UNIHOI: VIDEO AND MOTION CO-GENERATION VIA A UNIFIED LATENT SPACE FOR HAND–OBJECT INTERACTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Hand-object interaction (HOI) generation aims to synthesize physically plausible videos and 3D motion. Motion generation models capture structured 3D interaction dynamics but are limited by scarce high-quality data, while video foundation models learn rich visual priors yet lack explicit modeling of 3D geometric structure. To leverage these complementary strengths, recent works have explored video and motion co-generation frameworks that jointly model visual priors and 3D structural constraints. They often employ direct attention-based interactions to capture cross-modal dependencies between video and motion representations. However, these methods overlook the spatiotemporal granularity mismatch between heterogeneous representations, hindering fine-grained cross-modal correspondence and degrading native representations. Therefore, our key insight is to introduce a modality-shared yet interaction-comprehensive latent representation for cross-modal interaction. Motivated by this, we propose **UniHOI**, a unified video and motion co-generation framework that preserves modality-specific generation while constructing a unified latent space to abstract away modality-specific organization and retain complementary interaction information. Specifically, we organize this space along two complementary dimensions, visual dynamics and 3D structure, thereby establishing fine-grained spatiotemporal correspondence and enabling collaborative modeling of rich visual priors and structured 3D information. Furthermore, we introduce cross-modal rhythm alignment to promote consistent temporal dynamics between video and motion generation. Extensive experiments validate our superiority over state-of-the-art.

1 INTRODUCTION

Hand–object interaction (HOI) generation plays an important role in character animation Xu et al. (2023); Corona et al. (2024), digital human creation Wang et al. (2025) and dexterous robotic manipulation Kim et al. (2025); Goswami et al. (2026), requiring faithful generation of complex hand–object behaviors. Existing studies mainly pursue two directions: motion generation and video generation. Motion generation methods Zhou et al. (2025); Bensabath et al. (2026); Cha et al. (2024); Li et al. (2025a) aim to synthesize structured hand–object pose and trajectory sequences from given task descriptions, relying on high-quality 3D interaction data with accurate annotations collected by multi-view cameras or motion-capture systems. Due to the limited diversity of object types and interaction patterns in laboratory-collected datasets, motion generation models often exhibit inaccurate hand–object contacts (e.g., hand–object interpenetration and missing contacts) and struggle to generalize to unseen objects and interactions.

Recently, large-scale video foundation models Yang et al. (2025); Team et al. (2025); Kong et al. (2025) based on Diffusion Transformers (DiT) have shown promising capabilities in modeling visual appearance and temporal dynamics, typically generating with spatiotemporally compressed representations. For HOI video generation Chen et al. (2026b); Gao et al. (2026); Ren et al. (2026); Pang et al. (2025), existing methods inject 2D guidance signals (e.g., masks, skeleton maps, and optical flow) into the diffusion model to constrain structural consistency in the image space. However, these pixel-level video generation models still lack explicit modeling of underlying 3D geometric structures and physical constraints, often resulting in distorted hand–object geometry, implausible contact

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

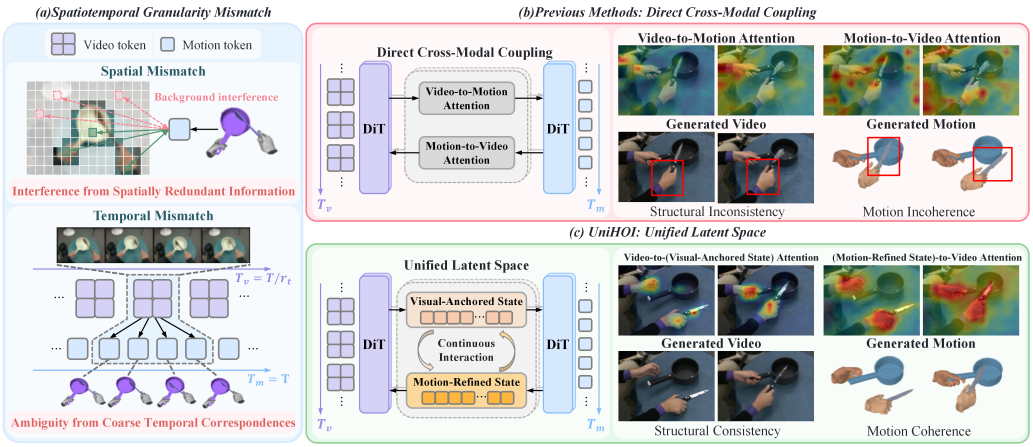


Figure 1: (a) Native video and motion representations exhibit spatiotemporal granularity mismatch. (b) Direct coupling of heterogeneous representations leads to diffuse bidirectional attention and coarse cross-modal correspondence, resulting in degraded visual interactions and inaccurate 3D motion. (c) UniHOI establishes fine-grained cross-modal correspondence within a unified latent space through complementary visual-anchored and motion-refined states, yielding interaction-focused attention and coherent video and motion co-generation.

relationships, and physically unrealistic motion. Overall, motion and video generation model offer complementary priors: motion generation models capture structured 3D interactions dynamics but are limited by scarce high-quality data, whereas video generation models learn rich visual and interaction priors from large-scale videos but lack explicit modeling of 3D structure. These strengths naturally motivate the co-generation of videos and motion, allowing rich visual priors and explicit structural constraints to be jointly modeled, thereby enhancing generalization in motion generation and the 3D structural awareness in video generation. However, due to high-dimensional hand articulation, complex hand–object occlusions, and rapidly changing contact states, jointly generating structurally consistent HOI videos and physically plausible motion remains challenging.

Existing works have explored video and motion co-generation for human animation Zhao et al. (2026); Xiu et al. (2026), hand–object interaction Sun et al. (2026); Dang et al. (2025), and robotic manipulation Bi et al. (2025); Li et al. (2025b). These methods mainly follow two architectural paradigms: either Ye et al. (2026); Kim et al. (2026) employing a shared backbone for joint denoising generation, or preserving modality-specific branches Bi et al. (2025); Yang et al. (2026) for separate denoising generation. To learn complementary dependencies between video and motion representations, they typically establish direct cross-modal interactions through attention-based mechanisms. However, these methods overlook **the inherent spatiotemporal granularity mismatch between heterogeneous representations**, resulting in interference from spatially redundant information and ambiguity arising from coarse temporal correspondences, which may further degrade their native representations. This mismatch manifests differently across spatial and temporal dimensions. **Spatially**, dense video representations jointly encode local interaction cues and global background information, whereas compact motion representations focus on 3D interaction dynamics. Consequently, redundant background visual tokens can dilute interaction-relevant visual cues, making it difficult for cross-modal interaction to focus on critical hand-object regions, as illustrated in Fig. 1. **Temporally**, video representations integrate visual information within local temporal windows for computational efficiency, whereas motion representations preserve frame-level states for generation accuracy. Consequently, temporal aggregation can reduce the discriminability of interaction-relevant visual dynamics, resulting in ambiguous temporal associations.

Overall, despite their distinct information densities, representation structures, and spatiotemporal granularities, video and motion representations describe the same underlying physical interaction process from distinct yet complementary modality-specific perspectives. Direct interaction between their native representations forces cross-modal correspondence to be learned across differently organized representation spaces, where mismatched spatiotemporal granularities make fine-grained alignment difficult to establish. Therefore, our key insight is to introduce a modality-shared yet

interaction-comprehensive shared latent representation for cross-modal interaction. It abstracts away modality-specific organization while retaining the complementary interaction information contributed by each modality. By modeling cross-modal dependencies within this shared latent, video and motion representations can establish spatiotemporal correspondence beyond the constraints of native granularities, enabling collaborative modeling of visual priors and 3D structural information.

