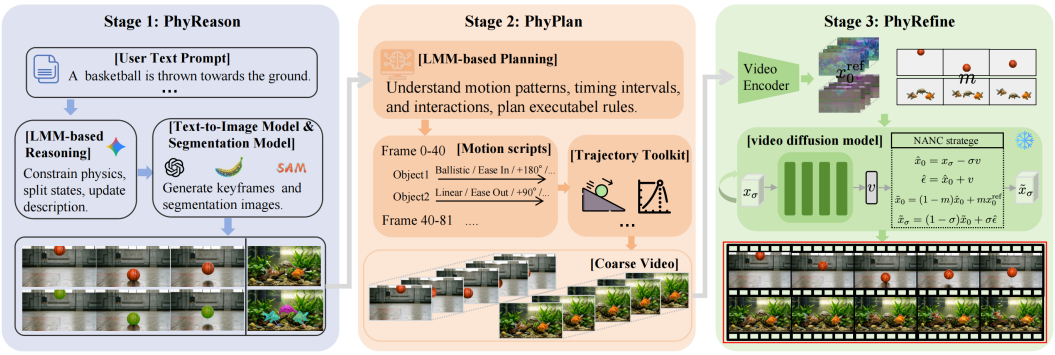

Figure 2: Overview of our training-free three-stage pipeline *PhyRPR*: *PhyReason*, *PhyPlan*, and *PhyRefine*. Stage 1 (see 3.1) outputs physically consistent keyframes and object states. Stage 2 (see 3.2) uses an LMM to select motion primitives and parameters, and deterministically renders a coarse motion video. Stage 3 (see 3.3) applies motion-aware noise-consistent injection(NANC) to enforce the planned kinematics while preserving visual coherence.

3.2 *PhyPlan*: Physics-Aware Motion Planning

Given the consistent keyframes produced by *PhyReason*, *PhyPlan* converts discrete states into continuous trajectories and renders a coarse video to guide refinement. Given I_i, $M_{i,k}$, and the prompt $\mathcal{P}$, it outputs a structured motion script with per-entity motion semantics and key states:

$$\mathcal{A} = \left\{ \left(\tau_k,\ \{\mathbf{s}_{i,k}\}_{i=1}^{L}\right) \mid o_k \in \mathcal{O} \right\}. \tag{1}$$

Here, $\mathcal{O}$ is the set of entities, τ_k denotes the primitive type (e.g., *Ballistic*, *Drifting*, *Linear*), and $\mathbf{s} * i, k = [x, y, s, r, \alpha]^\top$ encodes position, scale, rotation, and opacity at milestone i. We implement a toolkit for trajectory synthesis. For each pair of states, we instantiate $\mathcal{F} * \tau_k$ and fit parameters θ^* to satisfy boundary conditions. Let $\mathbf{c} * \text{start}$ and $\mathbf{c} * \text{end}$ denote positions; the trajectory is:

$$\mathbf{c}(t) = \mathcal{F}_{\tau_k}\left(\mathbf{c}_{\text{start}}, \mathbf{c}_{\text{end}}, t;\ \theta^*\right). \tag{2}$$

Finally, we render trajectories into a coarse visual signal by composing transformed objects onto an inpainted background. Let $\tilde{\mathbf{s}}_{t,k}$ be the per-frame state of entity o_k; we warp $M_{1,k}$ and appearance crop $K_{1,k}$ to obtain $\mathcal{M}_t$, then composite:

$$V_{\text{coarse}}^{(t)} = B^{(t)} \odot (1 - \mathcal{M}_t) + \sum\nolimits_k \mathcal{O}_{t,k}. \tag{3}$$

The coarse video V_{coarse} may be imperfect in texture but preserves topology and continuous trajectories, providing strong spatiotemporal guidance for the next stage.

