Chain of Event-Centric Causal Thought for Physically Plausible Video Generation

Supplementary Material

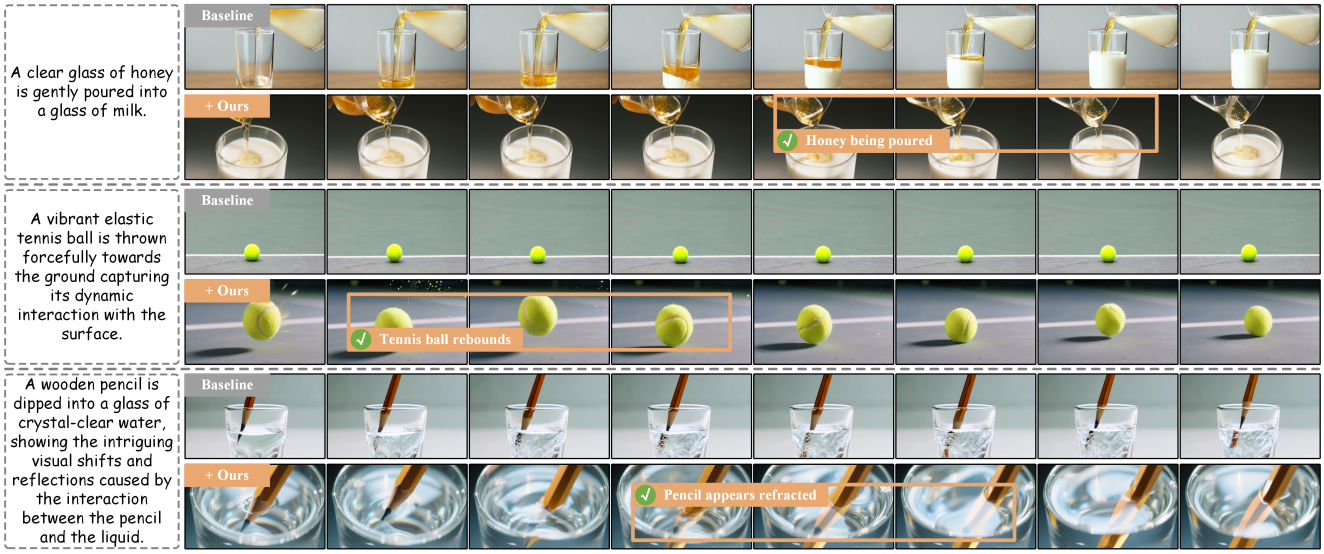

Figure 1. Visualization of physics-aware video generation results across diverse physical domains. Compared with baseline Wan2.1-14B [1], our approach yields causally coherent progressions of physical phenomena, *e.g.*, continuous honey flow, elastic rebound of the tennis ball, and optical distortion of the pencil. All prompts are sourced from PhyGenBench [2].

In this supplementary material, we provide additional information in the following aspects:

A. visual analysis of module effectiveness.
B. comprehensive evaluation on other baselines;
C. discussions of limitations and future works;

A. Visual Evaluation of TCP Module

We evaluate the effectiveness of the semantic prompts in the proposed TCP module. As shown in Fig. 2, using semantically compressed event descriptions as semantic prompts enables our framework to generate a physically plausible scene. The generated scene depicts dry-ice sublimation that gradually intensifies as the temperature increases. In contrast, the baseline produces a scene without a clear temporal progression of the phenomenon. This is because simply concatenating multiple event descriptions causes the model to ignore subtle cues that encode change. This indicates that **semantic condensation** is crucial for revealing the temporal evolution of physically grounded phenomena.

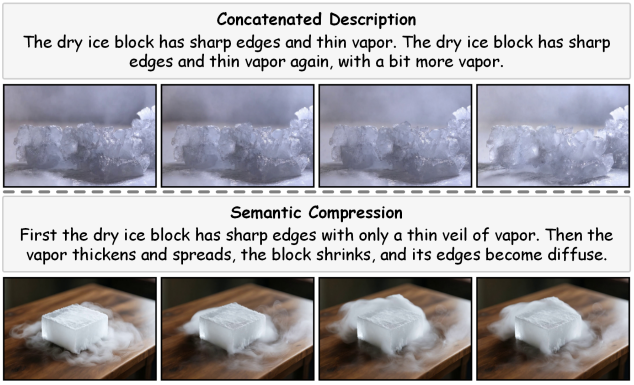

Figure 2. Ablation results of the semantic prompt in TCP module. Leveraging semantically compressed event descriptions as prompts enables the model to capture clear temporal evolution, while naive concatenation fails to reveal such changes.

B. Evaluations on Other Baselines

We use Wan2.1-14B [1] as the baseline to evaluate the effectiveness of our framework across diverse physical domains on PhyGenBench [2]. As shown in Tab. 1, our framework consistently improves performance across all four physical domains compared with the baseline. As shown in Fig. 1, our framework achieves visually realistic generation of honey pouring (top row), tennis-ball bounce (middle row), and pencil refraction (bottom row). These results demonstrate the effectiveness of our framework in enhancing the **physical fidelity** of diverse scenarios.

Table 1. Performance comparison on PhyGenBench [2]. These quantitative results demonstrate that our framework consistently improves physical plausibility across all domains, achieving the best average performance on PhyGenBench.

Methods	Physical domains (↑)				Avg.
	Mechanics	Optics	Thermal	Material	
Wan2.1-14B [1]	0.36	0.53	0.36	0.33	0.40
+ SGD [3]	0.47	0.60	0.51	0.40	0.50
+ Ours	**0.64**	**0.67**	**0.64**	**0.58**	**0.63**

C. Limitations and Future Works

As shown in Fig. 3, our framework achieves the best performance with 4 events across all physical domains. Fewer events (e.g., 1–3) provide weak temporal supervision, making it difficult for the model to follow instructions (describing physical process) accurately, thus reducing PCA scores. While, increasing the number of events (e.g., 5–6) introduces accumulated errors in keyframe generation due to editing-based propagation, leading to poor leading signals and degraded video quality. Cause of error accumulation is shown in Fig. 4. Under progressive keyframe editing, subsequent keyframes gradually deviate from earlier ones (*e.g.*, partial re-solidification). Since each edit applies only local adjustments, small deviations unavoidably arise per step. These deviations accumulate over successive edits, leading to **semantic drift in physical states** across keyframes. In future work, a more capable generative model could mitigate such an issue.

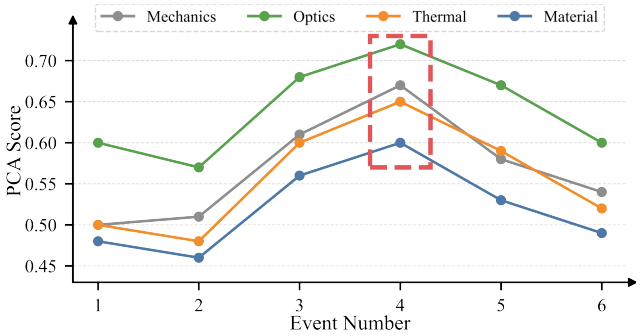

Figure 3. Effect of event number on Physical Commonsense Alignment (PCA) Score [2] across four physical domains: Mechanics, Optics, Thermal, and Material.

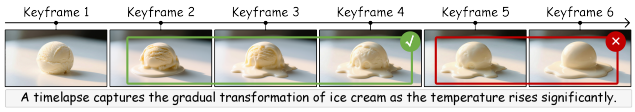

Figure 4. Visualization of progressive keyframe editing.

References

[1] 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. 1, 2

[2] 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. 1, 2

[3] 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. 2