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

Mingjin Chen[1 *] Junhao Chen[3 *] Zhaoxin Fan[2 †] Yujian Lee[4] Zichen Dang[1]
Lili Wang[5] Yawen Cui[1] Lap-Pui Chau[1] Yi Wang[1 †]

[1]Dept. of EEE, The Hong Kong Polytechnic University [2]Beijing Advanced Innovation Center for Future Blockchain and Privacy Computing, School of Artificial Intelligence, Beihang University
[3]Tsinghua University [4]Beijing Normal-Hong Kong Baptist University [5]State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University

Project Page: https://hvg3d.github.io

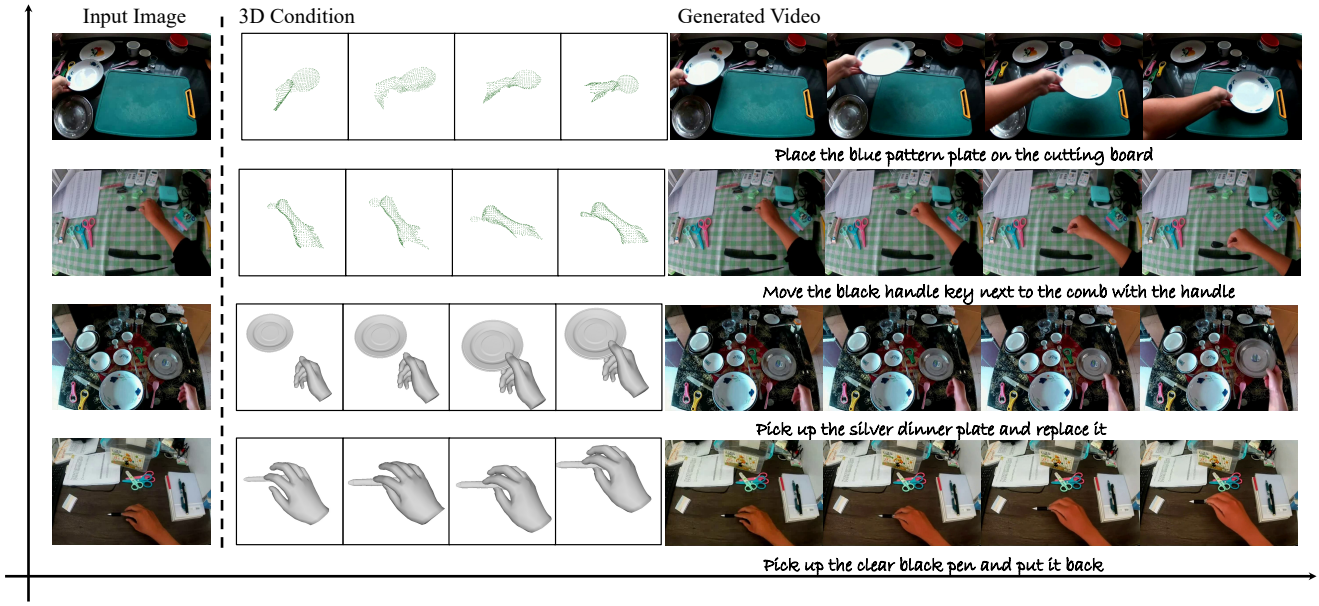

Figure 1. Illustration of 3D-conditioned hand–object interaction video generation with our proposed **HVG-3D** framework. HVG-3D synthesizes realistic and temporally coherent hand–object interaction videos by conditioning on explicit 3D signals. The top two rows display generated results using 3D point cloud and pose conditions extracted from real-world egocentric videos. The bottom two rows show results where 3D conditions are obtained from simulated hand–object sequences, demonstrating the framework's flexibility in accepting both real and synthetic 3D inputs. For each example, the leftmost column shows the input image and 3D condition, while subsequent columns depict selected frames from the generated video.

Abstract

Recent methods have made notable progress in the visual quality of hand-object interaction video synthesis. However, most approaches rely on 2D control signals that lack spatial expressiveness and limit the utilization of synthetic 3D conditional data. To address these limitations, we propose HVG-3D, a unified framework for 3D-aware hand-object interaction (HOI) video synthesis conditioned on explicit 3D representations. Specifically, we develop a diffusion-based architecture augmented with a 3D ControlNet, which encodes geometric and motion cues from 3D inputs to enable explicit 3D reasoning during video synthesis. To achieve high-quality synthesis, HVG-3D is designed with two core components: (i) a 3D-aware HOI video generation diffusion architecture that encodes geometric and mo-

* Equal Contribution † Corresponding Author.

tion cues from 3D inputs for explicit 3D reasoning; and (ii) a hybrid pipeline for constructing input and condition signals, enabling flexible and precise control during both training and inference. During inference, given a single real image and a 3D control signal from either simulation or real data, HVG-3D generates high-fidelity, temporally consistent videos with precise spatial and temporal control. Experiments on the TASTE-Rob dataset demonstrate that HVG-3D achieves state-of-the-art spatial fidelity, temporal coherence, and controllability, while enabling effective utilization of both real and simulated data.

1. Introduction

Recent breakthroughs in diffusion-based generative models have fundamentally advanced the field of video synthesis, with large-scale models such as Sora [6], CogVideo-X [76], Keling [32], Hunyuan Video [31], and Veo 3 [12] setting new standards for generating high-quality and temporally consistent videos. Leveraging the capabilities of these foundational models, a growing body of work has focused on the generation of hand-object interaction videos, which has garnered increasing interest for applications in training robotic grasping models [3, 5, 13, 14, 24, 43, 44, 66, 85, 88, 89].

However, while recent methods for hand-object interaction video generation [1, 11, 50, 58, 81, 84] have demonstrated impressive visual quality, their reliance on 2D conditioning signals remains a fundamental bottleneck. In particular, widely adopted controls, such as point trajectories [70, 86], optical flow [34, 39, 54, 79, 83], bounding boxes [25, 48, 53, 63], and masks [1, 9, 60, 65], are inherently limited in spatial expressiveness and temporal consistency. This absence of true 3D conditioning introduces two critical challenges: (1) Imperfect 3D Understanding: 2D signals provide only partial motion and geometry cues, frequently resulting in unrealistic deformations and physically implausible hand-object interactions; (2) High Data Cost: These 2D conditions are typically extracted from real-world videos, making it difficult to exploit synthetic data generated by efficient simulators, and thus substantially increasing the cost of data collection and annotation.

To address the issue, recent work such as Diffusion as Shader (DaS) [17] has begun to incorporate 3D tracking videos for richer motion guidance. Nevertheless, these 3D cues are ultimately projected into 2D video sequences for model input, which prevents full utilization of the spatial structure and depth relations intrinsic to 3D space. To overcome these limitations, it is essential to design methods that can intrinsically exploit 3D conditioning, thereby improving the realism and physical plausibility of hand-object interactions, as well as facilitating scalable data generation using simulators.

