

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

Ziyi Xu^{1,†*} Zejing Rao^{1,†*} Juan Cao² Xiaoqiang Liu³ Zhixue Fang³
 Haoxian Zhang³ Songlin Tang³ Fan Tang^{2,§}
¹University of Chinese Academy of Sciences
²Institute of Computing Technology, Chinese Academy of Sciences
³Kling Team, Kuaishou Technology



Figure 1. We propose GeoHOI, a human-object interaction video generation framework built on the DiT architecture. By introducing object geometry priors and hierarchical injection policy, GeoHOI enhances object perception and modeling, enabling the preservation of shapes and textures under complex motions and achieving high-quality interaction generation.

Abstract

Existing human-object interaction (HOI) video generation methods struggle to accurately perceive object structures, often producing poor results during object rotation or complex interactions. We found inspiration in visual geometry-grounded models, as their powerful ability to capture meaningful geometric-aware structures in learned representations. However, due to the inherent heterogeneity between models trained for different tasks, applying such models to HOI video generation presents significant challenges. We propose GeoHOI, a framework that encourages diffusion transformers to fully utilize geometry-aware priors to enhance object perception and physical consistency of HOI video generation. We introduce a Multi-View Geometry

Prior Extractor which captures object structural features and incorporates a Camera Pose Modulation module to provide viewpoint adaptation for multi-view object inputs. To enable effective injection of geometry-aware features and prevent feature entanglement, we design a Hierarchical Injection Policy that strategically integrates geometric, texture, and caption embeddings at different layers of the diffusion transformer. GeoHOI achieves a 26.2% relative improvement in Object-IoU and a 19.6% reduction in Hand Landmark Mean Distance (LMD) compared to our baseline model (Wan2.2-5B), outperforming state-of-the-art methods in object understanding and interaction plausibility.

1. Introduction

Video generation models have seen remarkable progress in recent years, particularly in human motion synthesis, leading to advanced methods like Animate-Anyone [9] and

[†] Work done during internship at Kling Team, Kuaishou Technology.

^{*} Equal contribution. xuziyi23@mails.ucas.ac.cn

[§] Corresponding author. tfan.108@gmail.com

Champ [46] that utilize pose as a driving modality. However, extending these successes to the Human-Object Interaction (HOI) domain introduces substantially greater complexity. HOI requires generation models not only to precisely control human actions but also to successfully integrate and dynamically handle a target object, necessitating a precise understanding of the object’s structure and dynamic geometry.

Existing HOI video generation frameworks, such as HunyuanVideo-HOMA [12], typically encode object images using generic 2D feature extractors (e.g., VAE [15] or DINO [28]), which primarily capture texture and color information but often neglect cross-view geometric correspondence. AnchorCrafter [41] attempts to preserve object appearance through multi-view image fusion. However, this reliance on 2D texture representation leads to a fundamental flaw: the model perceives the object as a “flat texture” rather than a “volumetric entity.” The catastrophic consequences include shape distortion during object rotation and inaccurate contact geometry during motion synthesis. This lack of volumetric understanding fundamentally violates basic physical plausibility.

Recent progress [8, 38] suggests that geometry-grounded vision models such as VGGT [34] can provide strong structural priors for generative modeling, enabling downstream models to reason about 3D structures rather than 2D textures. However, applying VGGT’s geometric representations to HOI video generation is non-trivial. HOI synthesis requires the model to dynamically integrate object geometry with human motion, demanding an understanding of both “shape consistency” and “interaction geometry.” The central challenge thus becomes one of architectural integration: How can the static, view-invariant 3D structural representations derived from VGGT be dynamically and coherently fused into the multi-modal, temporal latent space of the Diffusion Transformer (DiT), ensuring that object shape remains consistent across dynamic movement and that the crucial human-object contact geometry is physically accurate?

In this study, we introduce GeoHOI, a novel framework, that utilizes explicit 3D structural priors introduced by visual geometry grounded models to fundamentally resolve the object distortion problem and achieve physically consistent HOI video generation. To ensure that the object’s latent features robustly maintain consistency in dynamic scenes, we design a camera pose modulation (CPM) module for viewpoint-adaptive geometry encoding by leveraging predicted camera poses to perform feature-wise linear modulation, enabling the model to distinguish view-dependent appearance differences and learn consistent geometry across multiple views. Furthermore, instead of directly aligning the extracted intermediate representations with those of the diffusion model, which is unsuitable for HOI generation due

to the mismatch between geometric and generative feature spaces, we applied a decoupled multi-modal fusion following a hierarchical injection policy (HIP), which strategically injects geometric, texture, and semantic features into different hierarchical layers of the Diffusion Transformer based on the abstraction needs of the model. By decoupling the treatment of structural cues and appearance details, HIP avoids the representation interference problem commonly found in multi-modal feature fusion. In summary, our contributions are as follows:

- We propose GeoHOI, a novel HOI video generation framework that leverages visual geometry grounded models to introduce 3D structural priors and spatial reasoning capabilities, enabling more geometry-aware and physically consistent generation.
- We introduce a **Multi-View Geometry prior Extractor** with a **Camera Pose Modulation** module to deliver view-consistent 3D structural information. We further design a **Hierarchical Injection Policy** to alleviate representation interference in multi-modal feature injection.
- We conduct extensive quantitative and qualitative experiments demonstrating that GeoHOI outperforms state-of-the-art diffusion-based HOI video generation methods.

