RETAKE: A Training-free Framework for Physically Plausible Video Generation

Junchen Zhou*, Zhixian He*, Yaosi Hu, Ye Liu, and Chang Wen Chen†

The Hong Kong Polytechnic University

Abstract. Text-and-image-to-video (TI2V) generators can synthesize visually plausible videos, but they often violate elementary physical constraints such as object integrity, mass conservation, gravity, and contact dynamics. Existing research typically relies on domain-specific fine-tuning or applying the inference-time physics reward to guide video sampling or filtering, leaving the user prompt's physical ambiguity largely unresolved. In this work, we present RETAKE, a training-free framework ensuring physically plausible video generation. Our framework consists of two components: *prompt rewriting* before generation and *multi-take selection* after generation. The first component employs a vision-language model to infer the implied dynamics given the initial frame and rewrites the prompt into a concrete, temporally ordered description of predictable motion and interactions. This makes the intended physical evolution explicit and faithful to the original intent. The second component leverages a physical plausibility evaluator to select the best take from multiple generated candidates. In the Physics-IQ benchmark, RETAKE improves the aggregate Physics-IQ Score by 13% compared to the Wan2.2 baseline.

Keywords: Text-Image-to-Video Generation · Physical Plausibility

1 Introduction

Current video diffusion models can generate clips with striking visual fidelity [8, 12, 16]. Yet visual fidelity does not imply physical correctness. Recent benchmarks show a clear gap between perceptual realism and physical understanding. Physics-IQ benchmark [7] shows that visual realism is weakly correlated with the capability to follow physical principles, while VideoPhy and VideoPhy2 [3, 4] find that leading models still produce common violations such as solid objects passing through one another, abnormal fluid flow, and motions that ignore gravity or momentum. These errors are especially problematic because they often occur in videos that otherwise appear realistic.

Existing approaches to improving physical consistency fall into two main categories. The first fine-tunes the generator on physics-rich data [6] or physics-informed training objectives [14, 18], which is effective but expensive and typically tied to a single model. The second keeps the generator frozen and applies

* *Equal contribution.* † *Corresponding author.*

a physics reward at inference time, either to guide sampling or to select among multiple generated videos. A representative method is WMReward [17], which estimates physical plausibility with the surprise score of a latent world model.

Inference-time rewards, however, are limited by the candidates produced from the original prompt. They can guide or filter generated videos, but they do not directly resolve physical ambiguity before generation. This limitation is important in text-and-image-to-video (TI2V) generation, where the user prompt often describes only the intended event at a high level. However, the TI2V task starts from a concrete visual state, where the objects, their spatial relations, and the starting configuration of the scene are fixed. As a result, the physically relevant dynamics are not generic.

Therefore, we introduce RETAKE, a training-free framework that combines frame-grounded prompt rewriting with physics-aware multi-take selection. Given the user prompt and the initial frame, a Vision-Language Model (VLM) infers the physical dynamics from them and rewrites the original prompt into a physically grounded prompt that specifies the observable motion in concrete, temporally ordered terms while remaining faithful to the original intent. The frozen generator then samples N candidate videos conditioned on the refined prompt and initial frame, and a best-of-N selection stage returns the take with the highest physical rationality score, as measured by a physical plausibility evaluator [17]. Our contributions are summarized as follows:

- We propose RETAKE, a training-free TI2V framework that resolves physical ambiguity before generation via frame-grounded prompt rewriting: a VLM infers the implied dynamics from the initial frame and rewrites the prompt into concrete, temporally ordered motion.
- We show that our prompt enhancement solution can be paired with physics-aware multi-take selection techniques, thereby improving the physical plausibility of the generated videos.
- Extensive experiments on the Physics-IQ benchmark demonstrate the effectiveness and significance of the proposed scheme.

2 Related Work

Approaches to physically plausible video generation differ in the access they require to the generator. *Training-based* methods modify model parameters in various ways: distilling physical context into prompts before fine-tuning [18], integrating learned physics priors from V-JEPA-2 features or simulator-derived dynamics [13], or applying preference optimization against curated real videos [5]. While effective, these methods are model specific and costly to retrain. *Inference-time* methods instead keep the generator frozen. Reward-based approaches apply a physical-plausibility reward to guide denoising or to rank candidates via best-of-N [17], where guidance needs white-box access while best-of-N needs only completed videos. Prompt-rewriting approaches refine the prompt with a language model before generation [15].