To this end, we present HVG-3D, a unified framework that enables 3D-aware synthesis of hand-object interaction videos. Our key insight is to bridge the gap between visual realism and precise physical control by conditioning video generation on explicit 3D representations. Given a single real-world RGB image as appearance input and a 3D condition derived from a simulator (or collected from another real videos), HVG-3D is capable of generating high-fidelity, temporally consistent hand-object interaction videos. Specifically, the HVG-3D framework is composed of two key components. First, a 3D-aware HOI video generation diffusion architecture leverages a dedicated 3D ControlNet to encode geometric and motion cues from 3D point clouds or tracking sequences, injecting these features into a diffusion transformer via zero-initialized convolutional layers for explicit 3D reasoning. Second, a hybrid pipeline constructs input and condition signals by pairing real images with 3D conditions from simulation or another videos, supporting flexible and precise control throughout both training and inference. Both design together enable HVG-3D to generate realistic, 3D-consistent hand-object interaction videos from a single real image and a 3D control signal, as inllsurated in Fig. 1.

Extensive experiments on the TASTE-Rob dataset demonstrate that HVG-3D significantly outperforms state-of-the-art methods across multiple metrics, achieving superior spatial fidelity, temporal coherence, and controllability, highlighting its practical value for scalable and controllable video generation. Our contributions can be summarized as:

- We introduce a practical paradigm for hand-object interaction video generation that bridges real and simulated domains, enabling synthesis from a real input image and a 3D condition obtained from either simulation or another real video.
- We present HVG-3D, a unified framework featuring a 3D-aware diffusion-based architecture and a hybrid pipeline for constructing input and condition signals, achieving flexible and precise control.
- We validate our approach with comprehensive experiments, demonstrating state-of-the-art performance and effective integration of real and simulated data for scalable, controllable video generation.

2. Related Works

2.1. Controllable Video Generation

Controllable video generation leverages diffusion models pretrained on large-scale video datasets [4, 22, 23, 40, 55] to synthesize videos under user-specified constraints. Spatial control methods, such as ControlNeXt [52] and MimicMotion [82], use masks and keypoints to guide object appearance and pose, while Champ [87] incorporates optical flow for motion control. Temporal control is addressed by Tora [83], CameraCtrl [19], and MOFA-Video [49], which

introduce trajectory or camera motion cues for dynamic regulation. More recent efforts [9, 36, 37, 42, 68, 71] attempt to combine spatial and temporal signals for finer-grained control, extending to multi-person interactive scenarios with identity preservation. Meanwhile, diffusion-based approaches have also been applied to articulated character animation [56, 59] and temporally consistent human-centric dense prediction [29, 45, 57], further broadening the scope of controllable generation. Despite the effectiveness of existing methods, they operate primarily on 2D representations, limiting their ability to capture complex 3D geometry and hindering realistic and scalable video synthesis. In this work, we introduce explicit 3D conditioning into video diffusion models while focusing on the specific hand-object interaction video generation task.

2.2. Hand-Object Interaction Video Generation

Hand-object interaction video generation encompasses both 3D reconstruction and 2D synthesis methods. 3D approaches, such as ARCTIC [15], HOLD [16], ObMan [18], and HOIDiffusion [81], focus on reconstructing or generating 3D hand-object poses, while Text2HOI [7] enables text-driven 3D motion synthesis. However, these methods typically operate on isolated objects without incorporating full scene context, which limits their applicability to real-world scenarios. In the 2D domain, CosHand [58] synthesizes static HOI images, and InterDyn [1] as well as TASTE-Rob [84] extend generation to videos. Despite recent progress, 2D HOI generation still suffers from limited visual quality and physical inconsistency, which reduces its usefulness for downstream tasks such as robotic policy learning. We therefore study 3D-conditioned HOI generation to improve both visual fidelity and physical plausibility.

2.3. 3D Representation and Rendering

3D rendering methods provide the foundation for synthesizing visual content from geometric data. Traditional graphics pipelines render meshes or point clouds via rasterization [78], offering precise geometric control but requiring extensive manual asset preparation. Neural rendering techniques, such as NeRF [46] and 3D Gaussian Splatting [28], learn implicit or explicit 3D representations from multi-view images, enabling photorealistic novel view synthesis at the cost of scene-specific optimization. Recent works have also explored efficient 3D asset creation from diverse input modalities, including single-image reconstruction [10, 35, 38, 41, 77], interleaved multimodal 3D generation [8, 69, 72]. In controllable image and video generation, most approaches convert 3D information into 2D rendered maps, such as depth [26, 33, 51, 52], surface normals [71], or pose visualizations [36, 65, 82], to serve as conditioning signals for 2D diffusion models. While this rendering-then-generation paradigm is effective, it inevitably incurs information loss and struggles to capture complex 3D spatial relationships, especially under occlusion. In contrast, we directly incorporate 3D point clouds as conditioning signals, allowing the diffusion model to operate as a neural renderer and maintain full 3D structural information throughout generation.

3. Methods

3.1. Overview

We consider the task of 3D-conditioned hand-object interaction image-to-video (I2V) generation. Given a single input image $I_0 \in \mathbb{R}^{H\times W\times 3}$, a sequence of 3D point clouds $P = \{P_t\}_{t=1}^{T}$, $P_t \in \mathbb{R}^{N\times 3}$ representing hand-object geometry over T frames, and an optional 3D tracking sequence $\mathcal{T} = \{T_t\}_{t=1}^{T}$, the goal is to generate a video $V_{out} = \{I_t\}_{t=1}^{T}$, $I_t \in \mathbb{R}^{H\times W\times 3}$ that is visually realistic, temporally coherent, and faithfully respects the 3D spatial constraints.

To address this problem, as shown in Fig. 2, our proposed HVG-3D framework consists of two main components: (i) a 3D-aware diffusion-based architecture that encodes geometric and motion cues from the 3D point cloud and tracking sequence, enabling explicit 3D reasoning; and (ii) a hybrid pipeline for constructing input and condition signals, which flexibly integrates both real and simulated data to provide precise spatial and temporal control for both training and inference.

In the following sections, we first detail the 3D-aware diffusion-based architecture (Section 3.2), which builds upon a pretrained image-to-video diffusion backbone and incorporates 3D control signals. We then introduce the hybrid pipeline for constructing and aligning input and condition signals (Section 3.3), including the acquisition and processing of 3D cues in both real and synthetic settings.

3.2. 3D-aware HOI Diffusion Architecture

While recent diffusion-based image-to-video models achieve impressive visual quality, their reliance on 2D conditions limits spatial consistency and geometric fidelity, especially for hand–object interactions. To address this, we propose a 3D-aware diffusion framework that explicitly incorporates 3D structural and motion cues into the generation process. Our architecture consists of a strong image-to-video diffusion backbone and a dedicated 3D point cloud–guided ControlNet, as described below.

