GEOMOTION-HOI: FROM 3D INTERACTION MOTION TO CONTROLLABLE HAND–OBJECT VIDEOS

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent advances in hand–object interaction (HOI) video synthesis have improved visual realism, while 3D conditioning has strengthened control over motion and spatial structure. However, future 3D interactions are rarely available at inference time, and their motion, contact, and object-relative relations must remain coordinated to provide reliable video control. We propose GEOMOTION-HOI, a two-stage framework connecting future 3D interaction generation with controllable egocentric video synthesis. Conditioned on text, an observed interaction state, hand context, and object geometry, the first stage jointly predicts global motion, part-level contact, and object-relative relations. A dynamic binding module uses multiscale temporal convolutions to model evolving hand–object relationships, followed by contact–relation fusion and contact-aware residual refinement to produce a coordinated future 3D sequence. The second stage renders this sequence into two complementary conditions: Tracking Video conveys surface-point correspondences across frames, while Geometry Video combines depth-based surface structure and occlusion with stable, first-frame-derived colors that associate hand and object regions over time. A trainable control branch injects both conditions into a pretrained image-to-video model, allowing the predicted sequence to guide motion and geometry while the observed first frame anchors scene appearance. Experiments on OakInk2 and TACO reduce motion FID by 40.9% and 22.7%, respectively, relative to the best evaluated baselines; video evaluations with given 3D conditions show improved fidelity, and ablations confirm that combining both conditions outperforms either alone across all five video metrics. GEOMOTION-HOI enables 3D-guided HOI video synthesis without an externally specified future trajectory or an exact target hand pose.

1 INTRODUCTION

Hand–object interaction (HOI) videos must preserve not only realistic appearance but also object geometry and hand–object spatial relationships throughout manipulation. Deformation or drift during grasping, rotation, and occlusion can undermine an otherwise visually plausible interaction. Such videos are valuable for interaction learning and data augmentation in embodied AI. Pretrained video models have improved visual synthesis (Yang et al., 2025; Team Wan et al., 2025), while structured conditions have strengthened motion and geometry control . However, geometry-guided generation from an initial observation requires both a plausible future 3D interaction and an effective representation for conveying that interaction to the video model.

Existing conditional video methods address different aspects of this problem. DaS uses Tracking Video to convey cross-frame point correspondences, but projected tracks do not explicitly describe the continuous surfaces between sampled points (Gu et al., 2025). HVG-3D supplements tracking with 3D point-cloud features, while requiring future control sequences obtained from real data or simulation (Chen et al., 2026). PAM connects pose, appearance, and video generation without requiring an externally supplied complete trajectory or a real first frame (Gao et al., 2026). Its pipeline nevertheless requires a specified target hand pose, and its pose-generation stage is formulated for one hand and one object. Its depth, segmentation, and keypoint conditions also do not explicitly identify corresponding surface points across frames. We instead consider egocentric generation with an observed first frame but no prescribed future trajectory or exact endpoint.

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.

Obtaining the future 3D interaction requires more than predicting isolated poses. Text2HOI, DiffH2O, and LatentHOI have advanced contact-guided generation, staged manipulation synthesis, and generalization through grasp priors (Cha et al., 2024; Christen et al., 2024; Li et al., 2025). JointHOI further jointly generates dynamic contact and motion, while Dex2HOI models bimanual two-object interactions through hand-relative object motion (Song et al., 2026; Pratikaki et al., 2026). Building on these advances, our goal is to coordinate global motion, evolving contact, and object-relative relationships into a renderable future sequence, and use that sequence to guide appearance-conditioned video synthesis.

We propose **GEOMOTION-HOI**, a two-stage framework connecting future 3D interaction generation with controllable video synthesis. Given text, an initial interaction state, hand context, and object geometry, the first stage jointly predicts motion, part-level contact, and object-relative relationships. Multiscale temporal convolutions model dynamic hand–object binding, while contact–relation fusion and contact-aware residual refinement coordinate the predictions and improve local geometry. The second stage renders the resulting sequence into two complementary conditions: Tracking Video provides surface-point motion correspondences, and Geometry Video combines depth-based surface structure with stable, first-frame-derived region colors. A trainable control branch injects both into a pretrained image-to-video model, allowing predicted 3D interactions to guide motion and geometry while the observed first frame anchors appearance.