Motivated by this insight, we propose **UniHOI**, a unified framework for jointly generating HOI videos and 3D motion. We retain modality-specific generation branches while introducing a unified latent space for cross-modal interaction that is modality-shared and interaction-comprehensive. Along the two dimensions of visual dynamics and 3D structure, we organize the unified latent space into two structured states, preserving a comprehensive characterization of the underlying physical interaction dynamics. Specifically, the visual-anchored state focuses on interaction-relevant and temporally discriminative visual dynamics, establishing a visual grounding for cross-modal spatiotemporal correspondence. The motion-refined state incorporates structured motion dynamics, enabling explicit modeling of 3D geometric structure with in the latent space. Through the progressive interaction between the two states, complementary visual priors and 3D structural information continuously reinforce each other within the unified latent space, enabling collaborative modeling. Furthermore, we introduce a cross-modal rhythm alignment mechanism to promote consistent temporal dynamics between video and motion generation.

Compared with existing methods, UniHOI achieves consistent improvements across TI2V generation, pose-controlled video generation, and motion generation on unseen objects. In particular, under the unseen-tool generalization setting, UniHOI reduces the average relative motion error by 43.1% and 22.9% compared with dedicated hand–object motion generation and video–motion co-generation methods, respectively. Our main contributions are summarized as follows:

(1) We propose UniHOI, a unified video–motion co-generation framework that enable complementary visual and 3D motion priors to be collaboratively modeled within a single model.

(2) We introduce a modality-shared and interaction-comprehensive unified latent space to mitigate spatiotemporal misalignment across heterogeneous modalities and facilitate collaborative modeling of rich visual priors and 3D structural information.

(3) Extensive experiments and ablation studies demonstrate the effectiveness of our method, achieving superior performance with temporally coherent video generation and accurate motion generation compared to existing state-of-the-art approaches.

2 Related Work

2.1 Motion Generation

Text-guided hand–object motion generation typically employs diffusion or flow-matching models Lipman et al. (2023) to synthesize 3D interaction sequences. Existing methods improve interaction plausibility by explicitly modeling hand–object contact, decomposing grasping and manipulation stages Cha et al. (2024); Song et al. (2026); Christen et al. (2024), or learning compact interaction representations Li et al. (2025a); Chen et al. (2026c). However, limited object and interaction diversity in laboratory-captured datasets still constrains their generalization.

2.2 HOI Video Generation

Recent diffusion models can synthesize realistic hand–object interaction videos Gao et al. (2026); Akkerman et al. (2025); Zhao et al. (2025); Sun et al. (2026), with structural consistency commonly enhanced through 2D guidance to improve structural consistency such as masks, keypoints, depth, and normals. More recent methods Chen et al. (2026a); Zhang et al. (2026) further incorporate 3D point clouds or joint representations to strengthen structural awareness. However, without explicit 3D learning objectives, externally provided 3D cues are difficult to translate into intrinsic 3D structure and motion awareness in video representations.

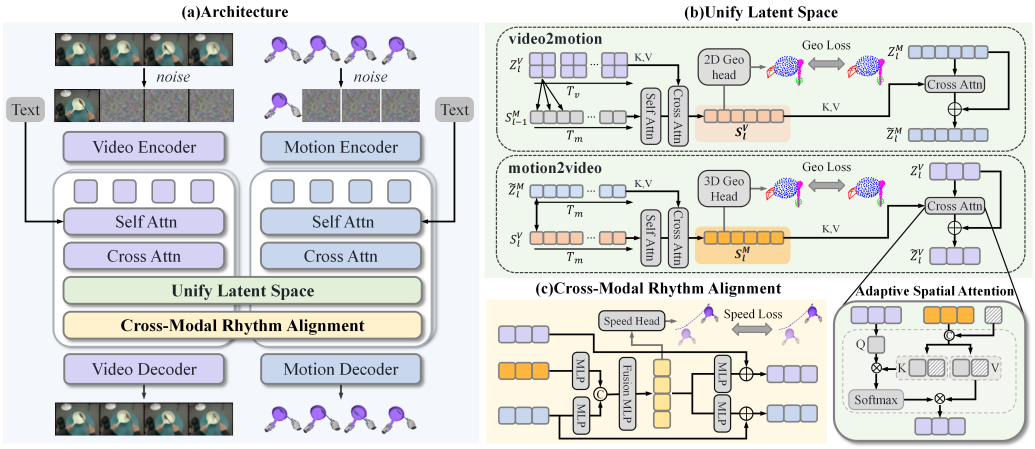


Figure 2: (a) Architecture of our method. (b)The Unified Latent Space is organized through a visual-anchored state and a motion-refined state, which progressively coordinate complementary visual priors and 3D structural information to establish fine-grained correspondence between heterogeneous representations. (c) The Cross-Modal Motion Rhythm Alignment module aligns the temporal dynamics of video and motion representations, promoting rhythm consistent generation.

2.3 Video and Motion Co-Generation

Video–motion co-generation has been explored in human motion Yang et al. (2026); Zhao et al. (2026); Wang et al. (2026), robotic manipulation Bi et al. (2025); Li et al. (2025b), and hand–object interaction Sun et al. (2026); Dang et al. (2025). Existing approaches generally follow two paradigms. One Ye et al. (2026); Kim et al. (2026) adapts motion representations to the video generative space and jointly models both modalities within a shared backbone; the former Yang et al. (2026); Fan et al. (2026); Yuan et al. (2026) jointly processes video and action tokens, whereas the latter represents actions as latent frames in the video diffusion space. The other methods Xiu et al. (2026); Yang et al. (2026) retains modality-specific generation branches and enables cross-modal interaction through attention. Notably, Dang et al. (2025) further introduces explicit 3D interaction diffusion and feeds structural predictions back into video generation. Despite their architectural differences, these methods largely establish cross-modal dependencies directly between heterogeneous native representations, with limited consideration of their spatiotemporal granularity mismatch, thereby hindering fine-grained cross-modal correspondence.

3 Method

Problem Definition. Given an initial hand-object interaction observation V_0, the corresponding initial 3D hand-object state M_0, and a text condition $\mathcal{C}$, our goal is to jointly generate the temporally synchronized future video sequence $V_{1:T}$ and 3D bimanual hand-object motion sequence $M_{1:T}$.

Architecture Overview Fig. 2 illustrates the overall architecture of our method. We employ modality-specific video and motion branches for joint generation, as detailed in Sec. 3.1. We further introduce a modality-shared yet interaction-comprehensive unified latent space for cross-model intraction, where visual-anchored and motion-refined states establish fine-grained cross-modal correspondence, as detailed in Sec. 3.2. Finally, cross-modal rhythm alignment promotes temporally consistent video–motion generation, as detailed in Sec. 3.3.

3.1 Synchronized Video-Motion Diffusion

Parameterized Hand-Object Motion Representation. We represent each hand using the MANO model Romero et al. (2022), $H_t = [\theta_t; \beta; r_t; t_t] \in \mathbb{R}^{109}$, including articulated pose, shape, global rotation, and translation, while each manipulated object is represented by its rigid pose $O_t = [r_t^o; t_t^o] \in \mathbb{R}^9$. Global hand and object motions are expressed relative to the first frame to emphasize local interaction dynamics. The resulting sequence is $M_{1:T} = [H_{1:T}^L, H_{1:T}^R, O_{1:T}^L, O_{1:T}^R] \in \mathbb{R}^{T\times 236}$. Different

motion components are projected into semantic motion tokens using dedicated MLPs, with the first-frame motion prepended as conditioning, yielding $Z^M \in \mathbb{R}^{(T+1)\times N_m \times C_m}$.

Modality-Specific Dual Diffusion. We preserve modality-specific representations using separate video and motion DiTs Peebles & Xie (2023). The video branch adopts Wan2.2 Wan et al. (2025) and operates on temporally compressed VAE latents $Z^V \in \mathbb{R}^{T_v \times H' \times W' \times C_v}$, while the motion branch operates on frame-level tokens $Z^M \in \mathbb{R}^{(T+1)\times N_m \times C_m}$ to retain fine-grained motion dynamics. Both branches share text conditioning and are jointly trained with Flow Matching Lipman et al. (2023):

$$\mathcal{L}_{\mathrm{FM}} = \mathbb{E}\left[\left\|v_\theta^V - v^V\right\|_2^2 + \lambda_m \left\|v_\theta^M - v^M\right\|_2^2\right],$$

where video and motion use independently sampled flow timesteps t_v and t_m.

