

Tempered Self-Similarity Alignment for Physically Plausible Video Generation

- *Supplementary Material* -

1. Additional Qualitative Results

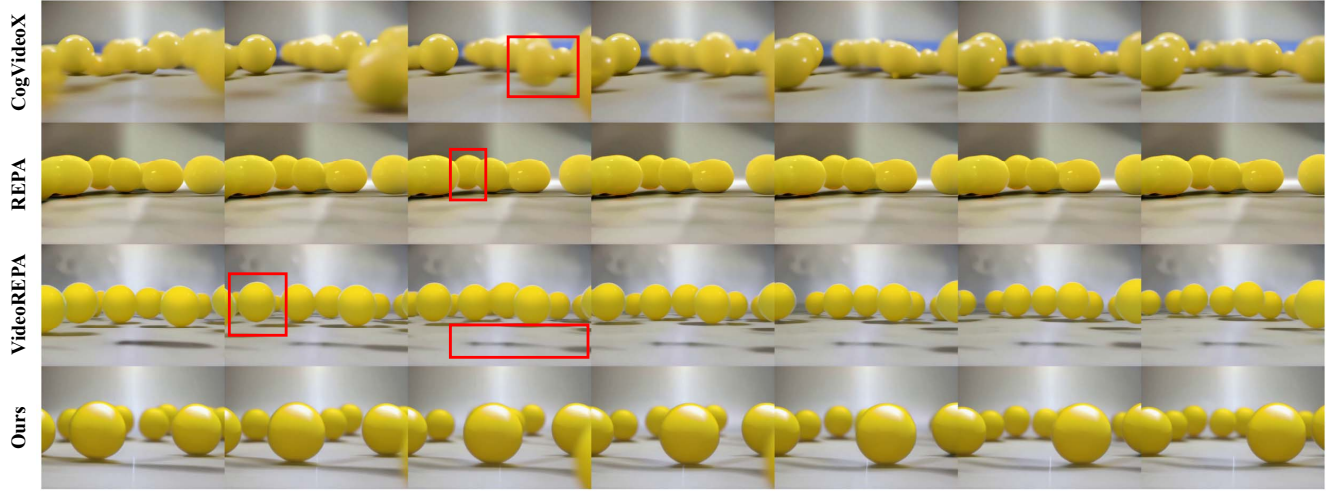
In Fig. 1, we provide additional qualitative comparisons of videos generated by CogVideoX [2], REPA [3], VideoREPA [4], and our method. In Fig. 1a, existing methods produce unrealistic dynamics such as floating balls or misplaced shadows, while our method generates physically grounded ball movement. In Fig. 1b, competing methods exhibit shape distortion or unnatural snowmobile motion, whereas our method produces temporally consistent and physically plausible motion dynamics. In Fig. 1c, existing methods suffer from artifacts such as elongated hockey sticks or spuriously appearing objects, while our method maintains realistic and coherent motion throughout. Full videos of the samples included in the paper are available in the accompanying PPT file submitted as part of our supplementary material.

Overall, our method consistently improves physical plausibility in generated videos. However, as shown in Fig. 1a and Fig. 1c, it still fails to faithfully follow the text prompt. We attribute this limitation to the relatively small scale of our base model, and expect that applying our method to larger models such as CogVideoX-5B [2] or HunyuanVideo [1] would alleviate this issue. We leave this as future work due to computational resource constraints.

orepa: Learning physics for video generation through relational alignment with foundation models. *arXiv preprint arXiv:2505.23656*, 2025. 1

References

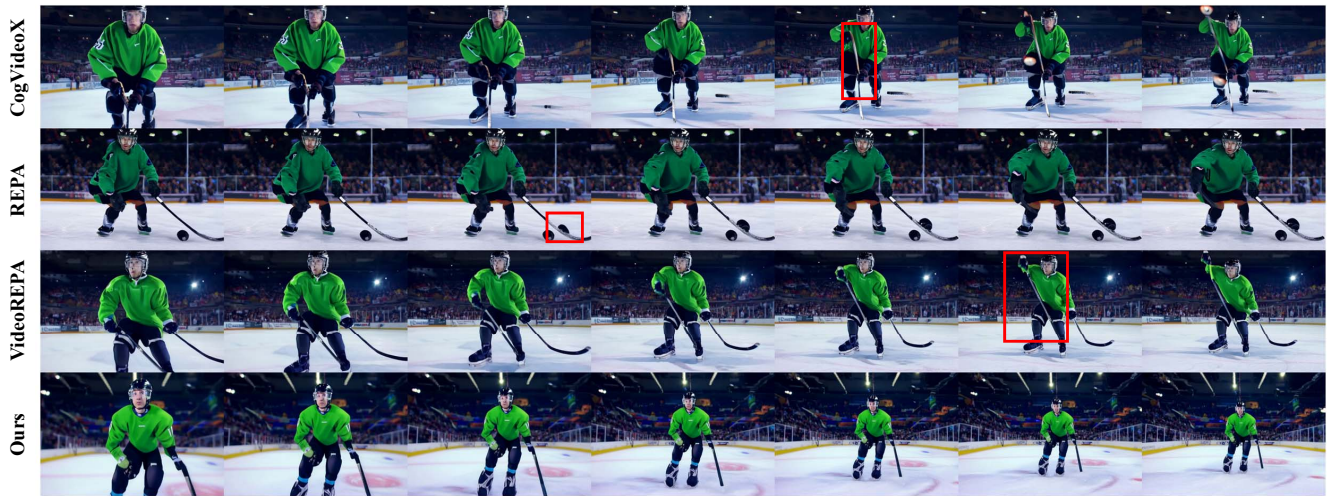
- [1] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024. 1
- [2] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 1
- [3] Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. *arXiv preprint arXiv:2410.06940*, 2024. 1
- [4] Xiangdong Zhang, Jiaqi Liao, Shaofeng Zhang, Fanqing Meng, Xiangpeng Wan, Junchi Yan, and Yu Cheng. Vide-



(a) Prompt: Multiple ping pong balls are shown, illustrating different stages of a rally.



(b) Prompt: A snowmobile drives across a snowy lake, kicking up a large spray of snow behind it.



(c) Prompt: A player performs a drag flick, hitting the ball forcefully from the ground towards the goal.

Figure 1. **Qualitative results.** Red rectangles highlight regions with physically implausible or temporally inconsistent motion. Our method generates videos with physically realistic dynamics.