2. Related work

2.1. Pose-Guided Human Video Generation

With the rapid maturation of video generation technologies [2, 16, 33], pose-guided human video generation has emerged as a pivotal research direction, aiming to synthesize high-fidelity and temporally coherent video sequences that accurately align with predefined pose trajectories. Early methods [11, 29, 36, 40] leveraged ControlNet [44] to inject skeletal motion cues, yet incurred substantial parameter overhead. To mitigate this, MimicMotion [45] introduced a lightweight confidence-aware pose guidance network, while Champ [46] and MoAnimate [6] enhanced motion controllability by fusing multiple types of pose representations. For appearance consistency, AnimateAnyone [9] introduced ReferenceNet via self-attention, and StableAnimator [30], which incorporates a facial encoder to preserve facial details.

The recent advent of Diffusion Transformer (DiT) architectures [7, 13, 19, 37] has further accelerated progress in this domain. HumanDiT [7] achieves pose transfer through a pose adapter, UniAnimate-DiT [37] injects both driving and reference pose features to enhance spatial alignment, OmniHuman [17] supports multimodal signals, and VideoJAM [4] generates complex motion via joint appearance–motion representations. Despite these advances, most approaches remain confined to single-subject generation, with limited modeling of objects and scene context, restricting controllable synthesis of human–object interactions.

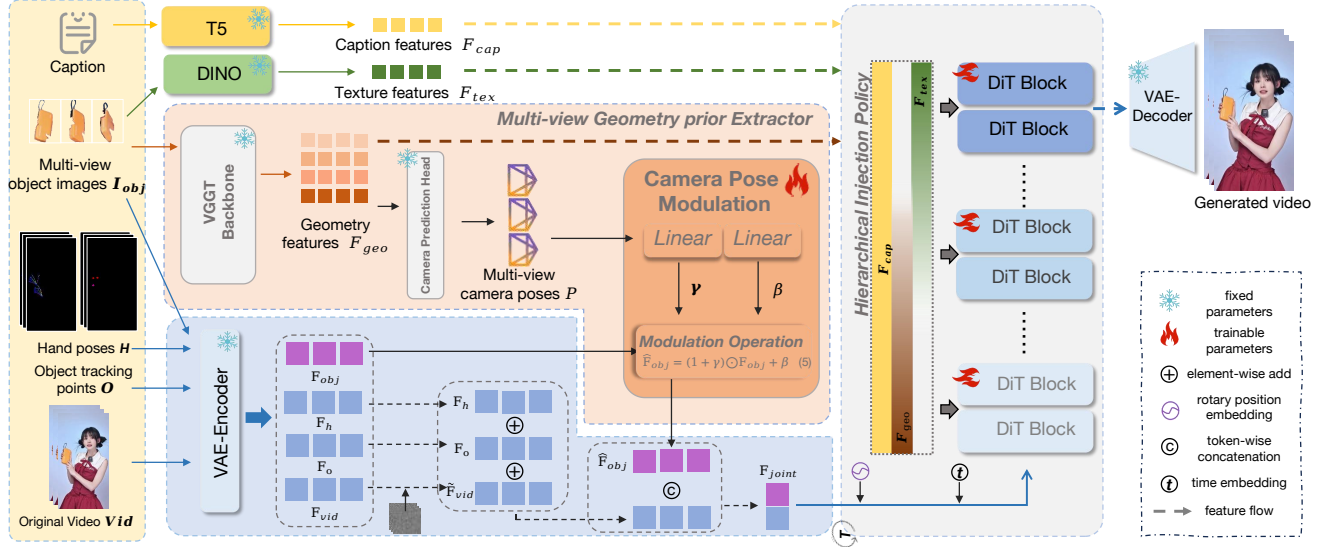


Figure 2. Training pipeline of GeoHOI: based on a video generation model, GeoHOI employs the Multi-view Geometry Prior Extractor to capture object geometry priors and designs a Hierarchical Injection Policy to effectively integrate multimodal information while preventing confusion among features from different sources.