3.2 Unified Latent Space

Existing co-generation methods directly couple heterogeneous video and motion representations, making fine-grained correspondence difficult to establish under their mismatched spatiotemporal granularities. Instead, we model their complementary information within a unified latent space that abstracts away modality-specific organization. The latent space is organized along two dimensions of visual dynamics and 3D structure, represented by a visual-anchored state and a motion-refined state. Through progressive interaction, these states establish fine-grained cross-modal correspondence and enable collaborative modeling of visual priors and structured information.

Complementary Latent States. To characterize visual and structure information within the unified latent space, we construct two latent states: a **visual-anchored state** and a **motion-refined state**. The visual-anchored state focuses on interaction-relevant and temporally discriminative visual dynamics, providing a visual basis for cross-modal spatiotemporal correspondence. The motion-refined state further incorporates motion information, enabling explicit modeling of 3D structure.

Specifically, we hierarchically update the two states across backbone layers. At the first layer, learnable frame-level queries S_{init}, which is aligned with the temporal resolution of motion representations, aggregate interaction-relevant visual dynamics from dense video latents to form S_l^V. This compact state suppresses redundant background information and propagates spatially and temporally aligned visual priors to the motion representation Z^M:

$$S_l^V = \mathrm{CrossAttn}\left(S_{\mathrm{init}}, Z_l^V, Z_l^V\right), \qquad \widetilde{Z}_l^M = Z_l^M + \mathrm{CrossAttn}\left(Z_l^M, S_l^V, S_l^V\right).$$

where $\widetilde{Z}_l^M$ denotes the updated motion representation at the l-th layer.

Building upon the interaction-relevant visual context established by the visual-anchored state S_l^V, we further refine it with structured 3D motion information to form the motion-refined state S_l^M. This process equips the unified latent space with more explicit 3D structural constraints while preserving the visual context, thereby strengthening cross-modal correspondence. Subsequently, S_l^M further refines the video representation, enhancing the awareness of 3D geometry structure:

$$S_l^M = S_l^V + \mathrm{CrossAttn}\left(S_l^V, \widetilde{Z}_l^M, \widetilde{Z}_l^M\right), \qquad \widetilde{Z}_l^V = Z_l^V + \mathrm{CrossAttn}\left(Z_l^V, S_l^M, S_l^M\right).$$

where $\widetilde{Z}_l^V$ denotes the refined video representation at the l-th layer. The motion-refined state S_l^M is then propagated to the next layer as the input state for form S_{l+1}^{V}, allowing visual dynamics and 3D structural information to continuously evolve across layers. Importantly, the visual-anchored and motion-refined states are not independent latent representations, but rather two successive refinement stages within the unified latent space. Through this continuous evolution, the unified latent space progressively refines cross-modal correspondence between video and motion representations.

Sparse Geometry Supervision. Although the unified latent space progressively establishes cross-modal associations, attention-based state interactions alone are insufficient to preserve fine-grained visual sensitivity in the visual-anchored state and explicit 3D structural awareness in the motion-refined state. For the visual-anchored state S^V, preserving temporal discriminability and spatial selectivity is critical. Temporal compression may obscure subtle frame-level dynamics, while uneven motion saliency can bias attention toward highly dynamic yet interaction-irrelevant regions. For S^M, implicit attention cannot explicitly enforce 3D hand-object structural relationships.

To complement implicit attention with explicit geometric constraints, we impose sparse 2D and 3D supervision on the visual-anchored and motion-refined states according to their semantic roles. Specifically, We use lightweight MLP heads D_{2D} and D_{3D} decode geometric structures from S^V and S^M, respectively. For 2D supervision, D_{2D} predicts hand joints and object keypoints in image space. The 2D coordinates are supervised by:

$$\hat{P}_{2D} = D_{2D}(S^V) = \left[\hat{P}_{H,2D}, \hat{P}_{O,2D}\right], \qquad \mathcal{L}_{\text{geo}}^{2D} = \left\|\hat{P}_{2D} - P_{2D}\right\|_1 .$$

where $\hat{P}_{H,2D}$ and $\hat{P}_{O,2D}$ denote the predicted 2D keypoints, and P_{2D} denotes ground-truth coordinates. Similarly, D_{3D} predicts sparse 3D hand-object geometry from S^M and is supervised with the same coordinate regression objective. The 3D coordinates are represented relative to the wrist root to emphasize relative hand-object structure.

The 2D supervision provides S^V with frame-level image-space anchors, while the 3D supervision imposes explicit structural constraints on S^M. Together, they enhance the latent states with fine-grained interaction dynamics and 3D structure, providing reliable complementary constraints for the unified latent space.

Adaptive Spatial Attention. To enhance the 3D structural awareness of video representations, we use the motion-refined state S^M to provide structural information to the video branch. However, such 3D constraints are primarily relevant to localized hand-object interaction regions rather than the entire visual field. Standard cross-attention requires every video token to distribute normalized attention over S^M, causing even interaction-irrelevant background regions to aggregate structural information. This may introduce unnecessary structural perturbations and compromise visual fidelity. To address this issue, we introduce adaptive spatial attention to selectively integrate 3D structural information into video representations according to spatial interaction relevance.

Specifically, at the l-th interaction layer, we augment the structural cross-attention space of the motion-refined state S_l^M with a fixed null state, which serves as an explicit zero-response reference rather than introducing additional structural semantics. The corresponding key–value sets are

$$K_l = [W_k S_l^M; K], \qquad V_l = [W_v S_l^M; V], \qquad K = V = 0.$$

For each visual token $Z_{l,i}^V$, the augmented attention distribution is

$$\boldsymbol{\alpha}_{l,i} = \text{Softmax}([s_{l,i1}, \ldots, s_{l,iN}, b]), \qquad s_{l,ij} = \frac{\langle W_q Z_{l,i}^V, W_k S_{l,j}^M \rangle}{\sqrt{d}}.$$

where $s_{l,ij}$ measures the structural relevance between the i-th visual token and the j-th motion-refined token at layer l, and b is the fixed logit of the null state. By jointly normalizing structural and null responses, interaction-relevant regions retain stronger structural responses, while irrelevant regions can shift attention mass toward the null state. Since $V = 0$, the video representation is updated as

$$\widetilde{Z}_{l,i}^V = Z_{l,i}^V + W_o \left(\sum_{j=1}^{N} \alpha_{l,ij} W_v S_{l,j}^M \right).$$

Accordingly, 3D structural information within the unified latent space can adaptively inject video representations according to spatial interaction relevance, thereby enhancing their awareness of 3D structural relationships in localized hand-object interactions.

3.3 Cross-Modal Rhythm Alignment.

Although cross-modal interaction establishes frame-wise correspondence between video and motion, it does not explicitly constrain their temporal evolution, which may still result in inconsistent motion speed and temporal rhythm. We therefore introduce cross-modal rhythm alignment to capture shared dynamic patterns from local temporal variations. Specifically, we compute first-order temporal differences of the motion-refined state S^M and the updated motion representation $\widetilde{Z}^M$, and fuse them through lightweight temporal encoders into a shared rhythm representation:

$$R_t = \mathcal{F}\Big(E_S(S_t^M - S_{t-1}^M), E_M(\widetilde{Z}_t^M - \widetilde{Z}_{t-1}^M)\Big).$$

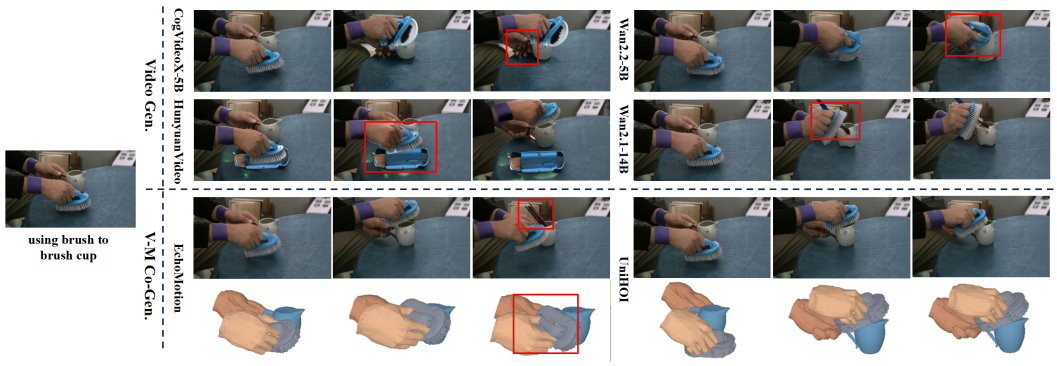


Figure 3: Qualitative comparison of TI2V generation and video-motion co-generation. UniHOI achieves coherent generation with structurally consistent videos and physically plausible 3D motion.

