DynamicsBoost: Dynamic Plausible Video Generation via Annotation-Free Continuation Preference Optimization

Supplementary Material

1. User Study

To evaluate human subjective preference, we conducted a user study on results from VBench, PhysGenBench, and VideoGenBench (corresponding to the evaluations in Sec. 4.2 of the main paper). From each benchmark, we randomly selected five videos, forming a total of 15 evaluation cases. For each case, participants were shown anonymous and randomly shuffled outputs from all compared methods and were asked to select the best result along three dimensions: overall video quality (VQ), text–video alignment (TA), and dynamic plausibility (DP). We distributed the questionnaire to 30 independent observers and recorded all selections. As shown in Table 1, our method is consistently preferred over existing state-of-the-art baselines, achieving the highest user preference across all three evaluation dimensions.

Table 1. User Study Results. Our method outperforms all baselines across all three evaluation dimensions.

	VQ (%)	TA (%)	DP (%)
Pretrain	10.00	3.33	10.00
SFT	6.67	3.33	6.67
Flow-DPO	13.33	10.00	13.33
Flow-Structural DPO	16.67	13.33	16.67
Flow-Dense DPO	13.33	16.67	10.00
Ours	**40.00**	**53.33**	**43.33**

2. Additional Experiments

2.1. Video Continuation

Continuation Pairs under VLM Perspective vs. Human Perspective. We further investigate how the continuation-induced partial order manifests under VLM scoring versus human judgment. We consider four continuation settings, $k \in 1, 4, 8, 13$, where k denotes the number of reference frames provided for latent-space continuation. From the evaluation data (corresponding to Sec. 4.3 in the main paper), we randomly sample 15 videos and generate continuations for each video under all four settings. For each video, the four continuation results are jointly evaluated by a VLM scorer and by 30 human participants. The VLM produces an automatic ranking based on its scoring function, whereas human evaluators are instructed to manually rank the four continuations by considering both *overall visual quality* and *dynamic plausibility*. We assign weighted scores according to the ranking positions and aggregate results across all videos to obtain a final score for each k. The statistics after normalization are reported in Table 2.

Table 2. Verification of continuation-induced partial order under VLM scoring and human evaluation. Darker green indicates higher scores.

Continuation Setting	VLM Ratio	Human Ratio
$k = 1$	0.152	0.088
$k = 4$	0.243	0.191
$k = 8$	0.285	0.289
$k = 13$	0.320	0.433

Results show that the continuation-induced partial order becomes substantially clearer under human evaluation: as k increases, the human preference score increases monotonically with a pronounced margin, while the gradient under VLM scoring is noticeably weaker. This indicates that continuation-based ordering aligns more strongly with human preference than with VLM preference, providing further evidence that continuation-driven, annotation-free preference signals are naturally well-suited for DPO-style alignment.

Table 3. Additional ablation results on continuation model design.

Setting	SSIM↑	LPIPS↓	CLIP-S↑	FVD↓
w/o continuation prompt	0.969	**0.012**	0.948	44.962
w/o timestep mask	0.971	0.014	0.940	46.245
Full setting	**0.973**	0.013	**0.954**	**42.586**

Continuation Model Design. As a complement to Sec. 4.3.4 in the main paper, we further evaluate different continuation model designs by measuring video-level reconstruction quality under the 4/13 conditioning-frame setting. Following the same evaluation split as the main text, we report SSIM[2], LPIPS[3], CLIP-S, and FVD[1] in Table 3. The results show that removing either the continuation prompt or the timestep mask leads to consistent degradation across multiple metrics, particularly in semantic similarity (CLIP-S) and temporal coherence (FVD). In contrast, the full model design achieves the best overall performance, indicating that both components are essential for stable and accurate continuation modeling.

Figure 1. Continuation results across diverse scenarios. (Please zoom in.)

3. Disscussion

3.1. Scene Performance

Performance differences across scenarios depend on the scene sensitivity of continuation. Fig. 1 illustrates (a) purely static, (b) camera-motion, and (c) dynamic cases. For (a), continuations under different contexts exhibit only minor variations; in practice, such excessive similarity can push the optimization toward instability during training. However, most real-world video backgrounds are low-dynamic rather than strictly static. In this regime, continuations still exhibit distinguishable and monotonic differences, remaining effective for DPO, although the performance gains are smaller compared to highly dynamic scenes (e.g., the second-best *background metrics* reported in Tables 1 and 2 of the main paper).

For (b), even when the primary subject remains static, camera motion can still introduce noticeable differences in continuation, suggesting that the idea behind DynamicsBoost may be extended to camera-conditioned tasks or other related regimes.

3.2. Limitation

Although our method provides substantial improvements in dynamic plausibility, temporal coherence, and semantic alignment, it still has several limitations. First, the quality of continuation-based preference pairs strongly depends on the quality of the source video (i.e., the ground-truth segment used for continuation). When the source video contains motion artifacts, severe blur, or structural inconsistencies, the resulting continuations may inherit these imperfections, weakening the reliability of the induced preference ordering. Second, for out-of-distribution (OOD) source videos, the continuation model may generate content that deviates significantly from the distribution of the base video generator, leading to unreliable preference pairs and, in extreme cases, optimization collapse during DPO training. As discussed in the paper, applying an SFT cold-start before the DPO stage helps stabilize training by partially realigning the continuation model with the pretrained generator.

References

[1] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018. 1

[2] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 1

[3] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 1