2.2. Human-Object Interaction Generation

Human-Object Interaction (HOI) aims to generate visual content that accurately depicts interactions between humans and objects. Some methods address this as a video editing problem by replacing elements within existing scenes. For example, MIMO [20] and AnimateAnyone2 [10] attempt to replace human subjects in complex dynamic interaction scenarios, whereas HOISwap [42], ReHold [5], and iDiT-HOI [27] specialize in substituting handheld objects in real videos.

The core challenge of true HOI video generation—synthesizing coherent interactions from conditional inputs alone—has now seen its initial progress [12, 35, 41]. AnchorCrafter [41] leverages multi-conditional inputs such as pose and depth to represent perceptual human motion during interactions, but requires fine-tuning for specific objects. DreamActor-H1 [35] retrieves optimal motion templates from a predefined pool and adjusts object bounding box scaling through joint analysis of reference human/product images. Hunyuan Video-HOMA [12] explores a sparse, user-editable input framework that relies on minimal yet semantically meaningful conditions for user-friendly generation. However, these approaches exhibit limited object awareness and are restricted to simple actions such as translation, struggling to maintain consistency when objects undergo rotation or complex interactions.

3. Method

As illustrated in Fig. 2, our system’s pipeline consists of two core components: Multi-view Geometry prior Extractor and

Hierarchical Injection Policy. Section 3.1 covers the Diffusion Transformer and VGGT fundamentals. Section 3.2 gives the framework overview. The following sections describe our core technical contributions. Section 3.3 details the Multi-view Geometry Prior Extractor with a Camera Pose Modulation module. Section 3.4 introduces the Hierarchical Injection Policy. Finally, Section 3.5 explains our Data Processing methodology.

3.1. Preliminary

Diffusion Transformer. The Diffusion Transformer (DiT) [21] combines the strengths of diffusion models with transformers [32], addressing the limited scalability and representation capacity of U-Net-based latent diffusion models. In our work, we follow the flow matching [18] paradigm for deterministic sample generation. Given a clean data sample $\mathbf{x}_0 \sim p_{\text{data}}(\mathbf{x}_0)$ and Gaussian noise $\mathbf{x}_1 \sim \mathcal{N}(0, \mathbf{I})$, an intermediate noisy sample is constructed by linear interpolation:

$$\mathbf{x}_t = (1 - t)\mathbf{x}_0 + t\mathbf{x}_1, \quad t \in [0, 1]. \quad (1)$$

The instantaneous target velocity along this trajectory is defined as

$$\mathbf{v}_t = \frac{d\mathbf{x}_t}{dt} = \mathbf{x}_1 - \mathbf{x}_0, \quad (2)$$

and the model $\mathbf{v}_\theta(\mathbf{x}_t, t)$ is trained to approximate this field using a mean squared error objective:

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{\mathbf{x}_0, \mathbf{x}_1, t} \left[\|\mathbf{v}_\theta(\mathbf{x}_t, t) - (\mathbf{x}_1 - \mathbf{x}_0)\|_2^2 \right]. \quad (3)$$

During inference, the model integrates the learned flow field in reverse (from $t = 1$ to $t = 0$) to deterministically generate clean samples from noise.

Visual Geometry Grounded Transformer. The Visual Geometry Grounded Transformer (VGGT) [34] is a large feed-forward transformer that predicts key 3D attributes—camera parameters, depth maps, point maps, and point tracks—from one or many input views in a single forward pass. Its geometry-aware representations transfer well to downstream tasks such as point tracking and feed-forward novel view synthesis, and can act as geometry priors for 2D generative models to stabilize viewpoint changes and preserve spatial layout in generated content.

3.2. Overview

As illustrated in Fig. 2, our method uses hand pose sequences and sparse object tracking points for fine-grained motion control, while object appearance is represented with a flexible set of images. A caption further describes the video content, providing semantic guidance for synthesis. Following a first-frame generation paradigm, the objective is to maintain geometric consistency across subsequent frames while synthesizing plausible interaction motions.

During training, the original T-frame video, hand pose sequences, object tracking points, and V multi-view object images are encoded by a shared VAE encoder into a unified latent space $\mathbf{F}_{vid}$, $\mathbf{F}_h$, $\mathbf{F}_o \in \mathbb{R}^{T \times H \times W \times C}$, and $\mathbf{F}_{obj} \in \mathbb{R}^{V \times H \times W \times C}$. Gaussian noise is added to $\mathbf{F}_{vid}$ except for the first frame, yielding $\tilde{\mathbf{F}}_{vid}$. The object features $\mathbf{F}_{obj}$ are further processed by the Camera Pose Modulation module to get $\hat{\mathbf{F}}_{obj}$. The joint representation is formulated as:

$$\mathbf{F}_{joint} = \text{Concat}(\hat{\mathbf{F}}_{obj}, \tilde{\mathbf{F}}_{vid} + \mathbf{F}_h + \mathbf{F}_o), \quad (4)$$

which is then fed into the Diffusion Transformer (DiT) for subsequent denoising and generation.

