HVG-3D: Bridging Real and Simulation Domains for 3D-Conditional Hand-Object Interaction Video Synthesis

Supplementary Material

This supplementary material provides additional details on our method and results that complement the main paper. Sec. A supplements the main text with details that are not covered elsewhere. Sec. B presents additional visualization results. Finally, Sec. C analyzes the limitations of our method and discusses directions for future work.

A. Additional Details.

A.1. Metrics

We conduct a quantitative analysis of the generated videos using the following metrics, which are evaluated based on the methods outlined in DanceTogether [1]. Because of the limited space in the manuscript, the implementation details of these metrics cannot be fully presented in Section 4.1 of the main text. We therefore provide additional explanations here.

- **L1↓** The L1 reconstruction error measures the average absolute difference between the predicted image or video and the ground-truth signal at the pixel level. We compute the per-pixel ℓ_1 distance over all frames:

$$\mathrm{L1} = \frac{1}{THWC} \sum_{t=1}^{T} \sum_{x=1}^{W} \sum_{y=1}^{H} \sum_{c=1}^{C} \left| I_t(x,y,c) - \hat{I}_t(x,y,c) \right|. \tag{1}$$

- **PSNR↑** PSNR is a signal fidelity measure that quantifies the ratio between the maximum possible pixel intensity and the mean squared reconstruction error in logarithmic (decibel) scale. For each frame t, we first compute the mean squared error

$$\mathrm{MSE} = \frac{1}{HWC} \sum_{x=1}^{W} \sum_{y=1}^{H} \sum_{c=1}^{C} \big(I_t(x,y,c) - \hat{I}_t(x,y,c) \big)^2, \tag{2}$$

and then

$$\mathrm{PSNR} = 20 \log_{10} \Big(\frac{255}{\sqrt{\mathrm{MSE}}} \Big). \tag{3}$$

- **SSIM↑** SSIM [8] assesses perceptual image quality by comparing local patterns of luminance, contrast, and structural information between a generated image and its reference. For each frame t and channel c,

$$\mathrm{SSIM}_t^c = \mathrm{SSIM}\big(I_t(\cdot,\cdot,c),\ \hat{I}_t(\cdot,\cdot,c) \big), \tag{4}$$

and the final score is

$$\mathrm{SSIM} = \frac{1}{TC} \sum_{t=1}^{T} \sum_{c=1}^{C} \mathrm{SSIM}_t^c. \tag{5}$$

- **LPIPS↓** LPIPS measures perceptual similarity by comparing deep feature representations extracted from pretrained neural networks for the generated and reference images. On a 720×480 crop, let $\phi_\ell(\cdot)$ be the ℓ-th layer feature map and w_ℓ the learned weights:

$$\mathrm{LPIPS} = \frac{1}{L} \sum_{\ell=1}^{L} \frac{1}{H_\ell W_\ell} \big\| w_\ell \odot \big(\phi_\ell(I) - \phi_\ell(\hat{I}) \big) \big\|_1. \tag{6}$$

- **CLIPScore↑** CLIPScore [2] evaluates semantic alignment between visual content and a corresponding text description by computing similarity in the joint embedding space of a pretrained CLIP model. We encode each frame from the ground-truth and generated videos into CLIP image embeddings $v_t, \hat{v}_t \in \mathbb{R}^d$, normalize them to unit vectors, and compute

$$s_t = \frac{v_t^\top \hat{v}_t}{\|v_t\| \cdot \|\hat{v}_t\|}. \tag{7}$$

The final CLIPScore averages over all T frames:

$$\mathrm{CLIPScore} = \frac{1}{T} \sum_{t=1}^{T} s_t. \tag{8}$$

- **ST-SSIM↑** ST-SSIM [5] extends the SSIM formulation from individual images to video sequences by jointly considering spatial structure and temporal evolution across frames. It measures how well local luminance, contrast, and structural patterns are preserved both within each frame and over time, thus capturing temporal consistency and motion coherence in addition to frame-wise quality. With window length $w = 3$, for each spatio-temporal block we compute

$$\mathrm{SSIM}_{3\mathrm{D}} = \mathrm{SSIM}\big(I_{t:t+w-1},\ \hat{I}_{t:t+w-1} \big), \tag{9}$$

and then

$$\text{ST-SSIM} = \frac{1}{T-w+1} \sum_{t=1}^{T-w+1} \mathrm{SSIM}_{3\mathrm{D}}. \tag{10}$$