Base Image-to-Video Model. Our framework adopts CogVideoX-5B-I2V [76] as the image-to-video generation backbone. CogVideoX-5B-I2V is a Transformer-based video diffusion model with 3D full attention, enabling high-fidelity and temporally consistent synthesis. As shown in Fig.2, the model takes an input image $I_0 \in \mathbb{R}^{H\times W\times 3}$ and a ground-truth video $V_{gt} \in \mathbb{R}^{T\times H\times W\times 3}$, which are encoded

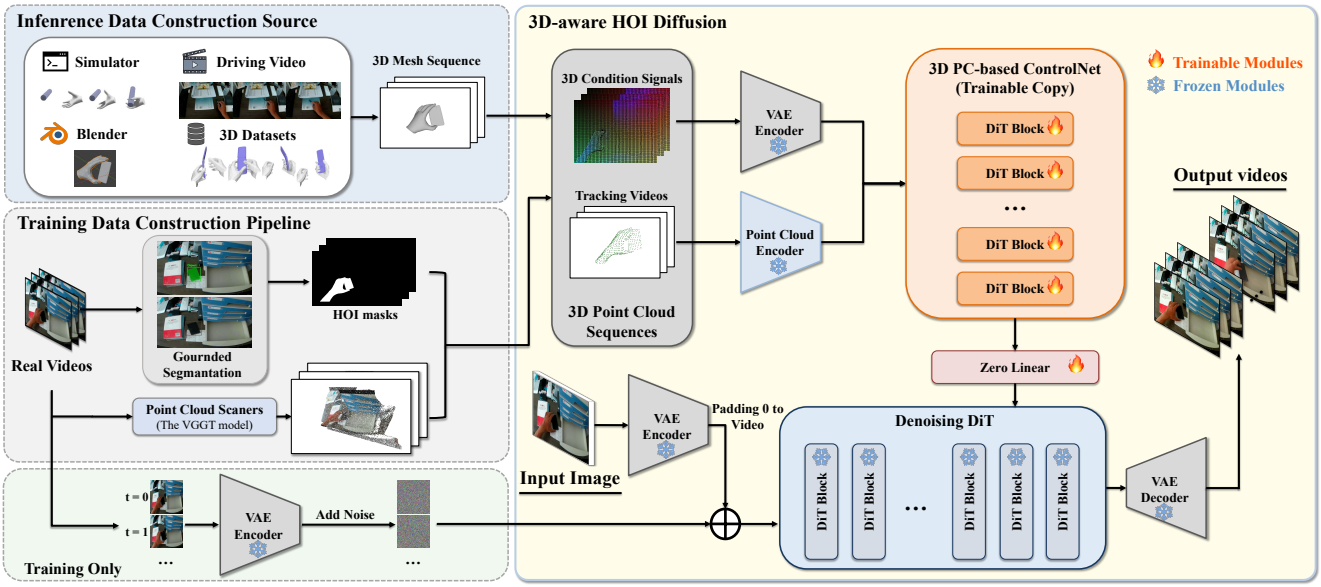

Figure 2. **Architecture of HVG-3D**. The left panel illustrates the hybrid training and inference pipeline, where egocentric driving videos, simulator outputs, and 3D HOI datasets are processed by grounded segmentation, key bounding-box extraction and a point-cloud scanner to construct paired input images, 3D tracking videos, and 3D point cloud sequences. The right panel depicts the 3D-aware HOI video generation diffusion architecture, in which the 3D point cloud and tracking signals are encoded by a trainable 3D ControlNet and injected into a frozen image-to-video diffusion backbone via zero-initialized layers, enabling the synthesis of temporally coherent videos that respect the underlying 3D hand–object interaction geometry.

into latent representations Z_{I_0} and $Z_{gt} \in \mathbb{R}^{T\times\frac{H}{8}\times\frac{W}{8}\times 16}$ via a VAE encoder[30]. The image latent Z_{I_0} is temporally zero-padded to match the length T and concatenated with a noised version of Z_{gt}. The resulting joint latent sequence is processed by a Diffusion Transformer, which performs iterative denoising to recover the clean video latent Z_ε. Finally, a 3D VAE decoder reconstructs the output video V_{out} from Z_ε. Such a diffusion-based architecture provides robust temporal modeling and fine-grained control over video dynamics, facilitating the generation of realistic and structurally consistent hand–object interaction sequences. Its unified latent space enables effective encoding of both appearance and motion, supporting generalization to diverse HOI scenarios.

3D Point Cloud-based ControlNet. To provide explicit 3D structural and motion guidance, we introduce 3D point clouds as conditioning signals within our diffusion framework. Given a sequence of point clouds $P \in \mathbb{R}^{T\times N\times 3}$ extracted from the input video, we employ a point cloud encoder [80] to obtain latent features $Z_{pc} \in \mathbb{R}^{T\times L\times 768}$, where N denotes the number of points and L the number of latent tokens. In parallel, 3D tracking information is encoded as $Z_{tracking} \in \mathbb{R}^{T\times\frac{H}{8}\times\frac{W}{8}\times 16}$.

To ensure compatibility among heterogeneous condition signals, we project Z_{pc} via a learnable linear layer and resample to match the dimensionality of $Z_{tracking}$ and Z_{gt}. The aligned latents are concatenated and serve as input to the 3D Point Cloud ControlNet. Architecturally, the ControlNet is constructed by replicating all pretrained DiT blocks, which are then specialized to encode the 3D conditioning. At each layer, the output of the ControlNet is modulated by a zero-initialized convolution and injected into the corresponding DiT block of the main diffusion backbone. By injecting 3D structural and motion cues at every denoising step, our design significantly improves spatial consistency and physical plausibility in synthesized hand–object interactions, especially under challenging scenarios such as severe occlusions and complex articulations.

3.3. Hybrid Pipeline for Input and Condition Signal Construction

Our hybrid pipeline is designed to flexibly support both real and synthetic 3D conditioning signals throughout training and inference. It consists of three stages: training data construction, model training, and inference and practical conditioning. Next, we detail each stages.

Training Data Construction. To address the lack of explicit mask and 3D point cloud annotations in TasteRob [84], we devise a data pipeline for recovering these signals from monocular egocentric RGB videos of hand-object interaction. Object and hand bounding boxes are extracted by combining inter-frame difference maps (for static backgrounds) and YOLOv8-X [27] detection (for dynamic hand regions). Instance masks for both hand and object

are generated using SAMURAI [75], with tracking initialized via bounding boxes and bidirectional refinement to ensure temporal consistency. The masks are fused to obtain a per-frame hand-object segmentation. 3D point cloud reconstruction is performed with VGGT [64]: given the video frames and their corresponding masks, VGGT produces a per-frame point cloud $P \in \mathbb{R}^{T\times N\times 3}$ representing the 3D geometry of the hand and object.

Model Training. With the training data constructed as described above, we proceed to optimize our 3D-aware diffusion model for hand-object interaction video generation. While explicit 3D conditioning provides strong geometric control, it may not fully suppress background distractions, especially in cluttered scenes. To address this, we augment the standard diffusion loss with a mask-based reconstruction term inspired by StableAnimator [60], which focuses learning on the regions of interest.

The final training objective is defined as:

$$L = \sum_{i=1}^{n} \mathbb{E}\varepsilon \left(\left| (Z_{gt} - Z_{\varepsilon}) \odot (1 + M^i) \right|^2 \right) \quad (1)$$

where Z_{gt} and Z_{ε} denote the ground-truth and predicted video latents respectively, and $M^i \in \{0,1\}^{1\times H\times W}$ is the hand–object mask for frame i. This loss formulation ensures that errors in the hand-object regions are emphasized, promoting accurate reconstruction of interaction-critical areas while mitigating the influence of background noise.