In addition, we employ VGGT and DINO to extract object features, with VGGT modeling geometry features $\mathbf{F}_{geo}$ and DINO capturing texture features $\mathbf{F}_{tex}$. The caption features $\mathbf{F}_{cap}$ is encoded by T5 [23]. These features are injected into DiT following our Hierarchical Injection Policy, thereby enhancing geometry-aware human-object interaction video generation.

3.3. Multi-view Geometry Prior Extractor

We introduce a Multi-view Geometry Prior Extractor to provide geometry-aware structural cues for human-object interaction synthesis. Given an arbitrary set of V object images, we model them using VGGT. Instead of relying on its final outputs (e.g., sparse point clouds), which are difficult to learn and exhibit a feature-level gap with our model, we adopt intermediate features as the geometry-aware representation $\mathbf{F}_{geo}$. Specifically, we collect feature tokens from

the 4th, 11th, 17th, and 23rd transformer blocks, which progressively capture geometric information ranging from local shapes and contours to global structural configurations. In addition, the VGGT camera head estimates camera pose tokens $P = \{p_i\}_{i=1}^V$, $p \in \mathbb{R}^{V \times 9}$ for each object image, providing explicit viewpoint cues. These poses are subsequently fed into the Camera Pose Modulation (CPM) module to perform feature-wise linear modulation.

Camera Pose Modulation. Following the in-context conditioning (ICC) paradigm, we concatenate the latent representations of multi-view object images with the input to enhance cross-view consistency. To further provide explicit viewpoint information, we introduce a Camera Pose Modulation module, which modulates multi-view object representations based on the predicted camera poses. Specifically, the predicted poses P are injected into two linear networks to produce scaling and shifting parameters, which are then applied to the object latent representations:

$$\hat{\mathbf{F}}_{obj} = (1 + \gamma) \odot \mathbf{F}_{obj} + \beta, \quad (5)$$

where $\odot$ denotes channel-wise multiplication, γ and β denote the scaling and shifting values derived from the camera poses, and $\mathbf{F}_{obj}$ represents the original object latent features. This modulation explicitly incorporates viewpoint information into the latent features, thereby enhancing the alignment of object representations across different views and reducing geometric distortions.

3.4. Hierarchical Injection Policy

Multi-modal features are often fused through simple concatenation or alignment, but directly combining heterogeneous sources can cause representation interference and hinder effective fusion. We propose to allocate different layers of DiT to specialize in processing distinct feature modalities, such as geometry and texture feature, thereby improving the disentanglement and alignment of multi-modal representations. For instance, the shallow layers of a diffusion transformer primarily capture spatial and motion structures, while the deep layers are responsible for refining appearance details and fine-grained features [39]. Consequently, we propose a Hierarchical Injection Policy that progressively integrates multi-modal conditions at different blocks of the DiT, as detailed below:

- **Shallow DiT blocks (layers $\leq m$):** These layers receive both geometry features $\mathbf{F}_{geo}$ and caption features $\mathbf{F}_{cap}$ via cross-attention. They focus on learning the object’s structural geometry and motion dynamics.
- **Middle DiT blocks ($m < \text{layers} < n$):** In addition to $\mathbf{F}_{geo}$ and $\mathbf{F}_{cap}$, these blocks incorporate texture features $\mathbf{F}_{tex}$ through cross-attention. The weight of geometry features gradually decreases, while the weight of texture features increases.

- **Deep DiT blocks (layers $\geq n$):** These layers are conditioned on $\mathbf{F}_{cap}$ and $\mathbf{F}_{tex}$, focusing on fine-grained appearance details, texture refinement, and semantic consistency.

In our work, with 30 DiT blocks in total, we set $m=12$ and $n=18$ to achieve a balanced integration of structure and appearance. The cross-attention operation at each DiT block can be formulated as:

$$\begin{aligned} \mathbf{y} = \mathbf{W}_o & \left[\text{Attn}(\mathbf{Q}, \mathbf{K}_{cap}, \mathbf{V}_{cap}) + \phi \text{Attn}(\mathbf{Q}, \mathbf{K}_{geo}, \mathbf{V}_{geo}) \right. \\ & \left. + \omega \text{Attn}(\mathbf{Q}, \mathbf{K}_{tex}, \mathbf{V}_{tex}) \right], \quad (6) \\ \text{Attn}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = & \text{softmax} \left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}} \right) \mathbf{V}, \end{aligned}$$

where $\text{Attn}(\mathbf{Q}, \mathbf{K}, \mathbf{V})$ denotes the multi-head attention mechanism, and $\mathbf{W}_o$ denotes the output projection matrix that maps the aggregated attention features back to the latent feature space. ϕ and ω are dynamically scheduled according to the layer depth such that shallow layers prioritize geometric feature, while deep layers focus on texture feature. This hierarchical injection policy enables the model to integrate geometry, texture, and semantic information in a layer-aware manner, achieving geometry-consistent and visually coherent HOI video generation.