- **GMSD-Temporal↓** GMSD-Temporal [9] is a video quality metric that evaluates discrepancies in gradient-based structural information across time. For $t = 2, \ldots, T$, let

$$g_t(x,y) = \big\| \nabla I_t(x,y) \big\|_2, \quad \hat{g}_t(x,y) = \big\| \nabla \hat{I}_t(x,y) \big\|_2, \tag{11}$$

and

$$\mathrm{GMS}_t(x,y) = \frac{2\, g_t\, \hat{g}_t + \varepsilon}{g_t^2 + \hat{g}_t^2 + \varepsilon}. \tag{12}$$

Then

$$\text{GMSD-Temporal} = \sqrt{\frac{1}{(T-1)HW} \sum_{t=2}^{T} \mathrm{Var}_{x,y}\big(\mathrm{GMS}_t(x,y) \big)}. \tag{13}$$

- **FID ↓** FID [3] measures the distributional discrepancy between real and generated images in a deep feature space, typically using activations from a pretrained Inception network. On all frames, we extract Inception-V3 features, form (μ_r, Σ_r) and (μ_f, Σ_f), and use

$$\mathrm{FID} = \|\mu_r - \mu_f\|_2^2 + \mathrm{Tr}\big(\Sigma_r + \Sigma_f - 2(\Sigma_r\Sigma_f)^{1/2}\big). \quad (14)$$

- **FVD ↓** FVD [6] extends the FID concept to videos by extracting spatio-temporal features from a pretrained video recognition network and comparing the distributions of real and generated sequences. From I3D features of each non-overlapping 16-frame clip, we compute means μ_r, μ_f and covariances Σ_r, Σ_f:

$$\mathrm{FVD} = \|\mu_r - \mu_f\|_2^2 + \mathrm{Tr}\big(\Sigma_r + \Sigma_f - 2(\Sigma_r\Sigma_f)^{1/2}\big). \quad (15)$$

- **CLIP-FID ↓** CLIP-FID is a distributional similarity metric that combines the Fréchet distance formulation with CLIP feature representations instead of Inception features. It evaluates how closely the semantic and high-level visual characteristics of generated images or videos match those of real data in a joint vision–language embedding space. Identical to FID but using CLIP embeddings instead of Inception features:

$$\mathrm{CLIP{-}FID} = \|\mu_r - \mu_f\|_2^2 + \mathrm{Tr}\big(\Sigma_r + \Sigma_f - 2(\Sigma_r\Sigma_f)^{1/2}\big). \quad (16)$$

A.2. Point cloud scanners

Our Point cloud scanner is built on top of VGGT [7]. We first apply VGGT to the entire video to obtain a per-frame reconstruction of the scene, resulting in a sequence of scene point clouds. In parallel, we obtain hand and object masks using SAMURAI [10], and use these masks to retain only the points within the masked regions. During the VGGT reconstruction process, we set the confidence threshold to 0.5. In addition, when scanning the scene, the total number of points is capped at $N_{max} = 50000$

A.3. Implement Details

We implement our model in PyTorch and train it on eight NVIDIA H20 GPUs with a batch size of (B = 4). We adopt the AdamW optimizer with a learning rate of 1×10^{-4} and train for 20 epochs. During training, we use the $learnable\_vec1024 \times 32\_dim1024\_depth24\_sdf\_nb$ configuration from 3DShape2VecSet [12] as the point cloud encoder. Before encoding, each point cloud is downsampled to (N = 8192) points. CogVideoX-5B-I2V [11] is employed as the base generative model. In the training process, the parameters of the denoising DiT are kept frozen, and only the parameters of the 3D point cloud ControlNet are updated.

In terms of dataset construction, Taste-rob [13] comprises a large collection of hand–object interaction videos recorded across diverse scenes. Specifically, we segment each original long video sequence into shorter clips of 49 frames at 20 FPS, and each clip contains a valid hand–object interaction episode. The training set consists of 50k such valid video clips. For evaluation, we first sample 2% of the data from each scene and then randomly select 100 hand–object interaction clips from this subset, covering different scenes, to form the test set.

B. More Evaluation Results

B.1. Qualitative Results for Comparisons

In Fig. S2 and Fig. S3, we present additional qualitative results that compare HVG-3D with existing state-of-the-art methods. For qualitative video comparisons, please refer to the supplementary video.

B.2. More Visualizations of Generated Videos

In Fig S4, we provide additional visualizations of hand–object interaction videos generated by our method. As shown, our model produces more plausible and coherent hand–object interaction motions.

