
Supplementary Materials for: Epipolar Geometry Improves Video Generation Models

Anonymous Author(s)

Affiliation

Address

email

1 Generalization to Dynamic Objects

To assess whether our approach generalizes beyond its training domain of static scenes with dynamic cameras, we evaluate the finetuned model on videos containing both dynamic objects and camera movement. We follow the same evaluation protocol, generating 200 videos using both baseline and epipolar-aligned models with captions describing scenes featuring moving objects alongside camera motion. As shown in Table 1, our model demonstrates robust generalization, achieving win rates of 57.0%, 58.0%, and 52.0% for Visual Quality, Motion Quality, and Text Alignment respectively, with an overall score of 58.5%. This generalization can be attributed to the fact that aligning the model with smoother and more geometrically consistent camera trajectories inherently improves video quality even when objects are also moving. Intuitively, the primary source of error in highly dynamic scenes stems from unstable motion trajectories, artifacts, and flickering, which become amplified by additional object movement and cause generation instability. By learning to produce more stable camera motion and reducing geometric inconsistencies, our epipolar-aligned model addresses these fundamental issues, automatically improving the quality of dynamic object generation as well. Some visual examples are presented in Figure 1.

Table 1: **Win-rate vs. original model** on the VideoReward [1] benchmark on a set of videos with dynamic objects [2].

Text-to-Video				
Method	Visual Quality	Motion Quality	Text Alignment	Overall
DPO-Epipolar	57.0%	58.0%	52.0%	58.5%

Table 2: **Win-rate vs. Wan-2.1-14B** [3] on the VideoReward [1] benchmark. The Baseline and Epipolar-Aligned Model contain only 1.3B parameters.

Text-to-Video				
Method	Visual Quality	Motion Quality	Text Alignment	Overall
Baseline	13.3%	14.4%	24.2%	8.6%
DPO-Epipolar	18.1%	21.8%	25.0%	13.8%

2 Scaling Analysis

To understand how our geometric alignment performs across different model scales, we compare both the baseline and epipolar-aligned 1.3B parameter models against the much larger Wan-2.1-14B model [3]. As shown in Table 2, while the performance gap remains substantial due to the 14B model’s



Figure 1: **Qualitative Evaluation:** Comparison of baseline and epipolar-aligned models on dynamic scenes featuring both camera movement and object motion. Our approach maintains improved geometric consistency and smoother trajectories, demonstrating generalization beyond static scene training. Best seen in the supplementary video.

higher resolution (720p) and superior base capabilities, our epipolar alignment helps close this gap meaningfully. The aligned 1.3B model achieves win rates of 18.1%, 21.8%, and 25.0% for Visual Quality, Motion Quality, and Text Alignment respectively, compared to 13.3%, 14.4%, and 24.2% for the baseline 1.3B model. Notably, the 14B model requires approximately 10× longer inference time than the 1.3B variant, making our alignment approach particularly valuable for applications where computational efficiency is critical. This suggests that geometric consistency improvements can partially compensate for scale limitations, offering a practical path toward better video quality without the computational overhead of significantly larger models.

3 Qualitative Evaluation

For comprehensive assessment of video quality and geometric consistency, we include an interactive webpage in the supplementary materials where readers can view the full video sequences and directly compare the baseline and epipolar-aligned model outputs.

References

- [1] Jie Liu, Gongye Liu, Jiajun Liang, Ziyang Yuan, Xiaokun Liu, Mingwu Zheng, Xiele Wu, Qiulin Wang, Wenyu Qin, Menghan Xia, et al. Improving video generation with human feedback. *arXiv preprint arXiv:2501.13918*, 2025.
- [2] Xuan Ju, Yiming Gao, Zhaoyang Zhang, Ziyang Yuan, Xintao Wang, Ailing Zeng, Yu Xiong, Qiang Xu, and Ying Shan. Miradata: A large-scale video dataset with long durations and structured captions. *Advances in Neural Information Processing Systems*, 37:48955–48970, 2024.
- [3] Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.