For training, each video is center-cropped and resized to 720×480, with a fixed length of 49 frames. Each training sample comprises: the input image I_0, ground-truth video V_{gt}, the hand–object mask sequence, 3D point cloud sequence P, and 3D tracking sequence (estimated via SpatialTracker [73]). We fine-tune only the copied condition DiT blocks, keeping all parameters of the original denoising DiT backbone frozen to preserve pre-learned video generation capabilities. Training is performed using the AdamW optimizer with a learning rate of 1×10^{-4} for 20 epochs, employing gradient accumulation to achieve an effective batch size of 4. All experiments are conducted on 8 H20 GPUs.

Inference and Practical Conditioning. At inference time, the model takes a single real input image I_0 together with a 3D conditioning signal. The 3D condition can be: (1)extracted from real video using the same pipeline as in training (detection, segmentation, and VGGT point cloud reconstruction); or (2) synthesized in simulation, e.g., by generating 3D hand-object sequences in Blender or other simulators, or by sampling from 3D HOI datasets such as ARCTIC [15] or HOT3D [2]. All 3D mesh sequences are processed to produce compatible point clouds and, if needed, tracking sequences, ensuring seamless integration with the model's conditioning interface.

4. Experiment

4.1. Experiment Setting

Datasets We train our HVG-3D on Taste-Rob [84]. This datasset is an egocentric hand object dataset. This dataset comprises egocentric videos of hand–object interactions collected across multiple scenes. For training, we selected the Single Hand subset and used samples with scene labels office, dining, bedroom, kitchen, dressing table. We crop all videos to a resolution of 720×480 and segment the original videos into clips of 49 frames each. In the evaluation stage, we first sampled 2% of the videos from each scene category as candidate test samples. For the Taste-Rob evaluation, we then randomly selected 100 videos from these candidates to construct our final test set. All evaluation metrics are reported based on this test set.

Metrics. We evaluate the performance of our model from two complementary aspects: *image quality* and *spatio-temporal similarity*. **Image quality.** To assess the fidelity of the generated frames, we adopt several commonly used image-based metrics, including L1, Peak Signal-to-Noise Ratio (PSNR), the Structural Similarity Index Measure (SSIM) [67], Learned Perceptual Image Patch Similarity (LPIPS), CLIP Score [20], Fréchet Inception Distance (FID) [21], and CLIP-FID. These metrics comprehensively measure the perceptual and structural consistency between generated and ground-truth frames. **Spatio-temporal similarity.** To further evaluate the overall video quality across both spatial and temporal dimensions, we measure the perceptual similarity between the generated and real video distributions using the Fréchet Video Distance (FVD) [61], Spatio-Temporal SSIM(ST-SSIM) [47] and Gradient Magnitude Similarity Deviation – Temporal(GMSD-T) [74]. These metrics quantify the coherence and realism of motion dynamics in the generated videos.

4.2. Baseline Comparisons

To ensure a comprehensive and fair comparison, we select three state-of-the-art video generation models, namely CogVideoX [76], Wan 2.2 [62], Kling [32] and DaS [17]. In addition, we compare our method with a specialized approache for hand–object interaction video generation, namly, InterDyn [1].

Quantitative comparison. To compare HVG-3D with existing methods, we divide each video in the test set into 49-frame clips and randomly select one clip containing hand–object interaction from each video for evaluation. For each baseline method, we extract the required conditional inputs from its corresponding video.

As shown in Tab. 1 and Tab 2, Tab. 1 reports the full-frame performance metrics, whereas Tab 2 presents the mask-aware metrics within the hand–object region. Under full-frame evaluation, benefiting from the precise con-

Table 1. **Quantitative comparison between HVG-3D and baselines on Full Frame evaluation metrics**. Most video generation metrics demonstrate that HVG-3D achieves superior performance.

Method	L1↓	PSNR↑	SSIM↑	LPIPS↓	CLIP↑	ST-SSIM↑	GMSD-T↓	FID↓	C-FID↓
Kling [32]	17.42	19.22	0.66	0.316	0.95	0.85	0.44	98.9	18.5
Wan2.2 [62]	14.26	20.76	0.77	0.25	0.95	0.87	0.45	122.6	18.1
CogVideoX [76]	22.85	17.10	0.67	0.334	0.93	0.77	0.45	174.2	27.6
InterDyn [1]	8.81	24.13	0.82	0.205	0.95	0.95	0.45	73.3	17.8
DAS [17]	**7.77**	**24.83**	**0.84**	**0.191**	**0.96**	0.96	0.44	75.5	**14.6**
Our	9.50	24.15	0.81	**0.193**	**0.96**	**0.97**	**0.40**	**58.2**	**14.6**

Table 2. **Quantitative comparison between HVG-3D and baselines on Hand Object Masked Region evaluation metrics**. All video generation metrics consistently indicate that HVG-3D delivers superior performance.

Method	L1↓	PSNR↑	SSIM↑	LPIPS↓	CLIP↑	ST-SSIM↑	GMSD-T↓	FID↓	C-FID↓
Kling [32]	41.86	14.74	0.95	0.055	0.94	0.77	0.17	182.4	24.7
Wan2.2 [62]	48.36	13.77	0.95	0.061	0.94	0.72	0.18	217.1	27.1
CogVideoX [76]	58.78	11.83	0.94	0.068	0.92	0.62	0.20	260.9	36.9
InterDyn [1]	20.99	19.03	0.96	0.034	0.96	0.92	0.16	104.5	14.5
DAS [17]	26.55	17.41	0.96	0.039	0.96	0.88	0.16	128.0	16.0
Our	**20.90**	**19.08**	**0.97**	**0.032**	**0.97**	**0.93**	**0.15**	**88.5**	**13.1**

trol provided by the 3D condition, our method achieves the lowest FVD (13.8) and FID (58.2), while also obtaining the highest CLIP Score (0.96) and GMSD-T (0.40). Although some full-frame metrics are slightly inferior to those of DaS, our method attains the best performance on all metrics within the hand–object mask region, which corresponds to the primary interaction area. In particular, FVD is reduced from 13.8 to 9.6 as shown in the Fig. 4, and C-FID decreases from 14.6 to 13.1, even as other methods exhibit a consistent degradation across all metrics in this region. Notably, these improvements are achieved without sacrificing low-level reconstruction fidelity. L1 and LPIPS are simultaneously reduced, whereas PSNR and SSIM are improved.

Qualitative comparison. The qualitative comparison is presented in Fig. 3. In the first case, we illustrate the process of unfolding a folded three-line checkered sheet. In the second case, we show a plate containing shrimp and scallions being moved to the left side of a plate with chicken breast. In the third case, we demonstrate placing a three-tone eyeshadow palette at the upper-left corner of a dressing table. In the fourth case, we depict placing a stapler on top of a blue book. For each example, we provide the input image, the text prompt, and the method-specific conditions (the tracking video for DaS and the mask for InterDyn). Moreover, since the output resolution of Sora2 is not aligned with that of the other baselines, a fair quantitative comparison is not feasible. We therefore only present qualitative results in this section.