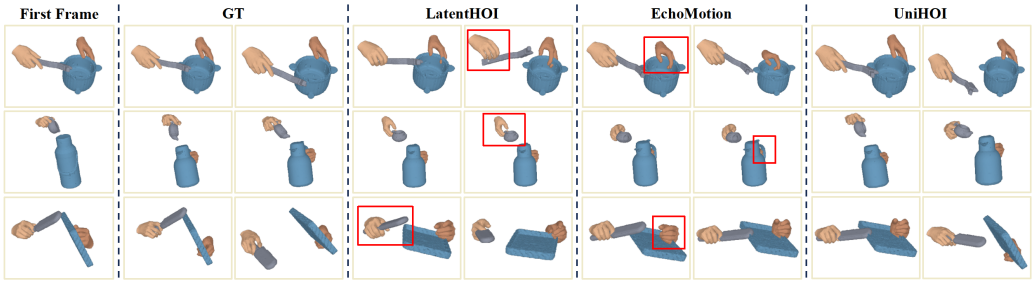


Figure 4: Qualitative comparison of 3D hand-object motion generation. UniHOI generates more physically plausible hand-object contacts and temporally coherent motions.

To endow R with explicit physical motion semantics, we further supervise it using ground-truth 3D wrist velocities. The shared rhythm representation is then projected back to both video and motion feature spaces through learnable gated residual updates, encouraging consistent motion speed and temporal rhythm during subsequent generation.

4 Expriments

4.1 Experimental Setup

Dataset. TACO dataset Liu et al. (2024) contains 2.5K bimanual tool-use sequences and 5.2M frames with multi-view videos and high-fidelity 3D hand–object motion annotations. It covers 20 object categories, 196 object models, 14 participants, and 15 interaction types. We report results on Test-2 for generalization to unseen tool geometries.

Evaluation Metrics. We evaluate video generation using PSNR and SSIM for reconstruction quality, together with four VBench metrics Huang et al. (2024) measuring subject consistency, background consistency, temporal smoothness, and motion dynamics. For 3D motion, we follow prior works Zhou et al. (2025); Li et al. (2025a) to evaluate hand and object motion accuracy and hand-object contact consistency. Detailed definitions are provided in the supplementary material.

Implementation Details. UniHOI is initialized from Wan2.2-TI2V-5B Wan et al. (2025), with both the video and motion DiTs consisting of 30 Transformer blocks. We adopt a two-stage training schedule, first training at 384×320 resolution and then fine-tuning at 480×704. The model is trained with an effective batch size of 8 on a single NVIDIA H200 GPU.

Table 1: Quantitative comparison of video generation results. Best and second-best results are marked in **bold** and underline, with green and blue highlights.

Method	Paradigm	PSNR ↑	SSIM ↑	Content Consistency		Motion Consistency	
				Subj. ↑	Bkg. ↑	TSmoo. ↑	Dyn. ↑
TI2V Generation							
CogVideoX-5B	Video Gen.	19.2501	0.6714	0.9389	0.9402	0.9492	0.9879
Wan2.2-5B	Video Gen.	19.4171	0.6827	0.9406	0.9521	0.9464	0.9817
Wan2.1-14B	Video Gen.	19.9302	0.6997	**0.9637**	**0.9649**	0.9805	0.6016
HunyuanVideo-13B	Video Gen.	19.9210	0.7129	0.9610	0.9603	0.9836	0.6834
EchoMotion (Wan2.2-5B)	V–M Co-Gen.	19.5029	0.6845	0.9483	0.9439	0.9512	0.9850
Ours (Wan2.2-5B)	**V–M Co-Gen.**	**19.9346**	**0.7148**	0.9635	0.9533	**0.9839**	**0.9892**
Pose-Controlled Video Generation							
VACE (Wan2.1-14B)	2D Pose Ctrl.	19.5923	0.6501	0.9495	0.9623	0.9845	0.9805
InterDyn (SVD)	2D Pose Ctrl.	18.4419	0.6006	0.9338	0.9635	0.9862	0.9749
JointControlVideo (Wan2.1-14B)	3D Pose Ctrl.	19.8886	0.7078	0.9511	**0.9821**	0.9931	0.9859
Ours (Wan2.2-5B)	**3D Pose Ctrl.**	**20.0342**	**0.7381**	**0.9592**	0.9741	**0.9954**	**0.9902**

4.2 Text to Video Generation

Baselines. We compare UniHOI with video foundation models, including CogVideoX-5B Yang et al. (2025), Wan2.2-5B Wan et al. (2025), Wan2.1-14B Wan et al. (2025), and HunyuanVideo-13B Kong et al. (2025), and further include EchoMotion Yang et al. (2026) as a video–motion co-generation baseline. Due to their higher training cost, Wan2.1-14B and HunyuanVideo-13B are evaluated using their publicly released pretrained checkpoints.

Quantitative and Qualitative Analysis. As shown in Tab. 1, UniHOI achieves stronger overall video generation performance. While some baselines achieve slightly higher content consistency under reduced motion variation, UniHOI preserves comparable visual consistency while supporting richer and more coherent interaction dynamics. Qualitative results further show that, although existing methods can generally maintain the global scene appearance, they still tend to produce hand deformations, object-geometry drift, and unstable hand–object relationships in fine-grained interaction regions. In contrast, UniHOI better preserves hand and object geometry across frames while maintaining substantial motion changes, resulting in more coherent and structurally consistent hand–object interactions.Additional qualitative analyses are provided in the supplementary material.

4.3 Motion Generation

Baselines. We compare UniHOI with LatentHOI Li et al. (2025a), a SOTA hand-object motion generation method, and EchoMotion Yang et al. (2026), a sSOTA video-motion co-generation baseline. We evaluate motion generation on the TACO Test-2 generalization split, designed to assess performance on tool geometries unseen during training.

Quantitative and Qualitative Analysis. As shown in Tab. 2, UniHOI consistently outperforms LatentHOI and EchoMotion, reducing the average relative motion error by **43.1% and 22.9%**. These results demonstrate that visual interaction priors and fine-grained cross-modal correspondence within the unified latent space improve the generalization of 3D interaction modeling to unseen tool geometries. Qualitatively, LatentHOI is prone to failed grasps, where the generated hand does not establish proper contact with the tool (Fig. 4, Col 5), while EchoMotion still exhibits hand-object interpenetration (Fig. 4, Col 6) and unstable contact relationships (Fig. 4, Col 7). In contrast, UniHOI better preserves the relative geometry among the hand, tool, and target object, producing temporally coherent motions that more faithfully follow the ground-truth interaction process.

4.4 Pose-Controlled Video Generation

Baselines. By jointly modeling video and 3D motion within a unified framework, UniHOI naturally supports pose-conditioned video generation in addition to TI2V generation. We compare UniHOI with SOTA pose-controlled video generation methods, including VACE Jiang et al. (2025), InterDyn Akkerman et al. (2025), and JointControlVideo Zhang et al. (2026), covering both 2D- and 3D-pose-guided paradigms. All baselines are fine-tuned under our experimental setting for fair.

Table 2: Quantitative comparison of motion generation results. Best and second-best results are marked in **bold** and underline, with green and blue highlights.