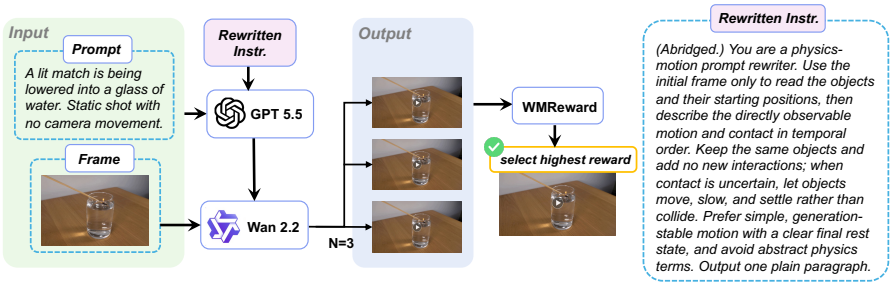

Fig. 1. Overview of the proposed RETAKE pipeline.

Progress is commonly evaluated on benchmarks such as Physics-IQ [7] and VideoPhy [3, 4]. We target the black box TI2V setting with closed-source generators. Unlike existing reward-based selection and text-only prompt rewriting, which ignore the scene physics in the visual input, RETAKE rewrites the prompt from the input frame before generation and then applies WMReward-based best-of-N selection.

3 Method

Given a text prompt T and an initial frame I, a frozen TI2V generator G that samples a clip $x = G(T, I, \epsilon)$ from a noise seed ϵ, a vision-language model V that reads images with a text instruction and returns text, and a physics reward r that scores a clip for physical plausibility. RETAKE leaves G, V, and r unchanged. The first stage uses V to rewrite the prompt before any clip is generated, with I as a physical context. The second stage samples several clips from the rewritten prompt and returns the one that r scores the highest, where r is instantiated with WMReward [17]. Figure 1 provides an overview of our pipeline.

3.1 Frame-Grounded Prompt Rewrite

A TI2V prompt typically specifies the intended event at a high level, while the actual movement of the scene is not mentioned. The frame I provides this context directly: the objects present, their arrangement, and the state from which motion begins. We leverage it by using the VLM to reason over I and infer the dynamics it implies, enriching rather than restating the prompt. Under a fixed and physics-oriented instruction, V reads I, anchors it to the intent in T, and returns a single grounded prompt.

$$T' = V(I, T). \tag{1}$$

The inference is verbalized as a concrete, temporally ordered chain of observable events, from initial setup, onset of motion, trajectory, and contact to final rest state, making the intended physical evolution explicit.

The instruction shapes this reasoning through four principles. *(i) Faithful grounding:* I is read only to identify which objects are present and where, so the

rewrite adds no new objects, collisions, or chain reactions beyond what T states or I shows. *(ii) Observable over abstract:* dynamics are cast as visible events (release, fall, slide, bounce, spill, settle) rather than textbook quantities such as force or momentum. *(iii) Conservative under uncertainty:* ambiguous outcomes commit to one plausible result, with objects moving, slowing, and settling rather than colliding. *(iv) Generator-friendly form:* a single short paragraph ending in a clear rest state, preserving the original camera and prioritizing motion over already-visible appearance. Together they turn the frame and prompt into one physically grounded specification before any clip is generated.

3.2 Best-of-N Selection

Given the rewritten prompt T' and the frame I, we draw N clips $\{x_1, \ldots, x_N\}$ from the frozen generator with independent noise seeds, $x_n = G(T', I, \epsilon_n)$. Each clip is scored by the reward r, and RETAKE returns the highest-scoring one,

$$x^\star = \arg\max_{n \in \{1,\ldots,N\}} r(x_n). \tag{2}$$

We adopt WMReward [17] as r, WMReward turns the prediction surprise of a pretrained latent world model into a physical-plausibility signal. It slides a window along the clip, uses the world model to predict the latent representation of upcoming frames from earlier context frames, and compares the prediction with the representation of the frames the generator actually produced. A clip whose continuation the world model predicts well has low surprise and, therefore, a high reward. Because r operates on the finished clip, the second stage needs no access to the sampler.

Together, the two stages improve the physical plausibility of the output while keeping the generator frozen.

4 Experiments

4.1 Evaluation Settings

We evaluate our method on the Physics-IQ benchmark [7], which assesses physical understanding by conditioning on a real frame and comparing the generated continuation against the ground truth. We report the four constituent metrics (Spatial IoU, Spatiotemporal IoU, Weighted Spatial IoU, and MSE) along with the aggregate Physics-IQ Score.

All experiments are conducted under the image-and-text conditioned (TI2V) setting. Given the benchmark conditioning frame I and the original text prompt, we use GPT-5.5 [9] to generate a refined prompt conditioned on both modalities. This refined prompt is then fed to Wan2.2 [12], which produces 5-second videos at a resolution of 1280×720. For each sample, we generate $N = 3$ candidate videos and apply WMReward [17] to score them, selecting the highest-scoring video for evaluation on Physics-IQ.