As shown in Fig. 3, only our method, HVG-3D, successfully executes the specified manipulation while preserving the shape of both the hand and the object. DaS performs reasonably well for in-plane parallel translations of the object, but once the task involves folding or motion perpendicular to the tabletop, it tends to introduce noticeable object deformation. InterDyn exhibits similar issues and is even less stable than DaS. In contrast, the recent powerful video generation models CogVideoX, Wan2.2, Kling, and Sora2 all struggle to reliably accomplish the required manipulation.

Beyond the aforementioned qualitative results, we further demonstrate the flexibility of HVG-3D in handling conditioning sources at inference time. As shown in Fig. 2, our framework not only accepts 3D point clouds scanned from videos, but also ingests 3D conditions derived from diverse pipelines, including physics-based simulators, real driving videos with 3D reconstruction, and pre-captured hand–object interaction datasets. In practical deployments, HVG-3D can be driven by a broad range of 3D inputs, including 3D mesh sequences edited in Blender to create novel hand–object motions, 3D trajectories or point clouds estimated from driving videos through standard reconstruction pipelines, ready-made 3D conditions from existing hand–object interaction datasets, and synthetic 3D hand–object interaction sequences rendered directly by simulators. As shown in the last two rows of Fig. 1, we further demonstrate that editing 3D mesh sequences in Blender enables the generation of new hand–object interaction videos. This unified interface for heterogeneous 3D inputs underscores the generality of our framework and supports seam-

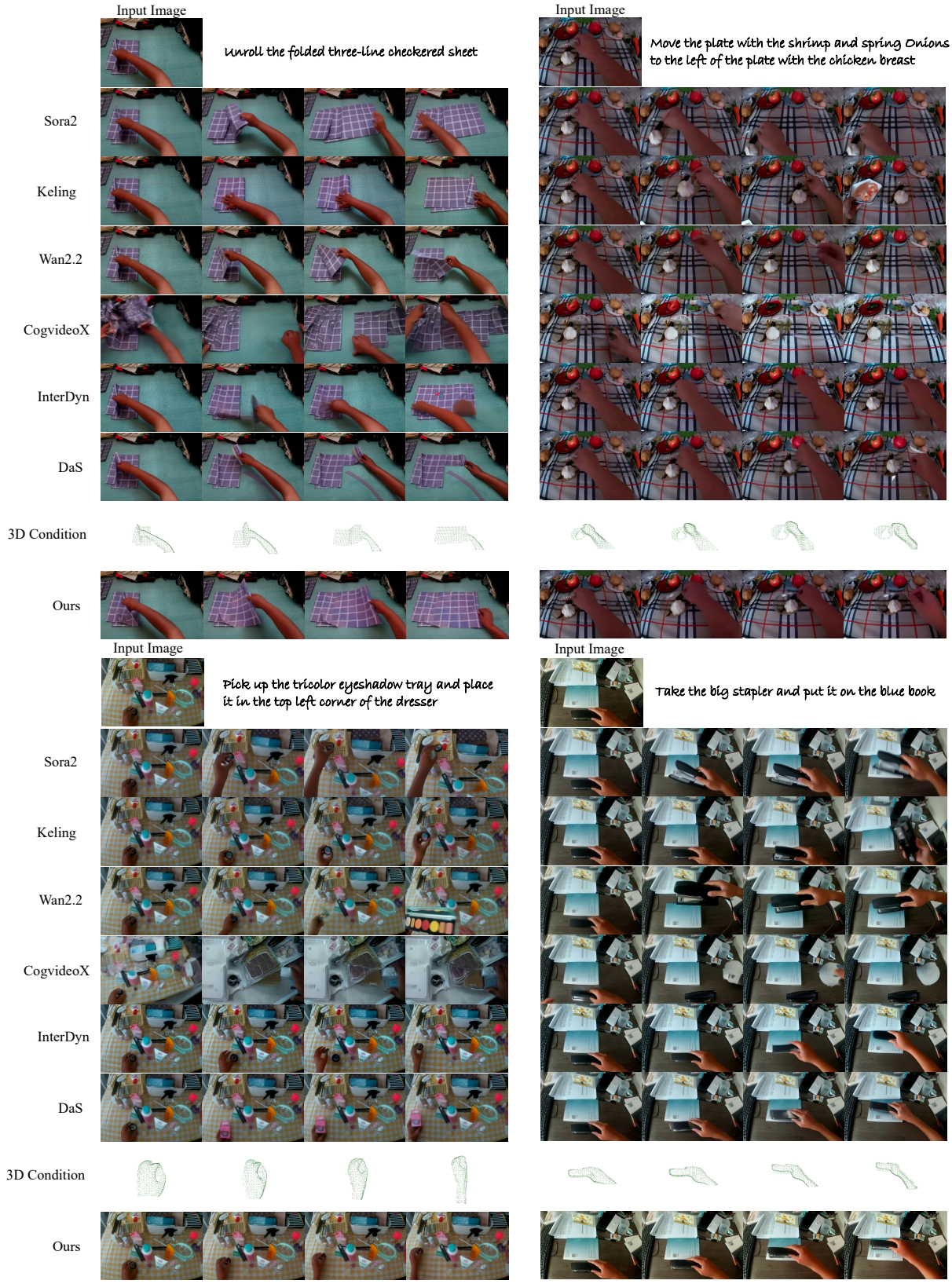

Figure 3. **Qualitative comparison of video generation performance**. HVG-3D is capable of generating videos with highly accurate motions and superior visual quality, while further ensuring that both the hand and the object remain free from geometric deformation. A level of performance that current state-of-the-art general-purpose video generation models are unable to achieve.

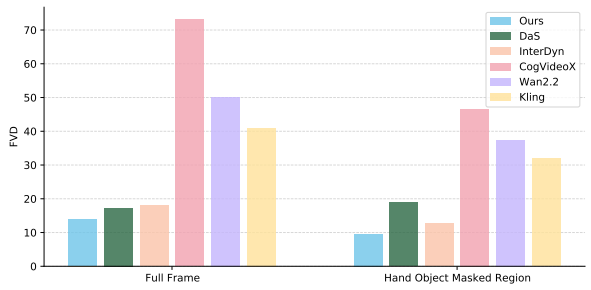

Figure 4. **Qualitative comparison between HVG-3D and baselines on FVD**. Our method achieves the best FVD scores in both the full-frame setting and the hand–object masked region.

less deployment across diverse 3D acquisition pipelines and downstream application scenarios.

4.3. Ablation Study

In this section, we present ablation studies to validate the effectiveness of our 3D point-cloud conditioning. Our ablation studies further demonstrate that combining the 3D tracking video with the proposed mask diffusion loss not only improves manipulation accuracy in human–object interaction scenarios and better preserves the shapes of the hand and the object, but also accelerates training and enables faster model convergence.