Method	Paradigm	Hand Motion				Object Motion		Contact Consistency		
		MPJPE ↓	PA-MPJPE ↓	MWTE ↓	MRE ↓	MWTE ↓	MRE ↓	IV ↓	ID ↓	CRE ↓
LatentHOI	Motion Gen.	8.77	1.06	10.01	0.67	12.15	0.84	6.73	1.13	1.85
EchoMotion	V–M Co-Gen.	5.04	0.62	5.93	0.32	9.10	0.63	7.43	1.39	2.89
Ours	**V–M Co-Gen.**	**4.39**	**0.56**	**4.40**	**0.25**	**7.74**	**0.50**	**5.90**	**1.06**	**1.47**

Table 3: Ablation study of the proposed components. Best and second-best results are marked in **bold** and underline, with green and blue highlights.

Variant	Video Quality		Hand Motion		Contact Consistency	
	PSNR ↑	SSIM ↑	MPJPE ↓	PA-MPJPE ↓	IV ↓	CRE ↓
w/o Video-Gen	–	–	6.0021	0.8192	8.1990	3.0427
w/o Motion-Gen	19.4126	0.6817	–	–	–	–
w/o ULS	19.4505	0.6808	4.6437	0.5961	6.9212	1.6349
w/o AGS	19.7136	0.7050	4.7119	0.5968	6.3334	1.8249
w/o ASA	19.3086	0.6978	4.5694	0.5603	5.9894	1.9053
w/o CMRA	19.4662	0.6812	4.6240	0.5810	6.0248	1.5328
Full UniHOI	**19.9346**	**0.7148**	**4.3942**	**0.5600**	**5.9021**	**1.4769**

Quantitative and Qualitative Analysis. As shown in Tab. 1, UniHOI achieves the best performance on all metrics except background consistency. Although JointControlVideo attains marginally higher background consistency, it remains weaker in interaction-centric visual quality and motion dynamics. In contrast, UniHOI better preserves scene fidelity while generating smooth and sufficiently dynamic hand–object interactions, demonstrating that the collaborative modeling of visual priors and structured 3D information provides more effective structural and dynamic constraints for fine-grained interaction generation. Qualitative comparisons are provided in the supplementary material.

4.5 Ablation Study

Tab. 3 analyzes the main components of UniHOI, where each variant is obtained by independently removing a single component from the full model. ULS, AGS, ASA, and CMRA denote Unified Latent Space, Auxiliary Geometry Supervision, Adaptive Spatial Attention, and Cross-Modal Rhythm Alignment, respectively. Removing either the video or motion generation branch degrades the other modality, validating the complementary roles of visual priors and structured 3D motion in joint generation. Removing ULS leads to clear degradation in both visual quality and motion consistency, demonstrating the importance of a shared interaction space for establishing fine-grained cross-modal correspondence. Ablating AGS substantially weakens motion generation, indicating that explicit geometric supervision strengthens 3D structural awareness in the latent states. In contrast, removing ASA causes a marked drop in visual quality, suggesting that spatial selectivity is critical for restricting structural guidance to interaction-relevant regions and avoiding unnecessary perturbations to irrelevant visual content. Finally, removing CMRA degrades video and motion generation, showing that frame-wise cross-modal interaction alone is insufficient to capture the non-uniform temporal dynamics of dexterous manipulation, while cross-modal rhythm alignment further regularizes their temporal evolution. Qualitative comparisons are provided in the supplementary material, where we further visualize the influence of removing AGS or ASA in the cross-modal attention patterns.

5 Conclusion

We presented **UniHOI**, a unified framework for joint HOI video and 3D motion generation. By introducing a unified latent space and cross-modal rhythm alignment, UniHOI enables effective collaboration between visual priors and structured 3D information despite their spatiotemporal granularity mismatch. Extensive experiments demonstrate consistent improvements in both structurally coherent video generation and physically plausible 3D motion.

REFERENCES

Rick Akkerman, Haiwen Feng, Michael J Black, Dimitrios Tzionas, and Victoria Fernández Abrevaya. InterDyn: Controllable Interactive Dynamics with Video Diffusion Models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12467–12479. IEEE, 2025.

Léore Bensabath, Mathis Petrovich, and Gül Varol. Text-Driven 3D Hand Motion Generation from Sign Language Data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23095–23105, 2026.

Hongzhe Bi, Hengkai Tan, Shenghao Xie, Zeyuan Wang, Shuhe Huang, Haitian Liu, Ruowen Zhao, Yao Feng, Chendong Xiang, Yinze Rong, Hongyan Zhao, Hanyu Liu, Zhizhong Su, Lei Ma, Hang Su, and Jun Zhu. Motus: A unified latent action world model, 2025.

Junuk Cha, Jihyeon Kim, Jae Shin Yoon, and Seungryul Baek. Text2HOI: Text-guided 3D Motion Generation for Hand-Object Interaction. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1577–1585. IEEE, 2024.

Mingjin Chen, Junhao Chen, Zhaoxin Fan, Yujian Lee, Zichen Dang, Lili Wang, Yawen Cui, Lap-Pui Chau, and Yi Wang. HVG-3D: Bridging Real and Simulation Domains for 3D-Conditional Hand-Object Interaction Video Synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15986–15997, 2026a.

Tianxiao Chen, Jintao Rong, Huajin Chen, Jingya Wang, Tao Zhou, Jiming Chen, and Qi Ye. MIMIC: Mask-Injected Manipulation Video Generation with Interaction Control. In *International Conference on Learning Representations*, 2026b.

Zerui Chen, Rolandos Alexandros Potamias, Shizhe Chen, Jiankang Deng, Cordelia Schmid, and Stefanos Zafeiriou. Ho-flow: Generalizable hand-object interaction generation with latent flow matching, 2026c. URL https://arxiv.org/abs/2604.10836.

Sammy Christen, Shreyas Hampali, Fadime Sener, Edoardo Remelli, Tomas Hodan, Eric Sauser, Shugao Ma, and Bugra Tekin. Diffh2o: Diffusion-based synthesis of hand-object interactions from textual descriptions. In *SIGGRAPH Asia 2024 Conference Papers*, pp. 1–11, 2024.

Enric Corona, Andrei Zanfir, Eduard Gabriel Bazavan, Nikos Kolotouros, Thiemo Alldieck, and Cristian Sminchisescu. Vlogger: Multimodal diffusion for embodied avatar synthesis, 2024.

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. *Advances in Neural Information Processing Systems*, 38:76689–76718, 2025.

Liaoyuan Fan, Zetian Xu, Chen Cao, Wenyao Zhang, Mingqi Yuan, and Jiayu Chen. Aim: Intent-aware unified world action modeling with spatial value maps. *arXiv preprint arXiv:2604.11135*, 2026.

Mingju Gao, Kaisen Yang, Huan-ang Gao, Bohan Li, Ao Ding, Wenyi Li, Yangcheng Yu, Jinkun Liu, Shaocong Xu, Yike Niu, Haohan Chi, Hao Chen, Hao Tang, Yu Zhang, Li Yi, and Hao Zhao. PAM: A Pose-Appearance-Motion Engine for Sim-to-Real HOI Video Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15954–15965, 2026.

Raktim Gautam Goswami, Amir Bar, David Fan, Tsung-Yen Yang, Gaoyue Zhou, Prashanth Krishnamurthy, Michael Rabbat, Farshad Khorrami, and Yann LeCun. World models for learning dexterous hand-object interactions from human videos, 2026.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Vbench: Comprehensive benchmark suite for video generative models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21807–21818. IEEE, 2024.

Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. VACE: All-in-One Video Creation and Editing. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17191–17202. IEEE, 2025.

Byungjun Kim, Taeksoo Kim, Junyoung Lee, and Hanbyul Joo. Dexterous world models, 2025.

Moo Jin Kim, Yihuai Gao, Tsung-Yi Lin, Yen-Chen Lin, Yunhao Ge, Grace Lam, Percy Liang, Shuran Song, Ming-Yu Liu, Chelsea Finn, and Jinwei Gu. Cosmos policy: Fine-tuning video models for visuomotor control and planning, 2026.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, Kathrina Wu, Qin Lin, Junkun Yuan, Yanxin Long, Aladdin Wang, Andong Wang, Changlin Li, Duojun Huang, Fang Yang, Hao Tan, Hongmei Wang, Jacob Song, Jiawang Bai, Jianbing Wu, Jinbao Xue, Joey Wang, Kai Wang, Mengyang Liu, Pengyu Li, Shuai Li, Weiyan Wang, Wenqing Yu, Xinchi Deng, Yang Li, Yi Chen, Yutao Cui, Yuanbo Peng, Zhentao Yu, Zhiyu He, Zhiyong Xu, Zixiang Zhou, Zunnan Xu, Yangyu Tao, Qinglin Lu, Songtao Liu, Dax Zhou, Hongfa Wang, Yong Yang, Di Wang, Yuhong Liu, Jie Jiang, and Caesar Zhong. Hunyuanvideo: A systematic framework for large video generative models, 2025.

