

GeoHOI: Geometry-Enhanced Human-Object Interaction Video Generation via Hierarchical Multi-Modal Injection

Supplementary Material



Figure S1. Qualitative comparisons with other methods on HunyuanVideo-HOMA test. Our method generates more realistic interactive motions while faithfully preserving the high-fidelity geometry and appearance of objects.

A. Results of HunyuanVideo-HOMA test set

The quantitative experimental results on the HunyuanVideo-HOMA [1] test set are presented in Table S1. The HunyuanVideo-HOMA model is specifically designed for generating videos that focus on human-object interactions. Wan2.2-5B [3] is modified to include hand pose and object trajectory, in alignment with our method, and is trained on our human interaction dataset. Our method achieves the best performance across FID, Object-IoU, Object-CLIP, and LMD (hand), clearly demonstrating superior object consistency and interaction plausibility. It is worth noting that, Unianimate-DiT [4] and VACE [2] are 14B-parameter models, while HunyuanVideo-HOMA is a 13B-parameter model. Across all evaluated metrics, our approach consistently demonstrates superior performance compared to Wan2.2-5B.

The qualitative results are illustrated in Figure S1. Com-

pared with other approaches, our method achieves more realistic interactive motion while maintaining both geometric and appearance consistency of objects. In contrast, other methods often suffer from issues such as object disappearance, object deformation, and incorrect actions.

Method	FID↓	FVD↓	Obj-IoU↑	Obj-CLIP↑	LMD (Hand)↓
MimicMotion [5]	40.53	389.52	0.59	84.00	17.32
VACE-14B [2]	40.35	271.05	0.59	87.30	22.14
UniAnimate-DiT [4]	31.88	200.04	0.66	88.13	10.19
HunyuanVideo-HOMA [1]	31.89	271.46	0.65	87.48	9.61
Wan2.2-5B [3]	35.74	311.78	0.63	87.14	10.97
Ours	30.65	308.77	0.70	89.60	8.94

Table S1. Quantitative results of our method compared with SO-TAs in HunyuanVideo-HOMA test set.

B. Model Size vs. Inference Speed Analysis

Method	Model Size (B)	Speed (second)
MimicMotion [5]	1.5	337
VACE-14B [2]	14.0	701
UniAnimate-DiT [4]	14.0	1154
HunyuanVideo-HOMA [1]	13.0	-
Wan2.2-5B [3]	5.0	156
Ours	5.3	181

Table S2. Comparison of model size and inference speed.

As shown in Table S2, we report the approximate parameter counts and the inference time for generating a video of resolution 480×864 with 129 frames. Our model demonstrates high efficiency: it introduces only 0.3B additional parameters compared to the baseline Wan2.2-5B, which is substantially fewer than VACE, UniAnimate-DiT, and HunyuanVideo-HOMA. In terms of inference, the latency increases by merely a few seconds relative to Wan2.2-5B, while delivering enhanced output quality, making our approach considerably faster than competing methods.

References

- [1] Ziyao Huang, Zixiang Zhou, Juan Cao, Yifeng Ma, Yi Chen, Zejing Rao, Zhiyong Xu, Hongmei Wang, Qin Lin, Yuan Zhou, Qinglin Lu, and Fan Tang. Hunyuanvideo-homa: Generic human-object interaction in multimodal driven human animation, 2025. 1
- [2] Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. Vace: All-in-one video creation and editing. *arXiv preprint arXiv:2503.07598*, 2025. 1

- [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. [1](#)
- [4] Xiang Wang, Shiwei Zhang, Longxiang Tang, Yingya Zhang, Changxin Gao, Yuehuan Wang, and Nong Sang. Unianimate-dit: Human image animation with large-scale video diffusion transformer. *arXiv preprint arXiv:2504.11289*, 2025. [1](#)
- [5] Yang Zhang, Jiaxi Gu, Li-Wen Wang, Han Wang, Junqi Cheng, Yuefeng Zhu, and Fangyuan Zou. Mimicmotion: High-quality human motion video generation with confidence-aware pose guidance. *arXiv preprint arXiv:2406.19680*, 2024. [1](#)