Ablations on 3D point cloud condition 3D point-cloud conditioning equips the model with more accurate 3D perception during training, thereby strengthening depth reasoning. To assess its contribution, we remove this condition and train the model for the same number of epochs as the full system, then compare the results. As shown in Tab. 3, omitting the 3D point cloud condition degrades the quality of hand–object interactions in the generated videos. The deterioration of these metrics further reflects that the absence of the 3D point cloud condition leads to noticeable shape distortions of the hand and object during interaction. Moreover, similar to the phenomenon observed in DaS, when the object needs to be folded or undergoes vertical motion perpendicular to the tabletop, meaning that the model must demonstrate a certain degree of depth awareness, the quality of the generated videos decreases substantially. These factors collectively contribute to the lower evaluation scores.

Ablations on 3D tracking video 3D tracking videos provide the model with more accurate viewpoint awareness and, when combined with 3D point-cloud conditioning, enhance control over hand–object interactions. To evaluate their effectiveness, we ablate the 3D tracking video condition, train for the same number of epochs as the full model, and then compare results. As shown in Tab. 3, the metrics indicate that removing the 3D tracking video leads to a degradation in the quality of hand–object interactions in the generated sequences. In the absence of 3D tracking video, the object shape may remain plausible, but the spatial alignment of the interaction (e.g., contact locations and motion trajectories) becomes less accurate. Moreover, since the model is trained with the 3D point cloud condition but without the 3D tracking video, the resulting performance drop suggests that the 3D tracking video encodes complementary camera-view information that helps the model learn more effective point cloud representations.

Table 3. Ablation Studies on 3D point cloud, 3D tracking video and mask diffusion loss. The experimental results demonstrate that these techniques enhance the quality of hand–object interaction video generation, improve the accuracy of the synthesized interaction process, and accelerate convergence during training.

Method	PSNR	SSIM	LPIPS	ST-SSIM
w/o 3D pc	18.44	0.37	0.5957	0.91
w/o 3D tracking	22.76	0.75	0.2054	0.95
w/o mask loss	22.09	0.80	0.199	0.96
full model	**24.15**	**0.81**	**0.193**	**0.97**

Ablations on mask diffusion loss Mask diffusion loss encourages the model to focus on hand–object interaction regions during training, thereby improving convergence. To assess its effectiveness, we remove the mask component from the diffusion loss and train for the same number of epochs as the full model, then compare the results. As shown in Tab. 3, removing the mask diffusion loss leads to a degradation in overall video quality. This occurs because, once the mask is excluded from the diffusion loss, the model tends to focus on the entire scene during training rather than emphasizing the hand–object interaction region. Under the same training configuration as the full model, the metrics obtained at the same number of epochs are consistently worse, indicating that explicitly guiding the model to focus on hand–object interactions enables faster training and more rapid convergence. At the same time, the model becomes more robust to distractions from other objects in complex scenes, resulting in more accurate generation of hand–object interactions.

5. Conclusion

We presented HVG-3D, a unified framework for 3D-conditioned hand–object interaction video generation. By incorporating a 3D ControlNet that encodes point cloud and tracking cues into a video diffusion backbone, together with a hybrid pipeline bridging real and simulated domains, HVG-3D achieves state-of-the-art spatial fidelity, temporal coherence, and controllability on the TASTE-Rob benchmark. Ablation studies confirm the complementary benefits of each component. Future work will extend to more diverse interaction scenarios, longer sequences, and closed-loop integration with robotic manipulation policies.

6. Acknowledgements

This work was supported by the New Generation Artificial Intelligence-National Science and Technology Major Project (2025ZD0122603). It was also supported by the Postdoctoral Fellowship Program and China Postdoctoral Science Foundation under Grant No. 2024M764093 and Grant No. BX20250485, the Beijing Natural Science Foundation under Grant No. 4254100, and by Beijing Advanced Innovation Center for Future Blockchain and Privacy Computing. It was also supported by the Young Elite Scientists Sponsorship Program of the Beijing High Innovation Plan (NO. 20250860).

The research work described in this paper was conducted in the JC STEM Lab of Machine Learning and Computer Vision funded by The Hong Kong Jockey Club Charities Trust. This research received partially support from the Global STEM Professorship Scheme from the Hong Kong Special Administrative Region.

References

[1] Rick Akkerman, Haiwen Feng, Michael J Black, Dimitrios Tzionas, and Victoria Fernández Abrevaya. Interdyn: Controllable interactive dynamics with video diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12467–12479, 2025. 2, 3, 5, 6

[2] Prithviraj Banerjee, Sindi Shkodrani, Pierre Moulon, Shreyas Hampali, Fan Zhang, Jade Fountain, Edward Miller, Selen Basol, Richard Newcombe, Robert Wang, et al. Introducing hot3d: An egocentric dataset for 3d hand and object tracking. *arXiv preprint arXiv:2406.09598*, 2024. 5

[3] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. π0: A vision-language-action flow model for general robot control. corr, abs/2410.24164, 2024. doi: 10.48550. *arXiv preprint ARXIV.2410.24164*. 2

[4] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 2

[5] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022. 2

[6] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. Video generation models as world simulators. *OpenAI Blog*, 1(8):1, 2024. 2

[7] Junuk Cha, Jihyeon Kim, Jae Shin Yoon, and Seungryul Baek. Text2hoi: Text-guided 3d motion generation for hand-object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1577–1585, 2024. 3

[8] Junhao Chen, Xiang Li, Xiaojun Ye, Chao Li, Zhaoxin Fan, and Hao Zhao. Idea23d: Collaborative lmm agents enable 3d model generation from interleaved multimodal inputs. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4149–4166, 2025. 3

[9] 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. 2, 3

[10] Mingjin Chen, Junhao Chen, Huan-ang Gao, Xiaoxue Chen, Zhaoxin Fan, and Hao Zhao. Ultraman: ultra-fast and high-resolution texture generation for 3d human reconstruction from a single image. *Machine Vision and Applications*, 37 (2):24, 2026. 3

[11] Lingwei Dang, Ruizhi Shao, Hongwen Zhang, Wei Min, Yebin Liu, and Qingyao Wu. Svimo: Synchronized diffusion for video and motion generation in hand-object interaction scenarios. *arXiv preprint arXiv:2506.02444*, 2025. 2

[12] Google DeepMind. Veo 3. https://deepmind.google/technologies/veo/, 2024. 2

[13] Yukiyasu Domae, Haruhisa Okuda, Yuichi Taguchi, Kazuhiko Sumi, and Takashi Hirai. Fast graspability evaluation on single depth maps for bin picking with general grippers. In *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1997–2004. IEEE, 2014. 2

[14] Shiying Duan, Pei Ren, Nanxiang Jiang, Zhengping Che, Jian Tang, Zhaoxin Fan, Yifan Sun, and Wenjun Wu. Robopara: Dual-arm robot planning with parallel allocation and recomposition across tasks. *arXiv preprint arXiv:2506.06683*, 2025. 2

[15] Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J Black, and Otmar Hilliges. Arctic: A dataset for dexterous bimanual hand-object manipulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12943–12954, 2023. 3, 5

[16] Zicong Fan, Maria Parelli, Maria Eleni Kadoglou, Xu Chen, Muhammed Kocabas, Michael J Black, and Otmar Hilliges. Hold: Category-agnostic 3d reconstruction of interacting hands and objects from video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 494–504, 2024. 3