3.3 *PhyRefine*: Motion-Aware Visual Refinement

The coarse video from *PhyPlan* provides a motion scaffold but lacks fine details. *PhyRefine* injects this scaffold as a latent constraint during sampling to improve motion adherence while letting the pretrained video model refine appearance. We downsample occupancy masks $\{\mathcal{M}_t\}$ to the latent resolution to obtain a broadcastable motion mask m, and encode V_{coarse} into a reference clean latent x_0^{ref}. We adopt the flow-matching interpolation path:

$$x_\sigma = (1 - \sigma)x_0 + \sigma\epsilon, \tag{4}$$

and at selected steps, given the current noisy latent x_σ and predicted velocity $v_\theta(x_\sigma, \sigma)$, we recover the implied clean/noise components and perform noise-consistent injection:

$$\hat{x}_0 = x_\sigma - \sigma v_\theta, \qquad \hat{\epsilon} = x_\sigma + (1 - \sigma)v_\theta, \tag{5}$$

$$\tilde{x}_0 = (1 - m)\hat{x}_0 + m x_0^{\text{ref}}, \qquad \tilde{x}_\sigma = (1 - \sigma)\tilde{x}_0 + \sigma\hat{\epsilon}. \tag{6}$$

This replaces only the clean component inside the motion region while preserving the noise $\hat{\epsilon}$, avoiding global distribution shift. The sampler continues from $\tilde{x}_\sigma$ (we apply injection in early steps, then sample normally), yielding videos that are both trajectory-consistent and visually coherent.

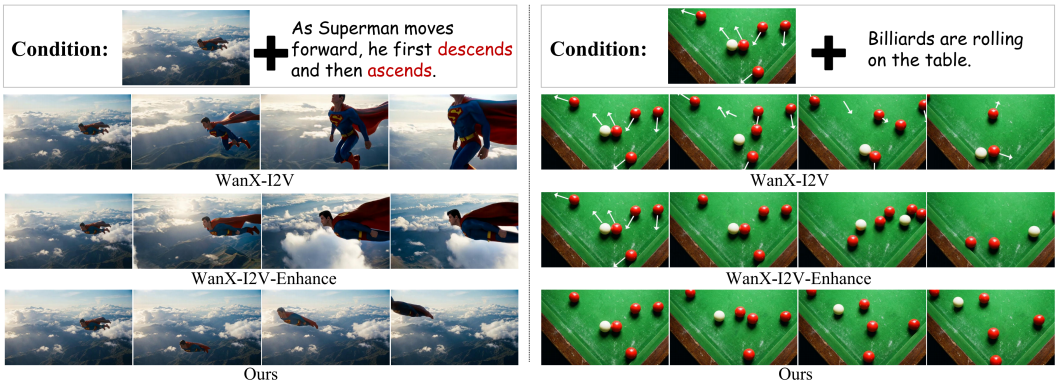

Figure 3: Qualitative comparison with baselines. We specify physical constraints via prompts or arrow guidance. Our method generates videos that more faithfully satisfy the physical constraints.

Method	VBench			LMM-as-judge (1-5)					User study (1-10)			
	Quality	Temporal	Overall	Physical plausibility	Trajectory compliance	Temporal consistency	Semantic alignment	Overall	Text alignment	Physics plausibility	Visual quality	Overall
WanX-T2V	59.79	95.78	83.78	2.95	2.50	3.10	2.15	2.68	6.00	5.51	6.92	6.14
WanX-T2V-Enhance	54.23	91.94	79.37	3.45	3.05	3.73	3.23	3.36	6.45	5.55	6.82	6.27
WanX-I2V	62.77	97.24	85.75	3.60	3.05	4.33	3.83	3.70	5.57	5.79	7.16	6.18
WanX-I2V-Enhance	60.88	95.26	83.80	3.85	3.63	4.45	4.13	4.01	6.18	5.99	7.49	6.56
LTX-Multi (-P)	59.59	97.54	84.89	3.65	3.88	4.10	4.00	3.91	6.32	5.86	6.02	6.07
SDEdit-style (-R)	61.66	97.71	85.70	3.55	3.43	4.38	3.95	3.83	6.13	6.01	6.93	6.35
PhyRPR (ours)	63.30	97.89	86.36	4.33	4.78	4.78	4.75	4.66	7.83	6.84	7.56	7.41