GEOMOTION-HOI connects coordinated 3D interaction generation with complementary motion and geometry conditioning, enabling egocentric video synthesis without a predefined future trajectory or an exact target hand pose. Experiments on OakInk2 and TACO show the lowest motion FID and maximum penetration depth among compared generation methods on both datasets. Video evaluations with given 3D conditions and condition ablations support the benefits of combining tracking and geometry cues, while qualitative joint inference illustrates how predicted interactions guide the complete synthesis pipeline.

2 Related Work

2.1 Hand–Object Interaction Motion Generation

HOI motion generation must coordinate hand articulation, object trajectories, and evolving contact. Text2HOI (Cha et al., 2024) generates a contact prior from text and object geometry, then uses it to guide motion diffusion and hand refinement. DiffH2O (Christen et al., 2024) decomposes generation into grasping and manipulation, separating hand approach from subsequent coupled hand–object motion. LatentHOI (Li et al., 2025) decouples global temporal motion from fine-grained grasp synthesis through a latent grasp representation, improving generalization to unseen object geometry. More recently, JointHOI (Song et al., 2026) jointly generates motion and dynamic distance-based contact maps, with contact-guided sampling to improve geometric agreement. Dex2HOI (Pratikaki et al., 2026) coordinates bimanual single- and two-object interactions through hand-relative object representations, dual-stream diffusion, and motion fusion. These approaches establish useful interaction priors but target 3D motion rather than RGB video synthesis. We build on this direction by coordinating motion, contact, and relative relations into future trajectories that serve as video-generation conditions.

2.2 Video Generation Models

Video Diffusion Models (Ho et al., 2022) extend diffusion modeling to temporally coherent image sequences. VideoCrafter2 (Chen et al., 2024) and CogVideoX (Yang et al., 2025) further improve video synthesis through advances in training and spatiotemporal modeling. Foundation models such as Wan (Team Wan et al., 2025) and Cosmos (NVIDIA et al., 2025) provide powerful generative priors under textual and visual conditioning. Although these models learn rich appearance and motion statistics, their standard conditioning interfaces do not explicitly specify hand–object contact or object-relative 3D states. Our method therefore retains the pretrained visual prior while introducing structured interaction conditions through a trainable control branch.

2.3 Controllable Hand–Object Interaction Video Generation

Action-conditioned generation has progressed from predicting individual state transitions to synthesizing continuous interactions. CosHand (Sudhakar et al., 2025) uses an input image and hand-mask cues to synthesize a post-interaction image. InterDyn (Akkerman et al., 2025) instead conditions video diffusion on an initial image and a driving mask sequence, allowing the model to infer the resulting object dynamics. Structured controls provide a complementary route: DaS (Gu et al., 2025) conveys 3D surface-point motion through Tracking Video, while HVG-3D (Chen et al., 2026) combines point-cloud and tracking features through a 3D ControlNet. PAM (Gao et al., 2026) connects pose, appearance, and motion generation and controls video synthesis using depth, segmentation, and hand keypoints. These methods differ in how future controls are obtained and represented: HVG-3D takes supplied 3D sequences, whereas PAM's pose-generation stage requires a specified target hand pose. GEOMOTION-HOI instead predicts future interactions from text and initial conditions, then combines point correspondences with depth-colored surface renderings to guide video synthesis without a prescribed future trajectory or exact target hand pose.

3 Method

Given text y, an observed hand–object state S_0, hand context H, object geometry G, and an egocentric RGB frame I_0, we generate future interaction videos without a predefined 3D trajectory or target hand pose. As shown in Figure 1, we first generate and refine a 3D interaction sequence, then use it to guide video synthesis.

3.1 3D Hand–Object Interaction Generation

Joint generation. We represent future interactions as $X = (M, P, R)$: M contains hand-root, hand-articulation, and object poses; P describes hand proximity to sampled object-surface regions; and R encodes frame-wise increments of object displacement relative to hand-surface anchors. Contact targets derive from surface distances, and relative increments are integrated from S_0.