3.5. Data Processing

We propose a systematic data processing pipeline designed to efficiently handle data. Specifically, we employ DW-Pose [43] to extract the hand pose sequence of the character in each video. For object-related conditions, we first utilize Qwen2.5-VL [1] to generate a video description together with a detailed object description specifying category and shape. Based on this description, Grounded-Segment-Anything [25] is applied to detect and segment the target object on first frame each videos. The segmentation is then propagated across the entire clip using SAM2 [24], resulting in an object mask sequence. From this mask sequence, we randomly sample the object across frames to obtain multi-view object images. Finally, we estimate the object tracking point sequence with CoTracker [14], conditioned on the first-frame mask provided by SAM2.

4. Experiments

4.1. Experimental Settings

Training and Inference Setup. We initialize our model with pre-trained weights from Wan2.2-5B. During training, we freeze the VAE, T5, DINO, and VGGT modules and train all other parameters for 4 epochs on eight 80GB GPUs, using a learning rate of 10^{-5} with 300 warm-up iterations. At inference time, generating a 125-frame video at 608×1088 resolution takes approximately 5 minutes and consumes around 15.3 GB of GPU memory.

Method	FID↓	FVD↓	Obj-IoU↑	Obj-CLIP↑	LMD (Hand)↓
MimicMotion-1.5B	37.03	486.39	0.43	83.54	19.31
VACE-14B	41.53	395.48	0.38	85.16	27.22
UniAnimate-DiT-14B	27.48	251.97	0.60	87.86	11.63
Wan2.2-Pose-5B	34.59	379.81	0.61	87.41	12.41
Ours-5B	30.25	351.28	0.77	90.15	9.98
w/o camera pose modulation	30.76	356.37	0.74	89.02	10.77
w/o geometry feature	34.71	398.49	0.75	88.64	10.76
w/o texture feature	35.83	387.62	0.76	88.75	11.29
Swapping injection layers	30.55	357.10	0.74	89.28	10.77

Table 1. Quantitative results of our method compared with SOTAs and ablation study. The results show that, under complex interaction tasks, our approach significantly outperforms prior methods in preserving object consistency, generating accurate motions, and controlling human gestures.

Dataset. We apply automated data processing to internal live streaming e-commerce video clips to support training under diverse interaction scenarios and finally collect approximately 150 hours of training data. For evaluation, we collected 100 web video clips featuring complex human-object interactions and diverse object appearances. Each test video was generated at 25 FPS with a duration of 5 seconds.

Comparison Methods. We compare our method with the following methods: Mimicmotion [45], VACE [13], UniAnimate-DiT [37], Wan2.2-5B, AnchorCrafter [41], and HunyuanVideo-HOMA [12]. MimicMotion and UniAnimate-DiT generate motion from a person image and a sequence of pose. VACE-14B is a general video editing framework and we use the pose control setting. Wan2.2-5B has the capability of generating videos from images. Similar to our method, we incorporate hand pose sequences and object trajectory control, and train it on our dataset. AnchorCrafter [41] requires fine-tuning for new objects, while HunyuanVideo-HOMA [12] is not publicly available; therefore, we only provide case demonstrations.

Metrics. We evaluate our method using several established metrics. For visual quality, we use FID [26] and FVD [31] to evaluate the image and video quality. For object generation, we employ Object-IoU [41], which utilizes SAM2 to extract object motion trajectory masks and compute the intersection over union (IoU) with the ground-truth mask. Additionally, Object-CLIP evaluates appearance consistency by computing the CLIP cosine similarity [22] between generated objects and reference objects. We measure hand consistency using Landmark Mean Distances (LMD), leveraging OpenPose [3].

4.2. Main Results

Quantitative Results. Table 1 reports the quantitative evaluation of our approach. Our method achieves the highest scores in both object consistency and motion accuracy. Although our FID and FVD are slightly lower than UniAnimate-DiT, it is important to emphasize that



Figure 3. Qualitative comparisons with other methods. Under challenging rotating objects and rapid jittering motions, alternative approaches fail to preserve visual object integrity or maintain plausible interaction dynamics, while our method demonstrates superior object structure perception and realistic interaction, even when given a single-view object image as input.