Table 1: We evaluate with VBench, LMM-as-judge, and User study. *-P,-R* denotes the ablation setting without PHYPLAN and PHYREFINE.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Implementation Details. In *PhyReason*/*PhyPlan*, we use `gemini-3-pro-preview` as the LMM and SAM3 Carion et al. (2025) for segmentation, and synthesize milestone keyframes with `Nano Banana Pro`. In *PhyRefine*, we use `Wan2.2-I2V-A14B` and perform training-free latent fusion during denoising. We evaluate both T2V and I2V settings with 40 diverse scenarios (text-only and image+prompt), using the same set for human evaluation, LMM-as-judge, and quantitative comparisons; I2V includes descriptive motion control and arrow-guided control.

Baseline Settings. We evaluate `Wan2.2-T2V-A14B` / `Wan2.2-T2V-Enhance` on T2V, and `Wan2.2-I2V-A14B` / `Wan2.2-I2V-Enhance` on all cases. All "-Enhance" variants use the same LMM-based rewriting to make physical/kinematic constraints explicit; for I2V, rewriting conditions on the reference first frame and prompt.

Evaluation Metrics. We use VBench for generic video quality, and Gemini as an LMM-as-judge (1–5) on *Physical plausibility*, *Trajectory compliance*, *Temporal consistency*, and *Semantic alignment*. We also run a user study with 12 participants (1–10) rating text alignment, physical plausibility, and visual quality.

4.2 BASELINES AND ABLATIONS

Table 1 reports quantitative results. I2V baselines generally achieve higher VBench scores than T2V baselines due to the reference first frame, while prompt enhancement ("-Enhance") improves constraint following (higher judge/user scores) at a small cost to VBench. Our method performs best across all metrics. Qualitatively (Fig. 3), baselines often drift from prompt-specified trajectories or confuse arrow directions, whereas our method follows the intended motion with better temporal coherence and physical plausibility.

We further ablate *PhyPlan* and *PhyRefine* (Table 1), where *-P* removes *PhyPlan* and *-R* removes *PhyRefine*; all variants share the same *PhyReason* outputs. For *-P,-R*, we use LTX-MultiHaCohen et al. (2024); for *-R*, we replace *PhyRefine* with an SDEdit-style refinement Meng et al. (2021) that globally denoises a noised V_{coarse} without selectively constraining motion regions.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Arslan Ali, Junjie Bai, Maciej Bala, Yogesh Balaji, Aaron Blakeman, Tiffany Cai, Jiaxin Cao, Tianshi Cao, Elizabeth Cha, Yu-Wei Chao, et al. World simulation with video foundation models for physical ai. *arXiv preprint arXiv:2511.00062*, 2025.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025.

Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, et al. Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. *arXiv preprint arXiv:2505.09568*, 2025.

Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, et al. Ltx-video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024.

Lanxiang Hu, Abhilash Shankarampeta, Yixin Huang, Zilin Dai, Haoyang Yu, Yujie Zhao, Haoqiang Kang, Daniel Zhao, Tajana Rosing, and Hao Zhang. Benchmarking scientific understanding and reasoning for video generation using videoscience-bench. *arXiv preprint arXiv:2512.02942*, 2025.

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

Minh-Quan Le, Yuanzhi Zhu, Vicky Kalogeiton, and Dimitris Samaras. What about gravity in video generation? post-training newton's laws with verifiable rewards. *arXiv preprint arXiv:2512.00425*, 2025.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021.

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

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

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. *arXiv preprint arXiv:2512.05564*, 2025.

Cong Wei, Quande Liu, Zixuan Ye, Qiulin Wang, Xintao Wang, Pengfei Wan, Kun Gai, and Wenhu Chen. Univideo: Unified understanding, generation, and editing for videos. *arXiv preprint arXiv:2510.08377*, 2025.

Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*, 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. *arXiv preprint arXiv:2509.21309*, 2025.