B.3. Result on TACO Dataset

We evaluate on TACO dataset [4]. TACO is a large-scale real-world 4D bimanual hand-object interaction dataset that captures diverse daily tool-use activities under both egocentric and multi-view third-person settings. It provides rich annotations, including accurate 3D ground truth for hand-object interactions, and serves as a valuable benchmark for action understanding, motion prediction, and manipulation-oriented perception. As shown in Tab. S1, HVG-3D outperforms DaS on all metrics, demonstrating 3D geometric priors generalize to bimanual interactions. As shown in the Fig. S1, we can

Table S1. **Quantitative comparison between HVG-3D and DaS on TACO dataset**.

Method, Results in TACO	PSNR↑	SSIM↑	LPIPS↓	ST-SSIM↑	FID↓	FVD↓
DaS (Full Frame)	20.97	0.66	0.341	0.87	111.7	47.9
Ours (Full Frame)	**24.81**	**0.78**	**0.256**	**0.95**	**68.0**	**19.3**
DaS (Masked Region)	16.50	0.96	0.045	0.95	98.2	16.7
Ours (Masked Region)	**20.43**	**0.97**	**0.026**	**0.98**	**53.8**	**6.8**

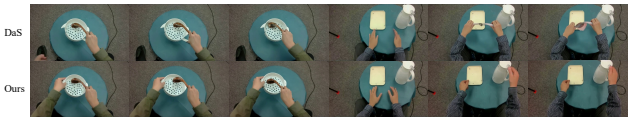

Figure S1. **Qualitative Comparison between HVG-3D and DaS on TACO Dataset**

C. Limitation and Future Work

Although HVG-3D can leverage 3D point cloud conditions to generate more plausible and high-quality hand–object interaction videos, as validated on the Taste-Rob dataset, sev-

eral key limitations remain and warrant further investigation in future work.

Limitation in application scenarios. At present, HVG-3D is primarily designed for hand–object interaction video generation. It has not yet been applied to generic video generation scenarios. This is mainly due to the lack of large-scale, high-quality 3D datasets for general scenes, which constrains the use of 3D conditions as control signals for generic video synthesis. Future work may explore how to effectively leverage 3D conditions to extend HVG-3D to more general scene settings.

Use of 3D point clouds as conditions. In HVG-3D, we adopt 3D point clouds as the primary conditioning signal. Compared with 3D meshes, point clouds inevitably lose part of the fine-grained geometric structure. Although some existing hand–object interaction datasets provide 3D mesh annotations, their scale remains limited, whereas 3D point clouds are considerably easier to obtain. In future work, we plan to investigate using 3D meshes as conditioning signals for video generation.

Future work Despite these limitations, HVG-3D provides a practical pathway for large-scale synthesis of 3D human–robot interaction data. Looking ahead, we aim to extend HVG-3D to more general scenarios and to employ more expressive 3D representations as conditioning signals for video generation. We believe that leveraging 3D conditions as control signals for video synthesis can open up new possibilities for future video content creation.

References

[1] Junhao Chen, Mingjin Chen, Jianjin Xu, Xiang Li, Junting Dong, Mingze Sun, Puhua Jiang, Hongxiang Li, Yuhang Yang, Hao Zhao, Xiao-Xiao Long, and Ruqi Huang. Dancetogether: Generating interactive multi-person video without identity drifting. In *The Fourteenth International Conference on Learning Representations*, 2026. 1

[2] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7514–7528, 2021. 1

[3] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. 2

[4] Yun Liu, Haolin Yang, Xu Si, Ling Liu, Zipeng Li, Yuxiang Zhang, Yebin Liu, and Li Yi. Taco: Benchmarking generalizable bimanual tool-action-object understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21740–21751, 2024. 2

[5] Anush K Moorthy and Alan C Bovik. Efficient motion weighted spatio-temporal video ssim index. In *Human Vision and Electronic Imaging XV*, pages 440–448. SPIE, 2010. 1

[6] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Fvd: A new metric for video generation. 2019. 2

[7] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5294–5306, 2025. 2

[8] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004. 1

[9] Peng Yan, Xuanqin Mou, and Wufeng Xue. Video quality assessment via gradient magnitude similarity deviation of spatial and spatiotemporal slices. In *Mobile Devices and Multimedia: Enabling Technologies, Algorithms, and Applications 2015*, pages 182–191. SPIE, 2015. 1

[10] Cheng-Yen Yang, Hsiang-Wei Huang, Wenhao Chai, Zhongyu Jiang, and Jenq-Neng Hwang. Samurai: Adapting segment anything model for zero-shot visual tracking with motion-aware memory. *arXiv preprint arXiv:2411.11922*, 2024. 2

[11] 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. 2

[12] Biao Zhang, Jiapeng Tang, Matthias Niessner, and Peter Wonka. 3dshape2vecset: A 3d shape representation for neural fields and generative diffusion models. *ACM Transactions On Graphics (TOG)*, 42(4):1–16, 2023. 2