Table 1. Evaluation results on the Physics-IQ benchmark (TI2V setting by default). Higher is better except for MSE. Spatial, ST-IoU, and W-IoU denote spatial IoU, spatiotemporal IoU, and weighted spatial IoU. VideoMAE, Qwen2.5-VL, and Qwen3-VL use BoN with $N = 16$ [17], while Rewrite + WMReward utilizes $N = 3$.

Method	BoN	Spatial↑	ST-IoU↑	W-IoU↑	MSE↓	Score↑
Sora (I2V) [8]	✗	0.138	0.047	0.063	0.030	10.0
Runway Gen 3 (I2V) [10]	✗	0.201	0.115	0.116	0.015	22.8
Wan2.2 [12]	✗	0.358	0.120	0.217	0.008	38.26
+ VideoMAE [11]	✓	0.347	0.124	0.216	0.008	37.91 (-0.35)
+ Qwen2.5-VL-7B [2]	✓	0.356	0.116	0.213	0.009	37.58 (-0.68)
+ Qwen3-VL-8B [1]	✓	0.353	0.128	0.216	0.008	38.51 (+0.25)
+ Rewrite	✗	0.363	0.149	0.227	0.008	40.71 (+2.45)
+ Rewrite + WMReward [17]	✓	**0.372**	**0.168**	**0.243**	0.009	**43.26** (+5.00)

4.2 Quantitative Results

Table 1 presents our evaluation results with Wan2.2 [12] on the Physics-IQ [7] benchmark. Our method RETAKE achieves the highest Physics-IQ Score of 43.26, outperforming the vanilla Wan2.2 baseline[17] by +5.00 points. The improvement is consistent across all IoU-based metrics, increasing Spatial IoU from 0.358 to 0.372, Spatiotemporal IoU from 0.120 to 0.168, and Weighted Spatial IoU from 0.217 to 0.243.

The ablation study shows that prompt rewriting alone already improves physical consistency, raising the overall physical score from 38.26 to 40.71 (+2.45). This suggests that explicitly emphasizing physically plausible object dynamics in the prompt helps the generator produce more realistic motion and interactions.

We further compare our method against alternative selection signals derived from foundation models, including VideoMAE [11], Qwen2.5-VL [2], and Qwen3-VL [1]. Despite using a substantially larger search budget ($N = 16$) [17], all three alternatives fail to consistently outperform the vanilla baseline. In contrast, WMReward, combined with prompt rewriting, achieves the best results with only $N = 3$ candidates. These results indicate that physics-aware prompt refinement and latent-world-model-based selection are complementary, jointly improving the physical realism of generated videos.

4.3 Qualitative Results

The examples in Figure 2 highlight key differences between our method and the baseline. In the upper example, a tangerine is cut by a knife. After the cut, the interior of the tangerine generated by the baseline no longer resembles a real tangerine. The baseline also produces an implausible final state in which the knife remains embedded in the fruit. In contrast, our method preserves the appearance of the tangerine throughout the sequence and generates a more realistic outcome after the cut.

The lower example depicts a potato being dropped into a cup of blue liquid. The conditioning image contains only a potato suspended above the cup, with

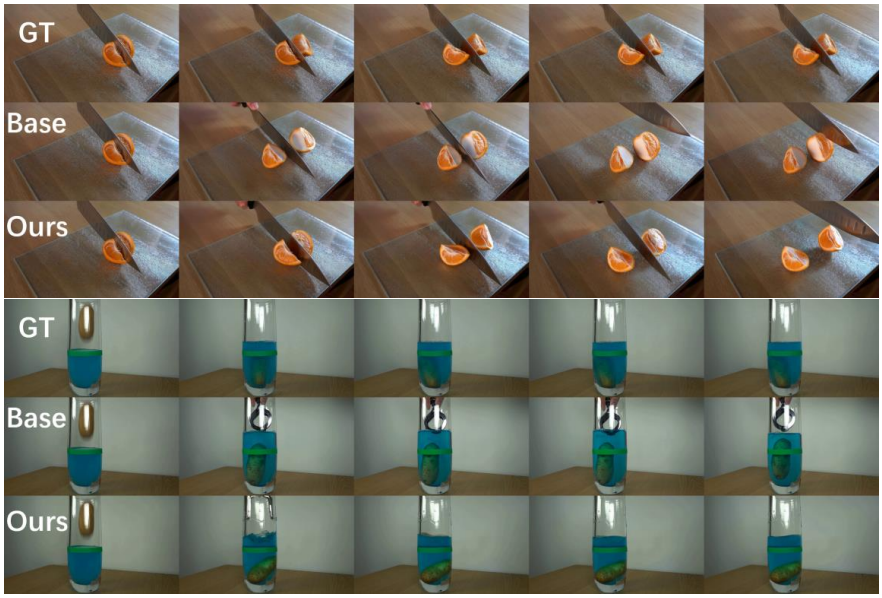

Fig. 2. Qualitative comparisons between our method and our baseline.