UniAnimate-DiT and VACE are built upon the 14B Wan model, whereas our framework is based on a 5B backbone. Relative to our baseline Wan2.2-5B, our approach achieves improvements in both image quality and video quality. MimicMotion and VACE achieve low scores in object-IoU

and LMD, highlighting their inability to align object and hand motions with the driving signals. Their object-CLIP scores are also low, further indicating a failure to generate consistent object appearances. By contrast, UniAnimate-DiT and Wan2.2-5B obtain considerably higher object-IoU



Figure 4. Comparison with HunyuanVideo-HOMA: while HunyuanVideo-HOMA fails to generate the shoes and misrepresents the bottom object’s size and structure, our method preserves both texture and geometric structures.



Figure 5. Comparison with AnchorCrafter: while AnchorCrafter requires fine-tuning on specific objects, our model achieves comparable results without any fine-tuning and further demonstrates superior human appearance consistency compared to AnchorCrafter.

and LMD scores, yet still remain far below our method. Their object-CLIP scores are only marginally lower than ours, which may be attributed to CLIP’s limited sensitiv-



Figure 6. Ablation study. Unlike other variants that exhibit object deformation or texture errors, our method preserves consistency and produces plausible motion interactions.

ity to fine-grained texture details, making it less capable of capturing subtle appearance inconsistencies.

These quantitative results demonstrate the effectiveness of our framework in advancing human–object interaction video generation beyond current state-of-the-art baselines.

Qualitative Results. As illustrated in Fig. 3, the qualitative results highlight the efficacy of our approach in handling complex interaction scenarios. Notably, our method retains its impressive performance during inference when given a single-view object image as input. In the figures, green boxes highlight inconsistent objects, orange boxes indicate unrealistic interactions, and red boxes mark incorrect motions. Whether handling complex interaction motions with rigid objects such as shoes, or rapid dynamic deformations of non-rigid objects such as clothing, our method consistently achieves superior object structure perception, better appearance preservation, and more plausible motion generation than existing approaches. In contrast, MimicMotion fails to preserve object appearance under large-scale motions, often resulting in severe distortions or even complete deformation. VACE produces unnatural human–object interactions that are misaligned with the intended motion, while also altering the geometric shape of the objects, thereby reducing video realism. Although Unianimate-DiT achieves accurate motion control, it struggles to maintain object texture and shape, leading to noticeable deformations during rotation. Similarly, Wan2.2-5B performs poorly when modeling complex interactions and fails to maintain consistent object appearance.

Fig. 4 presents our results with HunyuanVideo-HOMA. While our method generates reasonable interactive videos, HunyuanVideo-HOMA suffers from object loss, inaccurate textures and incorrect sizes. Fig. 5 compares our results with AnchorCrafter. It is noteworthy that AnchorCrafter requires fine-tuning for each specific object, whereas our

model achieves competitive object consistency without any fine-tuning. Furthermore, our approach demonstrates superior human appearance consistency, while AnchorCrafter often produces inconsistent facial representations.

By comparison, our method achieves precise perception of both object shapes and appearances, producing interaction videos that are geometrically coherent and visually faithful to the input references.

4.3. Ablation Study

Validation of Camera Pose Modulation. As presented in Table 1, removing Camera Pose Modulation results in a noticeable decline in both object-IoU and object-CLIP scores. Consistent with these quantitative findings, Fig. 6 illustrates that the absence of this component leads to object deformations and texture artifacts. Without the contextual conditioning provided by Camera Pose Modulation, object consistency deteriorates and the generated interactions appear less natural.

Validation of Geometry Features Injection. We evaluate the effectiveness of injecting geometry features integration. As reported in Table 1, removing geometry features leads to a reduction in both object-IoU and object-CLIP scores. Fig. 6 shows that objects exhibit noticeable deformations, particularly when rotation requires the synthesis of previously unseen regions in the first frame. These observations suggest that, without effective geometric priors, the model struggles to accurately capture and preserve object shapes. In contrast, incorporating geometry features provides strong structural priors that uphold 3D consistency and mitigate ghosting, thereby enabling stable geometry and physically coherent motion across multiple views.

Validation of Texture Features Injection. We evaluate the effectiveness of texture feature injection by removing this component. As shown in Table 1 and Fig. 6, the absence of texture feature injection leads to blurred or unstable textures when objects undergo rapid motion, even though their overall shapes are still preserved. This observation underscores the critical role of texture features in maintaining appearance fidelity under dynamic motion.

Validation of Swapping Injection Layers. To evaluate the effectiveness of our injection policy, we exchange the order of geometry and texture feature injection. As shown in Table 1, this reversal leads to a decrease in both object-IoU and object-CLIP scores. Fig. 6 further illustrates that although both modalities are injected, altering the order prevents the model from capturing accurate object structures or fine texture details, thereby confirming the effectiveness of injecting geometry feature before texture feature.