A Transformer predicts $\widehat{X} = F_\theta(X_\tau, \tau; y, S_0, H, G)$ from $X_\tau = (1-\tau)\epsilon + \tau X$, with Gaussian noise ϵ. Let $\mathcal{C}$ and $\mathcal{N}$ denote contact and non-contact locations defined by ground-truth distances, and $\langle\cdot\rangle_{\mathcal{A}}$ the normalized average over valid entries in $\mathcal{A}$. Joint supervision is

$$\mathcal{L}_{\text{joint}} = \mathbb{E}_{\tau,\epsilon}[\langle\|\widehat{M} - M\|_2^2\rangle + \langle\|\widehat{P} - P\|_2^2\rangle_{\mathcal{C}} + \langle\|\widehat{P} - P\|_2^2\rangle_{\mathcal{N}} + \langle\|\widehat{R} - R\|_2^2\rangle]. \tag{1}$$

For the fused trajectory, we supervise temporal changes in hand-joint positions J, object translations p^O, and rotation matrices Q^O. With $\dot{Y}_t = (Y_t - Y_{t-1})/(\vartheta_t - \vartheta_{t-1})$, frame timestamp ϑ_t, and ground truth denoted by $*$,

$$\mathcal{L}_{\text{motion}} = \langle\|\dot{\widehat{J}} - \dot{J}^*\|_2^2\rangle + \langle\|\dot{\widehat{p}}^{\,O} - \dot{p}^{O*}\|_2^2\rangle + \langle\|\dot{\widehat{Q}}^{\,O} - \dot{Q}^{O*}\|_F^2\rangle. \tag{2}$$

The rotation term measures matrix changes, not angular velocities.

Contact–relation fusion. To reconcile motion, contact, and relation predictions, we decode M into MANO hand meshes and object poses and integrate R from S_0. Anchor k on hand h yields an object-position candidate:

$$\widetilde{p}_{t,o}^{(h,k)} = a_{t,h,k} + Q_{t,h,k}\widehat{r}_{t,o,h,k}, \qquad w_{t,o,h} = w_{t,o,h}^{\text{geo}}\sigma(f_\phi(\xi_{t,o,h})). \tag{3}$$

Here, $(a_{t,h,k}, Q_{t,h,k})$ defines the anchor frame and $\widehat{r}_{t,o,h,k}$ is the integrated relation. The geometric weight w^{geo} uses proximity and predicted contact; the temporal module f_ϕ estimates binding strength from relative motion, pose, contact, and relation features. Fusion reconciles these candidates with global motion, adjusting object poses and necessary hand-root positions while preserving S_0.

For local gaps g and ground truth g^*, which defines $\mathcal{C}$ and $\mathcal{N}$, we impose

$$\mathcal{L}_{\text{gap}} = \langle\rho([g - \tau_c^{\text{gen}}]_+)\rangle_{\mathcal{C}} + \langle\rho([\tau_e - g]_+)\rangle_{\mathcal{N}}, \tag{4}$$

where ρ is a robust penalty and the thresholds define feasible contact and clear non-contact. Let $\Phi(g)$ map gaps to contact fields, $\Psi(\widehat{M}^f)$ re-encode relative increments from the fused trajectory, and b_t

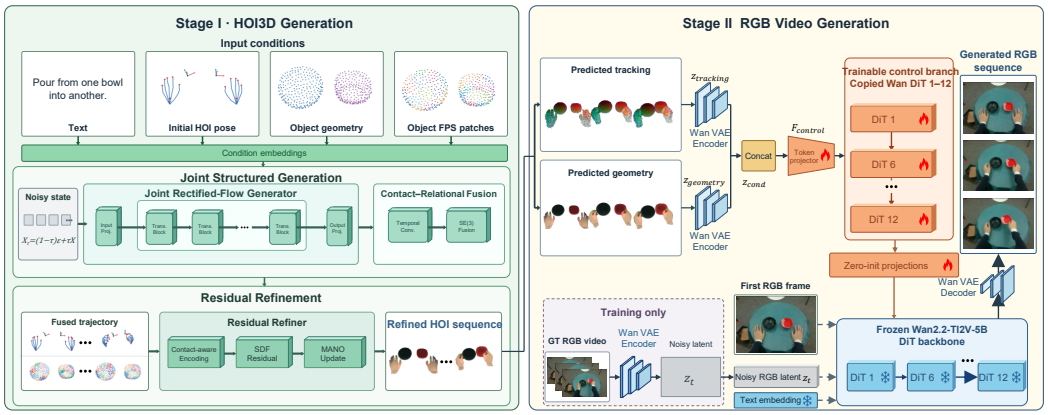


Figure 1: Overview of GEOMOTION-HOI. Stage I jointly generates motion, contact, and object-relative relations, followed by fusion and residual refinement. Stage II renders Tracking and Geometry Videos to guide a frozen video backbone through a trainable control branch, conditioned on the first RGB frame and text. Ground-truth RGB videos are used only during training.