[17] Zekai Gu, Rui Yan, Jiahao Lu, Peng Li, Zhiyang Dou, Chenyang Si, Zhen Dong, Qifeng Liu, Cheng Lin, Ziwei Liu, et al. Diffusion as shader: 3d-aware video diffusion for versatile video generation control. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pages 1–12, 2025. 2, 5, 6

[18] Yana Hasson, Gul Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11807–11816, 2019. 3

[19] Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for text-to-video generation. *arXiv preprint arXiv:2404.02101*, 2024. 2

[20] 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. 5

[21] 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. 5

[22] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2

[23] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in neural information processing systems*, 35:8633–8646, 2022. 2

[24] Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Manuel Y. Galliker, Dibya Ghosh, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Devin LeBlanc, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Allen Z. Ren, Lucy Xiaoyang Shi, Laura Smith, Jost Tobias Springenberg, Kyle Stachowicz, James Tanner, Quan Vuong, Homer Walke, Anna Walling, Haohuan Wang, Lili Yu, and Ury Zhilinsky. $\pi_{0.5}$: a vision-language-action model with open-world generalization, 2025. 2

[25] Yash Jain, Anshul Nasery, Vibhav Vineet, and Harkirat Behl. Peekaboo: Interactive video generation via masked-diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8079–8088, 2024. 2

[26] Sangwon Jang, Taekyung Ki, Jaehyeong Jo, Jaehong Yoon, Soo Ye Kim, Zhe Lin, and Sung Ju Hwang. Frame guidance: Training-free guidance for frame-level control in video diffusion models. *arXiv preprint arXiv:2506.07177*, 2025. 3

[27] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Yolo by ultralytics. https://github.com/ultralytics/ultralytics, 2023. 4

[28] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 3

[29] Rawal Khirodkar, Timur Bagautdinov, Julieta Martinez, Su Zhaoen, Austin James, Peter Selednik, Stuart Anderson, and Shunsuke Saito. Sapiens: Foundation for human vision models. *arXiv preprint arXiv:2408.12569*, 2024. 3

[30] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 4

[31] 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. 2

[32] Kuaishou. Kling. https://kling.kuaishou.com/, 2024. 2, 5, 6

[33] Ariel Lapid, Idan Achituve, Lior Bracha, and Ethan Fetaya. Gd-vdm: Generated depth for better diffusion-based video generation. *arXiv preprint arXiv:2306.11173*, 2023. 3

[34] Yujian Lee, Peng Gao, Yongqi Xu, and Wentao Fan. How do optical flow and textual prompts collaborate to assist in audio-visual semantic segmentation? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23342–23352, 2025. 2

[35] Biwen Lei, Yang Li, Xinhai Liu, Shuhui Yang, Lixin Xu, Jingwei Huang, Ruining Tang, Haohan Weng, Jian Liu, Jing Xu, et al. Hunyuan3d studio: End-to-end ai pipeline for game-ready 3d asset generation. *arXiv preprint arXiv:2509.12815*, 2025. 3

[36] Hongxiang Li, Yaowei Li, Yuhang Yang, Junjie Cao, Zhihong Zhu, Xuxin Cheng, and Long Chen. Dispose: Disentangling pose guidance for controllable human image animation. *arXiv preprint arXiv:2412.09349*, 2024. 3

[37] Junlong Li, Huaiyuan Xu, Sijie Cheng, Kejun Wu, Kim-Hui Yap, Lap-Pui Chau, and Yi Wang. Building egocentric procedural ai assistant: Methods, benchmarks, and challenges. *arXiv preprint arXiv:2511.13261*, 2025. 3

[38] Peng Li, Wangguandong Zheng, Yuan Liu, Tao Yu, Yangguang Li, Xingqun Qi, Xiaowei Chi, Siyu Xia, Yan-Pei Cao, Wei Xue, et al. Pshuman: Photorealistic single-view human reconstruction using cross-scale diffusion. 2024. 3

[39] Yaowei Li, Xintao Wang, Zhaoyang Zhang, Zhouxia Wang, Ziyang Yuan, Liangbin Xie, Ying Shan, and Yuexian Zou. Image conductor: Precision control for interactive video synthesis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5031–5038, 2025. 2

[40] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022. 2

[41] Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. Wonder3d: Single image to 3d using cross-domain diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9970–9980, 2024. 3

[42] Yuxuan Luo, Zhengkun Rong, Lizhen Wang, Longhao Zhang, and Tianshu Hu. Dreamactor-m1: Holistic, expressive and robust human image animation with hybrid guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11036–11046, 2025. 3

[43] Jeremy Maitin-Shepard, Marco Cusumano-Towner, Jinna Lei, and Pieter Abbeel. Cloth grasp point detection based on multiple-view geometric cues with application to robotic towel folding. In *2010 IEEE International Conference on Robotics and Automation*, pages 2308–2315. IEEE, 2010. 2

[44] Priyanka Mandikal and Kristen Grauman. Dexvip: Learning dexterous grasping with human hand pose priors from video. In *Conference on Robot Learning*, pages 651–661. PMLR, 2022. 2

[45] Xingyu Miao, Junting Dong, Qin Zhao, Yuhang Yang, Junhao Chen, and Yang Long. From frames to sequences: Temporally consistent human-centric dense prediction. *arXiv preprint arXiv:2602.01661*, 2026. 3

[46] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 3

[47] 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. 5

[48] Koichi Namekata, Sherwin Bahmani, Ziyi Wu, Yash Kant, Igor Gilitschenski, and David B Lindell. Sg-i2v: Self-guided trajectory control in image-to-video generation. *arXiv preprint arXiv:2411.04989*, 2024. 2

[49] Muyao Niu, Xiaodong Cun, Xintao Wang, Yong Zhang, Ying Shan, and Yinqiang Zheng. Mofa-video: Controllable image animation via generative motion field adaptions in frozen image-to-video diffusion model. In *European Conference on Computer Vision*, pages 111–128. Springer, 2024. 2

[50] Youxin Pang, Ruizhi Shao, Jiajun Zhang, Hanzhang Tu, Yun Liu, Boyao Zhou, Hongwen Zhang, and Yebin Liu. Manivideo: Generating hand-object manipulation video with dexterous and generalizable grasping. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12209–12219, 2025. 2

[51] Yatian Pang, Bin Zhu, Bin Lin, Mingzhe Zheng, Francis EH Tay, Ser-Nam Lim, Harry Yang, and Li Yuan. Dreamdance: Animating human images by enriching 3d geometry cues from 2d poses. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14039–14050, 2025. 3

[52] Bohao Peng, Jian Wang, Yuechen Zhang, Wenbo Li, Ming-Chang Yang, and Jiaya Jia. Controlnext: Powerful and efficient control for image and video generation. *arXiv preprint arXiv:2408.06070*, 2024. 2, 3

[53] Haonan Qiu, Zhaoxi Chen, Zhouxia Wang, Yingqing He, Menghan Xia, and Ziwei Liu. Freetraj: Tuning-free trajectory control in video diffusion models. *arXiv preprint arXiv:2406.16863*, 2024. 2