Muchen Li, Sammy Christen, Chengde Wan, Yujun Cai, Renjie Liao, Leonid Sigal, and Shugao Ma. LatentHOI: On the Generalizable Hand Object Motion Generation with Latent Hand Diffusion. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17416–17425. IEEE, 2025a.

Shuang Li, Yihuai Gao, Dorsa Sadigh, and Shuran Song. Unified video action model, 2025b.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling, 2023.

Yun Liu, Haolin Yang, Xu Si, Ling Liu, Zipeng Li, Yuxiang Zhang, Yebin Liu, and Li Yi. Taco: Benchmarking generalizable bimanual tool-action-object understanding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21740–21751. IEEE, 2024.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12209–12219. IEEE, 2025.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182. IEEE, 2023.

Zhongwei Ren, Yunchao Wei, Xiao Yu, Guixun Luo, Yao Zhao, Bingyi Kang, Jiashi Feng, and Xiaojie Jin. VideoWorld 2: Learning Transferable Knowledge from Real-world Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 40569–40580, 2026.

Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *arXiv preprint arXiv:2201.02610*, 2022.

Mingyeong Song, Jungbin Cho, Jisoo Kim, Ananya Bal, Kartik Sharma, Youngjae Yu, Laszlo A Jeni, and Junhyug Noh. JointHOI: Jointly Generating Contact Maps Enhances Hand Object Interaction Generation. In *European Conference on Computer Vision*, pp. 97–114. Springer, 2026.

Zhihao Sun, Zhiying Du, Xitong Yang, and Zuxuan Wu. HandWorld: Hand-Centric Unified Video Action Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15976–15985, 2026.

Kling Team, Jialu Chen, Yikang Ding, Zhixue Fang, Kun Gai, Yuan Gao, Kang He, Jingyun Hua, Boyuan Jiang, Mingming Lao, Xiaohan Li, Hui Liu, Jiwen Liu, Xiaoqiang Liu, Yuan Liu, Shun Lu, Yongsen Mao, Yingchao Shao, Huafeng Shi, Xiaoyu Shi, Peiqin Sun, Songlin Tang, Pengfei Wan, Chao Wang, Xuebo Wang, Haoxian Zhang, Yuanxing Zhang, and Yan Zhou. Klingavatar 2.0 technical report, 2025.

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.

Boyuan Wang, Xiaofeng Wang, Chaojun Ni, Guosheng Zhao, Zhiqin Yang, Zheng Zhu, Muyang Zhang, Yukun Zhou, Xinze Chen, Guan Huang, Lihong Liu, and Xingang Wang. Humandreamer: Generating controllable human-motion videos via decoupled generation, 2025.

Ziyi Wang, Xinshun Wang, Shuang Chen, Yang Cong, and Mengyuan Liu. UniMotion: A Unified Framework for Motion-Text-Vision Understanding and Generation. In *European Conference on Computer Vision*, pp. 664–682. Springer, 2026.

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. In *International Conference on Learning Representations*, 2026.

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, 2023.

Yuxiao Yang, Hualian Sheng, Sijia Cai, Jing Lin, Jiahao Wang, Bing Deng, Junzhe Lu, Haoqian Wang, and Jieping Ye. EchoMotion: Unified Human Video and Motion Generation via Dual-Modality Diffusion Transformer. In *International Conference on Learning Representations*, 2026.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. Cogvideox: Text-to-video diffusion models with an expert transformer, 2025.

Seonghyeon Ye, Yunhao Ge, Kaiyuan Zheng, Shenyuan Gao, Sihyun Yu, George Kurian, Suneel Indupuru, You Liang Tan, Chuning Zhu, Jiannan Xiang, Ayaan Malik, Kyungmin Lee, William Liang, Nadun Ranawaka, Jiasheng Gu, Yinzhen Xu, Guanzhi Wang, Fengyuan Hu, Avnish Narayan, Johan Bjorck, Jing Wang, Gwanghyun Kim, Dantong Niu, Ruijie Zheng, Yuqi Xie, Jimmy Wu, Qi Wang, Ryan Julian, Danfei Xu, Yilun Du, Yevgen Chebotar, Scott Reed, Jan Kautz, Yuke Zhu, Linxi "Jim" Fan, and Joel Jang. World action models are zero-shot policies, 2026.

Tianyuan Yuan, Zibin Dong, Yicheng Liu, and Hang Zhao. Fast-wam: Do world action models need test-time future imagination? *arXiv preprint arXiv:2603.16666*, 2026.

Chenyangguang Zhang, Botao Ye, Boqi Chen, Alexandros Delitzas, Fangjinhua Wang, Marc Pollefeys, and Xi Wang. Controllable Egocentric Video Generation via Occlusion-Aware Sparse 3D Hand Joints. In *European Conference on Computer Vision*, pp. 668–687. Springer, 2026.

Chengfeng Zhao, Jiazhi Shu, Yubo Zhao, Tianyu Huang, Jiahao Lu, Zekai Gu, Chengwei Ren, Zhiyang Dou, Qing Shuai, and Yuan Liu. CoMoVi: Co-Generation of 3D Human Motions and Realistic Videos. *arXiv preprint arXiv:2601.10632*, 2026.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27683–27693. IEEE, 2025.

Bohan Zhou, Yi Zhan, Zhongbin Zhang, and Zongqing Lu. MEgoHand: Multimodal Egocentric Hand-Object Interaction Motion Generation. *Advances in Neural Information Processing Systems*, 38:49464–49490, 2025.

AI USE STATEMENT

In this work, we used generative AI tools to assist with language polishing and translation. All AI-assisted content, including experimental analysis and textual material, was carefully reviewed and verified by the authors against the actual implementation and experimental results. Generative AI was not used to generate or modify experimental results. The authors take full responsibility for all content, claims, and associated outputs presented in this work.

ETHICS STATEMENT

This work studies generative modeling of hand–object interaction using existing research datasets and does not involve new human-subject data collection. We use the datasets solely for research purposes and follow their respective usage protocols. As with general video and human-motion generation systems, the proposed model may potentially be misused to synthesize misleading visual or motion content. Our work is intended for research applications such as interaction modeling, digital humans, and robotic manipulation, and we encourage responsible use of the resulting models and generated content.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the model architecture, motion representation, training objectives, and cross-modal interaction mechanisms in the main paper. The supplementary material further specifies implementation details, evaluation metrics, baseline configurations, ablation settings, and additional qualitative results. These materials document the training and evaluation protocols used throughout our experiments to facilitate reproduction of the reported results.

A APPENDIX

B SUPPLEMENTARY MATERIAL

The supplementary material is organized as follows:

- **Sec. A: Video–Motion Co-Generation Analysis.** Qualitative comparison with EchoMotion, focusing on cross-modal consistency, hand–object contact stability, and coordinated video–motion dynamics.
- **Sec. B: Ablation Study.** Additional qualitative analyses and cross-modal attention visualizations for the proposed components.
- **Sec. C: Text-and-Image-to-Video (TI2V) Generation.** Additional qualitative comparisons for TI2V generation.
- **Sec. D: Pose-Controlled Video Generation.** Additional qualitative comparisons under pose-controlled video generation settings.
- **Sec. E: Evaluation Metrics.** Detailed definitions and formulations of the evaluation metrics.