4.4. User Study

We conducted a user study to evaluate the quality of the generated videos using five subjective criteria. For humans,

we assessed appearance consistency and hand motion accuracy, while for objects we evaluated appearance consistency and motion accuracy. In addition, the overall video quality was evaluated. For each test case, the results of all methods were randomly displayed, and participants were asked to rate them on a scale from 1 to 5. A total of 30 participants were recruited, representing a diverse range of backgrounds and experiences, with varying levels of familiarity with video generation technology. This group included 15 males and 15 females, ensuring balanced gender representation. The final scores were obtained by averaging the ratings provided by the participants. As shown in Table 2, our method consistently outperforms all other approaches across every evaluation criterion, with particularly pronounced advantages on object-related metrics.

Method	Human		Object		Overall
	Appearance	Hand Motion	Appearance	Motion	
MimicMotion	2.98	2.15	1.51	1.80	1.84
VACE	3.50	3.14	2.88	3.42	3.38
UniAnimate-DiT	3.26	3.59	3.12	3.11	3.68
Wan2.2-Pose	3.64	3.70	2.54	2.47	3.22
Ours	4.11	4.29	4.55	4.63	4.48

Table 2. User study scores. The rating score is on a scale from one to five, where one is the lowest, and five is the highest score.

4.5. Limitations

Due to limitations of the training set, GeoHOI is trained on single-person datasets and may therefore struggle to generate high-quality results in more complex multi-person scenarios. In addition, the system performs poorly when handling transparent objects. Furthermore, inaccuracies in pre-processing models (e.g., DWPose) can introduce errors into the generated results.

5. Conclusion

We present GeoHOI, a generative system based on the DiT architecture for producing human-object interaction videos with complex motions. Our system introduces both geometric and texture features, and leverages a Multi-view Geometry Prior Extractor together with a Hierarchical Injection Policy to enable effective multimodal feature integration. Extensive experiments demonstrate that, through novel geometric feature extraction and injection, GeoHOI outperforms existing methods. It ensures high-quality video generation, preserving object shapes and texture details while maintaining realistic interaction dynamics, even under complex motions such as rapid object deformation, rotation, and interactions with non-rigid objects. This ability enables GeoHOI to deliver consistent realism and coherence across a wide range of challenging scenarios.

Acknowledgements

This work was partly supported by the National Natural Science Foundation of China (NSFC) under Grant 62572458 and the Beijing Major Science and Technology Project under No. Z251100008125009

References

- [1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhao-hai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. 5
- [2] 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
- [3] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017. 5
- [4] Hila Chefer, Uriel Singer, Amit Zohar, Yuval Kirstain, Adam Polyak, Yaniv Taigman, Lior Wolf, and Shelly Sheynin. Videojam: Joint appearance-motion representations for enhanced motion generation in video models. *arXiv preprint arXiv:2502.02492*, 2025. 2
- [5] Yingying Fan, Quanwei Yang, Kaisiyuan Wang, Hang Zhou, Yingying Li, Haocheng Feng, Errui Ding, Yu Wu, and Jingdong Wang. Re-hold: Video hand object interaction reenactment via adaptive layout-instructed diffusion model. *CVPR*, 2025. 3
- [6] Haipeng Fang, Sheng Tang, Zhihao Sun, Ziyao Huang, Juan Cao, Fan Tang, and Yongdong Zhang. Moanimate: Bridging the motion-oriented latent representation gaps in human video animation. *IEEE Transactions on Circuits and Systems for Video Technology*, 36(4):5453–5466, 2026. 2
- [7] Qijun Gan, Yi Ren, Chen Zhang, Zhenhui Ye, Pan Xie, Xiang Yin, Zehuan Yuan, Bingyue Peng, and Jianke Zhu. Humandit: Pose-guided diffusion transformer for long-form human motion video generation. *arXiv preprint arXiv:2502.04847*, 2025. 2
- [8] Shijia Ge, Yinxin Zhang, Shuzhao Xie, Weixiang Zhang, Mingcai Zhou, and Zhi Wang. Vggt-dp: Generalizable robot control via vision foundation models, 2025. 2
- [9] Li Hu. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8153–8163, 2024. 1, 2
- [10] Li Hu, Guangyuan Wang, Zhen Shen, Xin Gao, Dechao Meng, Lian Zhuo, Peng Zhang, Bang Zhang, and Liefeng Bo. Animate anyone 2: High-fidelity character image animation with environment affordance. *arXiv preprint arXiv:2502.06145*, 2025. 3
- [11] Ziyao Huang, Fan Tang, Yong Zhang, Xiaodong Cun, Juan Cao, Jintao Li, and Tong-Yee Lee. Make-your-anchor: A diffusion-based 2d avatar generation framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6997–7006, 2024. 2
- [12] 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. 2, 3, 5
- [13] 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. 2, 5
- [14] Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. CoTracker: It is better to track together. 2023. 5
- [15] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2
- [16] 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
- [17] Gaojie Lin, Jianwen Jiang, Jiaqi Yang, Zerong Zheng, Chao Liang, Yuan Zhang, and Jingtuo Liu. Omnihuman-1: Rethinking the scaling-up of one-stage conditioned human animation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR)*, pages 13847–13858, 2025. 2
- [18] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022. 3
- [19] Yuxuan Luo, Zhengkun Rong, Lizhen Wang, Longhao Zhang, Tianshu Hu, and Yongming Zhu. Dreamactor-m1: Holistic, expressive and robust human image animation with hybrid guidance. *arXiv preprint arXiv:2504.01724*, 2025. 2
- [20] Yifang Men, Yuan Yao, Miaomiao Cui, and Liefeng Bo. Mimo: Controllable character video synthesis with spatial decomposed modeling. *arXiv preprint arXiv:2409.16160*, 2024. 3
- [21] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023. 3
- [22] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 5
- [23] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. 4
- [24] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman

- Rädlle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 5
- [25] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, Zhaoyang Zeng, Hao Zhang, Feng Li, Jie Yang, Hongyang Li, Qing Jiang, and Lei Zhang. Grounded sam: Assembling open-world models for diverse visual tasks, 2024. 5
- [26] Maximilian Seitzer. pytorch-fid: FID Score for PyTorch. <https://github.com/mseitzer/pytorch-fid>, 2020. Version 0.3.0. 5
- [27] Zhelun Shen, Chenming Wu, Junsheng Zhou, Chen Zhao, Kaisiyuan Wang, Hang Zhou, Yingying Li, Haocheng Feng, Wei He, and Jingdong Wang. idit-hoi: Inpainting-based hand object interaction reenactment via video diffusion transformer, 2025. 3
- [28] Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3, 2025. 2
- [29] Shuyuan Tu, Qi Dai, Zhi-Qi Cheng, Han Hu, Xintong Han, Zuxuan Wu, and Yu-Gang Jiang. Motioneditor: Editing video motion via content-aware diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7882–7891, 2024. 2
- [30] Shuyuan Tu, Zhen Xing, Xintong Han, Zhi-Qi Cheng, Qi Dai, Chong Luo, and Zuxuan Wu. Stableanimator: High-quality identity-preserving human image animation. *arXiv preprint arXiv:2411.17697*, 2024. 2
- [31] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges, 2019. 5
- [32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3
- [33] 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. 2
- [34] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer, 2025. 2, 4
- [35] Lizhen Wang, Zhurong Xia, Tianshu Hu, Pengrui Wang, Pengfei Wei, Zerong Zheng, Ming Zhou, Yuan Zhang, and Mingyuan Gao. Dreamactor-h1: High-fidelity human-product demonstration video generation via motion-designed diffusion transformers, 2025. 3
- [36] Tan Wang, Linjie Li, Kevin Lin, Yuanhao Zhai, Chung-Ching Lin, Zhengyuan Yang, Hanwang Zhang, Zicheng Liu, and Lijuan Wang. Disco: Disentangled control for realistic human dance generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9326–9336, 2024. 2
- [37] 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. 2, 5
- [38] Haoyu Wu, Diankun Wu, Tianyu He, Junliang Guo, Yang Ye, Yueqi Duan, and Jiang Bian. Geometry forcing: Marrying video diffusion and 3d representation for consistent world modeling. *arXiv preprint arXiv:2507.07982*, 2025. 2
- [39] Jingqiao Xiu, Fangzhou Hong, Yicong Li, Mengze Li, Wentao Wang, Sirui Han, Liang Pan, and Ziwei Liu. Egotwin: Dreaming body and view in first person. *arXiv preprint arXiv:2508.13013*, 2025. 4
- [40] Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Hanshu Yan, Jia-Wei Liu, Chenxu Zhang, Jiashi Feng, and Mike Zheng Shou. Magicanimate: Temporally consistent human image animation using diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1481–1490, 2024. 2
- [41] Ziyi Xu, Ziyao Huang, Juan Cao, Yong Zhang, Xiaodong Cun, Qing Shuai, Yuchen Wang, Linchao Bao, and Fan Tang. Anchorcrafter: Animate cyber-anchors selling your products via human-object interacting video generation. *IEEE Transactions on Visualization and Computer Graphics*, 2026. 2, 3, 5
- [42] Zihui Xue, Mi Luo, Chen Changan, and Kristen Grauman. Hoi-swap: Swapping objects in videos with hand-object interaction awareness. *arXiv preprint arXiv:2406.07754*, 2024. 3
- [43] Zhendong Yang, Ailing Zeng, Chun Yuan, and Yu Li. Effective whole-body pose estimation with two-stages distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4210–4220, 2023. 5
- [44] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3836–3847, 2023. 2
- [45] 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, 5
- [46] 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