[54] Xiaoyu Shi, Zhaoyang Huang, Fu-Yun Wang, Weikang Bian, Dasong Li, Yi Zhang, Manyuan Zhang, Ka Chun Cheung, Simon See, Hongwei Qin, et al. Motion-i2v: Consistent and controllable image-to-video generation with explicit motion modeling. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 2

[55] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. 2

[56] Chaoyue Song, Jianfeng Zhang, Xiu Li, Fan Yang, Yiwen Chen, Zhongcong Xu, Jun Hao Liew, Xiaoyang Guo, Fayao Liu, Jiashi Feng, et al. Magicarticulate: Make your 3d models articulation-ready. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 15998–16007, 2025. 3

[57] Yuejiao Su, Yi Wang, Lei Yao, Yawen Cui, and Lap-Pui Chau. Interaction-aware representation modeling with co-occurrence consistency for egocentric hand-object parsing. *arXiv preprint arXiv:2602.20597*, 2026. 3

[58] Sruthi Sudhakar, Ruoshi Liu, Basile Van Hoorick, Carl Vondrick, and Richard Zemel. Controlling the world by sleight of hand, 2024. 2, 3

[59] Mingze Sun, Junhao Chen, Junting Dong, Yurun Chen, Xinyu Jiang, Shiwei Mao, Puhua Jiang, Jingbo Wang, Bo Dai, and Ruqi Huang. Drive: Diffusion-based rigging empowers generation of versatile and expressive characters. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 21170–21180, 2025. 3

[60] Shuyuan Tu, Zhen Xing, Xintong Han, Zhi-Qi Cheng, Qi Dai, Chong Luo, and Zuxuan Wu. Stableanimator: High-quality identity-preserving human image animation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 21096–21106, 2025. 2, 5

[61] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Fvd: A new metric for video generation. 2019. 5

[62] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 5, 6

[63] Jiawei Wang, Yuchen Zhang, Jiaxin Zou, Yan Zeng, Guoqiang Wei, Liping Yuan, and Hang Li. Boximator: Generating rich and controllable motions for video synthesis. *arXiv preprint arXiv:2402.01566*, 2024. 2

[64] 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. 5

[65] Qilin Wang, Zhengkai Jiang, Chengming Xu, Jiangning Zhang, Yabiao Wang, Xinyi Zhang, Yun Cao, Weijian Cao, Chengjie Wang, and Yanwei Fu. Vividpose: Advancing stable video diffusion for realistic human image animation. *arXiv preprint arXiv:2405.18156*, 2024. 2, 3

[66] Youzhuo Wang, Jiayi Ye, Chuyang Xiao, Yiming Zhong, Heng Tao, Hang Yu, Yumeng Liu, Jingyi Yu, and Yuexin Ma. Dexh2r: A benchmark for dynamic dexterous grasping in human-to-robot handover. *arXiv preprint arXiv:2506.23152*, 2025. 2

[67] 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. 5

[68] Zhenzhi Wang, Yixuan Li, Yanhong Zeng, Youqing Fang, Yuwei Guo, Wenran Liu, Jing Tan, Kai Chen, Tianfan Xue, Bo Dai, et al. Humanvid: Demystifying training data for camera-controllable human image animation. *Advances in Neural Information Processing Systems*, 37:20111–20131, 2024. 3

[69] Zhengyi Wang, Jonathan Lorraine, Yikai Wang, Hang Su, Jun Zhu, Sanja Fidler, and Xiaohui Zeng. Llama-mesh: Unifying 3d mesh generation with language models. *arXiv preprint arXiv:2411.09595*, 2024. 3

[70] Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 2

[71] Zhenzhi Wang, Yixuan Li, Yanhong Zeng, Yuwei Guo, Dahua Lin, Tianfan Xue, and Bo Dai. Multi-identity human image animation with structural video diffusion. *arXiv preprint arXiv:2504.04126*, 2025. 3

[72] Fangsheng Weng, Junhao Chen, Xiang Li, Jie Qin, Hanzhong Guo, Xiaoguang Han, et al. Garmentgpt: Compositional garment pattern generation via discrete latent tokenization. In *The Fourteenth International Conference on Learning Representations*, 2026. 3

[73] Yuxi Xiao, Qianqian Wang, Shangzhan Zhang, Nan Xue, Sida Peng, Yujun Shen, and Xiaowei Zhou. Spatialtracker: Tracking any 2d pixels in 3d space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20406–20417, 2024. 5

[74] 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. 5

[75] 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. 5

[76] 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, 3, 5, 6

[77] Chongjie Ye, Yushuang Wu, Ziteng Lu, Jiahao Chang, Xiaoyang Guo, Jiaqing Zhou, Hao Zhao, and Xiaoguang Han. Hi3dgen: High-fidelity 3d geometry generation from images via normal bridging. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 25050–25061, 2025. 3

[78] Wang Yifan, Felice Serena, Shihao Wu, Cengiz Öztireli, and Olga Sorkine-Hornung. Differentiable surface splatting for point-based geometry processing. *ACM Transactions On Graphics (TOG)*, 38(6):1–14, 2019. 3

[79] Shengming Yin, Chenfei Wu, Jian Liang, Jie Shi, Houqiang Li, Gong Ming, and Nan Duan. Dragnuwa: Fine-grained control in video generation by integrating text, image, and trajectory. *arXiv preprint arXiv:2308.08089*, 2023. 2

[80] 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. 4

[81] Mengqi Zhang, Yang Fu, Zheng Ding, Sifei Liu, Zhuowen Tu, and Xiaolong Wang. Hoidiffusion: Generating realistic 3d hand-object interaction data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8521–8531, 2024. 2, 3

[82] Yuang 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. 2, 3

[83] Zhenghao Zhang, Junchao Liao, Menghao Li, Zuozhuo Dai, Bingxue Qiu, Siyu Zhu, Long Qin, and Weizhi Wang. Tora: Trajectory-oriented diffusion transformer for video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2063–2073, 2025. 2

[84] 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, 3, 4, 5

[85] Yiming Zhong, Qi Jiang, Jingyi Yu, and Yuexin Ma. Dexgrasp anything: Towards universal robotic dexterous grasping with physics awareness. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 22584–22594, 2025. 2

[86] Haitao Zhou, Chuang Wang, Rui Nie, Jinlin Liu, Dongdong Yu, Qian Yu, and Changhu Wang. Trackgo: A flexible and efficient method for controllable video generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10743–10751, 2025. 2

[87] Shenhao Zhu, Junming Leo Chen, Zuozhuo Dai, Zilong Dong, Yinghui Xu, Xun Cao, Yao Yao, Hao Zhu, and Siyu Zhu. Champ: Controllable and consistent human image animation with 3d parametric guidance. In *European Conference on Computer Vision*, pages 145–162. Springer, 2024. 2

[88] Yufei Zhu, Yiming Zhong, Zemin Yang, Peishan Cong, Jingyi Yu, Xinge Zhu, and Yuexin Ma. Evolvinggrasp: Evolutionary grasp generation via efficient preference alignment. *arXiv preprint arXiv:2503.14329*, 2025. 2

[89] Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, pages 2165–2183. PMLR, 2023. 2