B.1 VIDEO–MOTION CO-GENERATION ANALYSIS

Qualitative Comparison with EchoMotion. As shown in Fig. **??**, UniHOI maintains more consistent interaction dynamics between the generated video and 3D motion. In the left example, as the pitcher is lifted and displaced in the video, the corresponding 3D motion consistently reflects its pose change while preserving a stable hand–object configuration. In contrast, EchoMotion exhibits noticeable discrepancies between the visual dynamics and the predicted object trajectory, together with local structural instability. In the bimanual tool-manipulation example on the right, EchoMotion further suffers from contact drift and implausible tool–object geometry, whereas UniHOI preserves more stable hand–tool contacts and coordinated relative motion. These results demonstrate that UniHOI better synchronizes visual dynamics with structured 3D motion, yielding more coherent video–motion co-generation.

B.2 ABLATION STUDY

Tab. 3 analyzes the main components of UniHOI, where ULS, AGS, ASA, and CMRA denote Unified Latent Space, Auxiliary Geometry Supervision, Adaptive Spatial Attention, and Cross-Modal Rhythm Alignment, respectively. Fig. 6 further visualizes representative failure cases.

Complementary Video–Motion Generation. Removing either generation branch degrades the other modality, confirming their complementary roles in modeling the same interaction. As shown

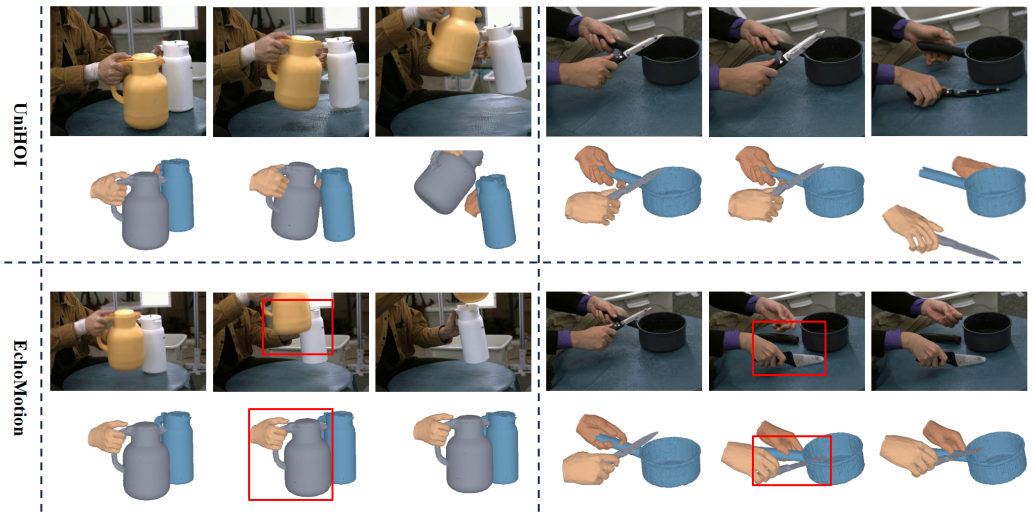


Figure 5: Qualitative comparison with EchoMotion. UniHOI better preserves cross-modal consistency, stable hand–object contacts, and coordinated video–motion dynamics.

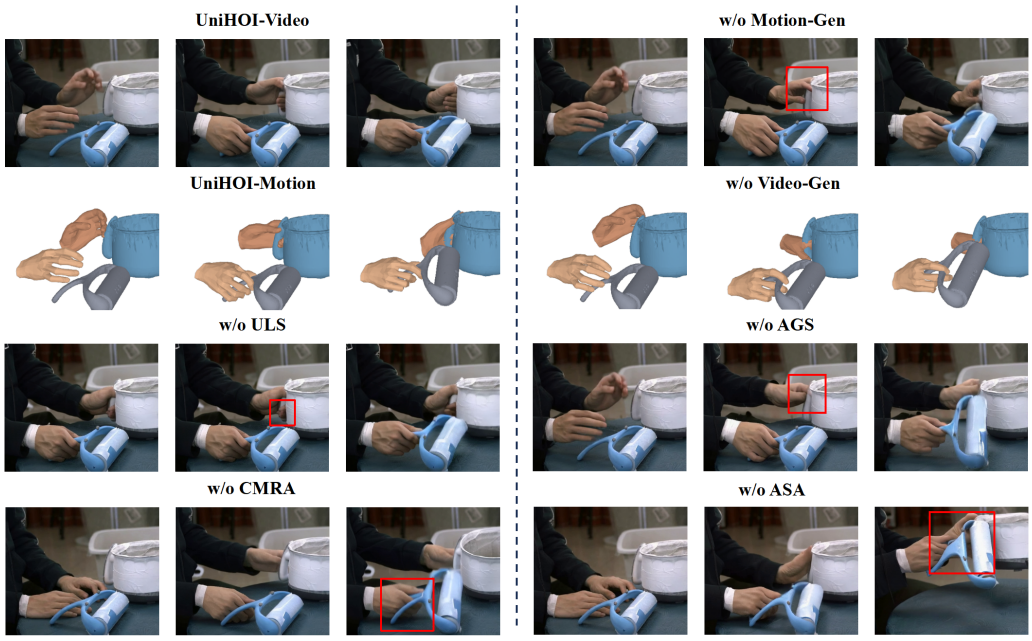


Figure 6: Qualitative comparison of Ablation Study.

in Fig. 6, without Motion-Gen, the generated video exhibits local hand deformation around the hand–object contact region, whereas removing Video-Gen leads to less accurate relative hand–object configurations in the generated 3D motion. These results show that visual interaction priors and structured 3D motion mutually regularize the two generation processes.

Unified Interaction Space. Removing ULS causes consistent degradation in both modalities and visibly introduces local geometric inconsistencies around the interaction region. Without AGS, the hand–object configuration becomes less structurally constrained, resulting in noticeable contact and geometry errors. Removing ASA further produces pronounced structural drift in later video frames, as structural information is no longer selectively concentrated on interaction-relevant regions. These

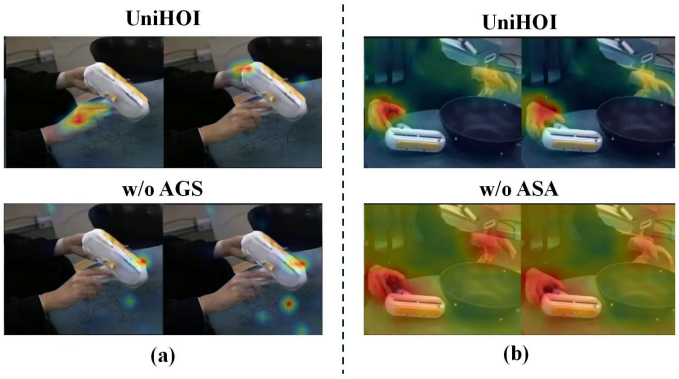


Figure 7: Visualization of cross-modal attention under AGS and ASA ablations. AGS improves geometric grounding, while ASA enhances spatial selectivity toward interaction-relevant regions.

observations demonstrate that the unified latent space requires both explicit geometric anchoring and spatially selective guidance to maintain reliable fine-grained cross-modal correspondence.

Cross-Modal Rhythm Alignment. Without CMRA, the overall spatial structure can still be preserved, but the hand–object evolution becomes less temporally stable, with visible inconsistencies in relative motion across frames. This indicates that frame-wise correspondence alone is insufficient for the non-uniform dynamics of dexterous manipulation; aligning local motion rhythms further improves the temporal coordination between video and motion.

Effect of Geometric Supervision and Spatial Selectivity on Cross-Modal Attention. As shown in Fig. 6, the full UniHOI consistently concentrates cross-modal attention around the hand, object, and their local interaction regions. Without AGS, the attention responses become less aligned with the underlying interaction geometry and exhibit spurious activations around object boundaries and background regions, indicating that explicit geometry supervision provides reliable structural anchors for cross-modal correspondence. In contrast, removing ASA causes the attention to spread broadly over interaction-irrelevant regions, showing that structural information is no longer selectively routed according to spatial relevance. The full model instead confines structural responses to interaction-centric regions while suppressing unnecessary background activations. These visualizations demonstrate that AGS and ASA play complementary roles in improving geometric grounding and spatial selectivity, respectively.

B.3 Text-and-Image-to-Video (TI2V) Generation.