no visible gripper. Nevertheless, the baseline introduces a gripper in subsequent frames despite its absence from the input image. Our method avoids this artifact and remains faithful to the conditioning frame. Furthermore, the baseline predicts the potato remaining suspended in the liquid at the end of the video, whereas both the ground truth and our result show the potato settling at the bottom of the cup. This suggests that our method more accurately captures the underlying physical behavior of the scene.

Overall, our method generates results that are more faithful to the conditioning image while better preserving realistic object dynamics and state transitions.

5 Conclusion

We presented RETAKE, a training-free framework that improves the physical plausibility of TI2V generation. By combining an input-side prompt rewrite mechanism based on the conditioning frame with an output-side best-of-N selection through WMReward, RETAKE requires only black-box access and applies directly to closed generators. Together, these complementary steps yield consistent Physics-IQ gains while preserving prompt adherence and visual quality. Under white-box access, the rewrite can also pair with WMReward guidance to potentially further boost performance.

References

1. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
2. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., et al.: Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
3. Bansal, H., Lin, Z., Xie, T., Zong, Z., Yarom, M., Bitton, Y., Jiang, C., Sun, Y., Chang, K.W., Grover, A.: Videophy: Evaluating physical commonsense for video generation. In: International Conference on Learning Representations (2025)
4. Bansal, H., Peng, C., Bitton, Y., Goldenberg, R., Grover, A., Chang, K.W.: Videophy-2: A challenging action-centric physical commonsense evaluation in video generation. In: International Conference on Learning Representations (2026)
5. Chen, H.H., Huang, H., Chen, Q., Yang, H., Lim, S.N.: Hierarchical fine-grained preference optimization for physically plausible video generation. In: Advances in Neural Information Processing Systems. vol. 38, pp. 133919–133951 (2026)
6. Li, C., Michel, O., Pan, X., Liu, S., Roberts, M., Xie, S.: Pisa experiments: Exploring physics post-training for video diffusion models by watching stuff drop. In: International Conference on Machine Learning (2025)
7. Motamed, S., Culp, L., Swersky, K., Jaini, P., Geirhos, R.: Do generative video models understand physical principles? In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 948–958 (2026)
8. OpenAI: Video generation models as world simulators (2024), https://openai.com/index/video-generation-models-as-world-simulators/
9. OpenAI: Gpt-5 system card (2025), https://openai.com/index/gpt-5-system-card/
10. Runway: Building ai to simulate the world (2024), https://runwayml.com/
11. Tong, Z., Song, Y., Wang, J., Wang, L.: Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In: Advances in neural information processing systems. vol. 35, pp. 10078–10093 (2022)
12. Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., Zeng, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
13. Wang, C., Chen, C., Huang, Y., Dou, Z., Liu, Y., Gu, J., Liu, L.: Physctrl: Generative physics for controllable and physics-grounded video generation. In: Advances in Neural Information Processing Systems. vol. 38, pp. 167907–167932 (2026)
14. Wang, P., Wang, W., Li, Q.: Physcorr: Dual-reward dpo for physics-constrained text-to-video generation with automated preference selection. arXiv preprint arXiv:2511.03997 (2025)
15. Xue, Q., Yin, X., Yang, B., Gao, W.: Phyt2v: Llm-guided iterative self-refinement for physics-grounded text-to-video generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 18826–18836 (2025)
16. Yang, Z., Teng, J., Zheng, W., et al.: Cogvideox: Text-to-video diffusion models with an expert transformer. In: International Conference on Learning Representations (2025)
17. Yuan, J., Zhang, X., Friedrich, F., Beltran-Velez, N., Hall, M., Askari-Hemmat, R., Han, X., Ballas, N., Drozdzal, M., Romero-Soriano, A.: Inference-time physics alignment of video generative models with latent world models. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (2026)
18. Zhang, K., Xiao, C., Xu, J., Mei, Y., Patel, V.M.: Think before you diffuse: Infusing physical rules into video diffusion. arXiv preprint arXiv:2505.21653 (2025)