[13] Hongxiang Zhao, Xingchen Liu, Mutian Xu, Yiming Hao, Weikai Chen, and Xiaoguang Han. Taste-rob: Advancing video generation of task-oriented hand-object interaction for generalizable robotic manipulation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 27683–27693, 2025. 2

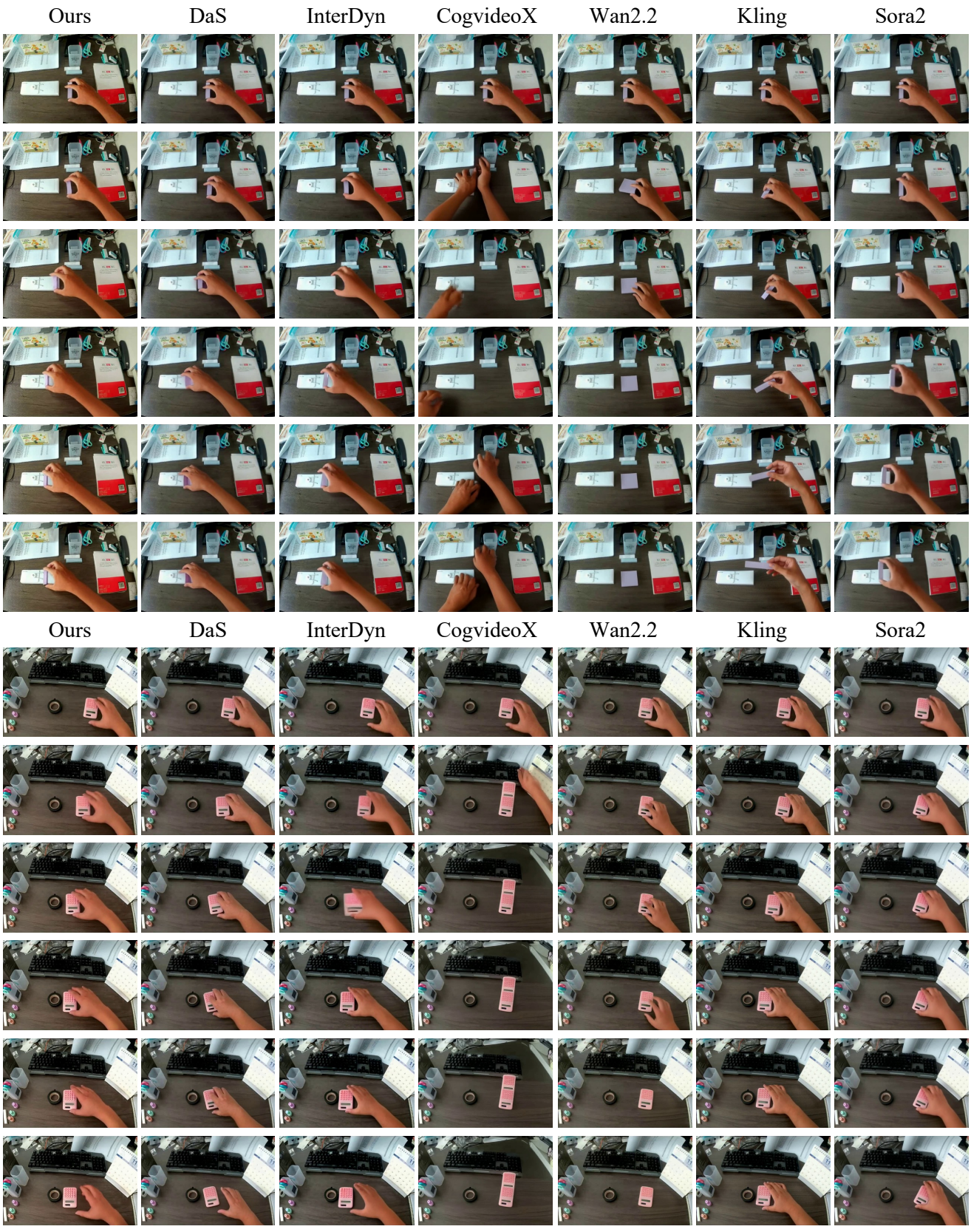

Figure S2. Qualitative Comparison of DaS, InterDyn, CogvideoX, Wan2.2, Kling, Sora2 and Our Method. The first frame of each baseline serves as the reference image. Zoom in for more details.

Figure S3. Qualitative Comparison of DaS, InterDyn, CogvideoX, Wan2.2, Kling, Sora2 and Our Method. The first frame of each baseline serves as the reference image. Zoom in for more details.

Figure S4. More Visualizations of Generated Videos