Quantitative Evaluation. As shown in Tab. 1, UniHOI achieves the best PSNR, SSIM, temporal smoothness, and dynamic degree while maintaining strong content consistency. Notably, Wan2.2-14B Wan et al. (2025) and HunyuanVideo-13B Kong et al. (2025) exhibit strong content consistency and temporal smoothness but considerably weaker interaction dynamics, whereas CogVideoX-5B Yang et al. (2025) and EchoMotion Yang et al. (2026) achieve higher dynamic degrees at the cost of visual quality or temporal consistency. Compared with its Wan2.2-5B backbone, UniHOI consistently improves all metrics, demonstrating that jointly modeling structured 3D motion complements the visual prior with more effective interaction dynamics.

Qualitative Evaluation. As illustrated in Fig. 8, the qualitative comparisons further reveal three advantages. **(1) Appearance and structural fidelity.** Existing methods frequently exhibit hand or object deformation, appearance drift, and local artifacts during large motions, as highlighted by the red boxes. For example, CogVideoX-5B shows noticeable geometric distortion around the handheld tool during the interaction and appearance inconsistency of the target object in later frames, whereas UniHOI better preserves both geometry and visual identity throughout the sequence. **(2) Dynamic quality and temporal continuity.** Some large-scale video models tend to produce conservative motion, while more dynamic baselines often suffer from motion blur, structural distortion, or abrupt temporal changes. UniHOI generates sufficiently expressive interaction dynamics while maintaining smooth and coherent motion evolution. **(3) Hand–object interaction consistency.** Several baselines exhibit unstable hand–tool relationships as the interaction evolves; for instance, EchoMotion shows

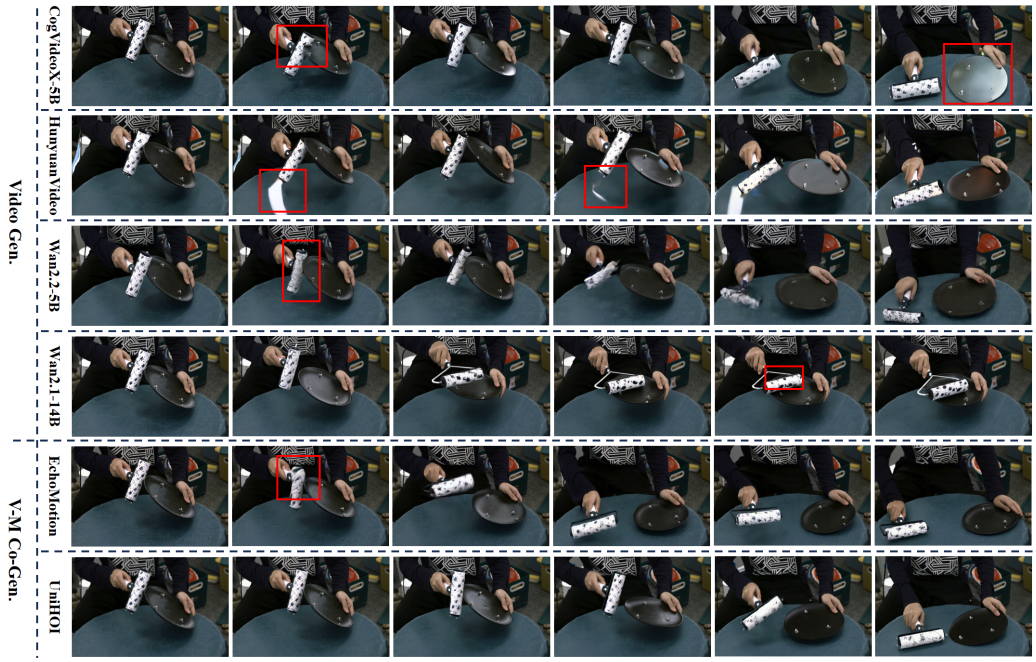


Figure 8: Qualitative comparison of TI2V generation. Red boxes highlight representative structural and appearance artifacts. UniHOI better preserves hand-object geometry, visual fidelity, and temporal coherence throughout the interaction.

local deformation around the hand–tool contact region, indicating that visually plausible motion does not necessarily preserve fine-grained interaction structure. In contrast, UniHOI maintains coherent relative relationships among the hand, tool, and target object throughout the sequence, demonstrating the benefit of structured 3D motion for fine-grained interaction modeling.

B.4 Pose-Controlled Video Generation

Quantitative Evaluation. As shown in Tab. 1, UniHOI achieves the best performance across all metrics except background consistency. Compared with 2D pose-controlled methods, UniHOI consistently improves visual quality, subject consistency, and motion consistency. Compared with the 3D pose-controlled JointControlVideo, UniHOI further improves subject consistency, temporal smoothness, and dynamic degree while maintaining competitive background consistency. These results indicate that jointly modeling structured 3D motion provides richer interaction constraints than treating pose merely as an external control signal.

Qualitative Evaluation. As illustrated in Fig. 9, the qualitative comparisons reveal three main advantages. **(1) Local structural fidelity.** 2D pose-guided methods are more prone to hand blur, local deformation, and object-geometry drift during complex interactions, as highlighted by the red boxes, whereas UniHOI better preserves local hand and object structures. **(2) Temporal structural consistency.** Even with stronger 3D pose guidance, existing methods may exhibit unstable appearance or local geometry across frames. UniHOI maintains more coherent structural evolution throughout the interaction. **(3) Stable hand–object relationships.** While existing methods primarily rely on pose as an external motion constraint, UniHOI jointly models video and structured 3D motion, allowing it to better preserve coherent relative relationships among the hand, tool, and target object over time

B.5 Evaluation Metric.

For video evaluation, we adopt four dimensions from VBench, including Subject Consistency and Background Consistency for content quality, and Motion Smoothness and Dynamic Degree for dynamic quality. We further report PSNR and SSIM to measure pixel and structural fidelity against the ground-truth videos. For motion evaluation, we assess the generated 3D interactions from three

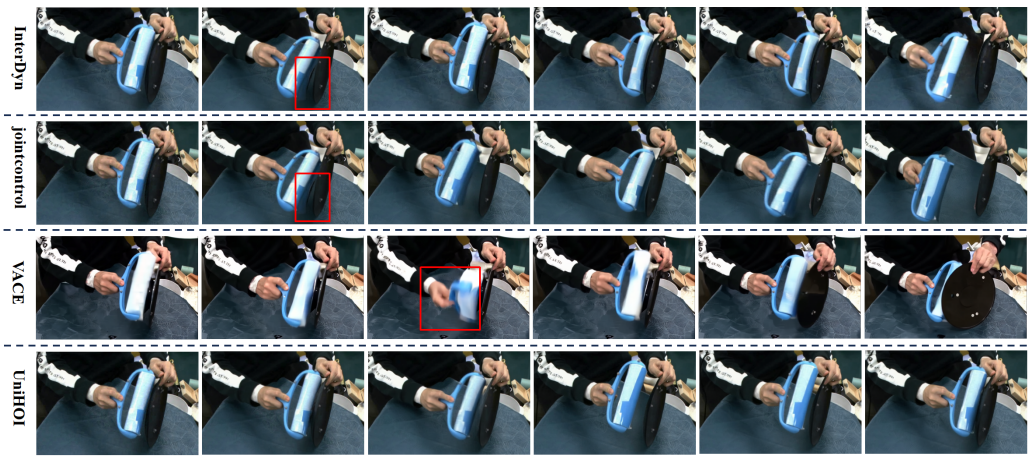


Figure 9: Qualitative comparison of Pose-Controlled Video Generation.

aspects: hand motion, object motion, and contact consistency. Following prior motion-generation works, hand motion is evaluated using MPJPE, PA-MPJPE, Mean Wrist Translation Error (MWTE), and Mean Rotation Error (MRE), measuring joint-position accuracy, globally aligned pose accuracy, wrist translation, and rotational accuracy, respectively. Object motion is evaluated using MWTE and MRE for translation and rotation accuracy. For hand–object contact consistency, we follow and report Interpenetration Volume (IV), Interpenetration Depth (ID), and Contact Ratio Error (CRE), which characterize geometric penetration and contact consistency between the generated hands and objects.