denote a stable hand-surface point in object-local coordinates. The coupling loss is

$$\mathcal{L}_{\text{coupling}} = \mathcal{L}_{\text{gap}} + \langle \rho(\widehat{P} - \Phi(g)) \rangle + \langle \rho(\widehat{R} - \Psi(\widehat{M}^f)) \rangle + \langle \rho(\dot{\widehat{b}} - \dot{b}^*) \rangle_{\text{stable}}. \tag{5}$$

The final term penalizes sliding during persistent contact. The generator objective is

$$\mathcal{L}_{\text{gen}} = \mathcal{L}_{\text{joint}} + \lambda_m \mathcal{L}_{\text{motion}} + \lambda_c \mathcal{L}_{\text{coupling}}. \tag{6}$$

Contact-aware residual refinement. To correct remaining gaps and penetrations, a refiner uses predicted hand states, joint motion, hand–object distances and directions, contact and relation features, and their temporal changes. Joint-to-region attention, joint-wise interaction, and temporal convolutions predict bounded wrist and finger-pose residuals. Only the hands are updated; object trajectories and the initial frame remain fixed. Refined poses are decoded through MANO for geometric supervision.

Let V denote hand-surface positions, U finger-joint rotation matrices, and A stable-contact anchors in object-local coordinates. The pose and temporal losses are

$$\begin{aligned} \mathcal{L}_{\text{pose}} &= \lambda_s \langle \rho_s(\widehat{V} - V^*) \rangle + \lambda_u \langle \rho_u(\widehat{U} - U^*) \rangle + \lambda_r \mathcal{L}_{\text{reg}}, \\ \mathcal{L}_{\text{temporal}} &= \lambda_v \langle \rho_v(\dot{\widehat{V}} - \dot{V}^*) \rangle + \lambda_a \langle \rho_a(\dot{\widehat{A}} - \dot{A}^*) \rangle_{\text{stable}} + \lambda_d \mathcal{L}_{\text{drift}}. \end{aligned} \tag{7}$$

Here, $\rho_\cdot$ are scale-normalized SmoothL1 penalties. $\mathcal{L}_{\text{reg}}$ penalizes large residuals, especially away from contact; $\mathcal{L}_{\text{drift}}$ compares anchor displacements from the start of each stable-contact interval.

We measure contact gaps at surface points and hand regions, $q \in \{\text{point}, \text{region}\}$. For refined gaps $\widehat{g}_i^q$, pre-refinement gaps $g_i^{q,0}$, and ground-truth contact sets $\mathcal{C}_q$, define the fixed bound $\bar{g}_i^q = \min(g_i^{q,0} + \varepsilon_c, \tau_c^{\text{ref}})$:

$$\mathcal{L}_{\text{contact}} = \sum_q \lambda_q \left\langle \rho_c \left(\frac{[\widehat{g}_i^q - \bar{g}_i^q]_+}{s_c} \right) \right\rangle_{\mathcal{C}_q}. \tag{8}$$

This one-sided loss corrects missed contacts without pulling feasible surfaces closer; region terms distinguish the palm and individual fingers.

Local signed distances from object samples to the refined hand mesh give penetration depths $\widehat{d}_i$. With pre-refinement depths d_i^0 and $m(d) = [d - \epsilon_p]_+$, we penalize final penetration and increases over the fixed baseline:

$$\mathcal{L}_{\text{pen}} = \left\langle \rho_p \left(\frac{m(\widehat{d}_i)}{s_p} \right) + \lambda_{\text{inc}} \rho_p \left(\frac{[m(\widehat{d}_i) - m(d_i^0)]_+}{s_p} \right) \right\rangle_i. \tag{9}$$

The refiner objective is

$$\mathcal{L}_{\text{ref}} = \mathcal{L}_{\text{pose}} + \mathcal{L}_{\text{temporal}} + \lambda_g (\mathcal{L}_{\text{contact}} + \mathcal{L}_{\text{pen}}). \tag{10}$$

3.2 GEOMETRY- AND MOTION-CONDITIONED HOI VIDEO GENERATION

Using camera information, we align the refined 3D sequence with video frames and render Tracking and Geometry Videos as complementary motion and geometry conditions.

Base image-to-video model. We use Wan2.2-TI2V-5B with a video VAE, text encoder, and latent DiT. The encoded first frame anchors scene layout and appearance, while a condition projector and control branch incorporate the rendered 3D conditions.

Dual-stream visual conditions. Following DaS (Gu et al., 2025), Tracking Video C^{trk} projects fixed object-surface samples and MANO vertices across frames, using a fixed color per point to identify correspondences. To complement these sparse tracks with continuous surfaces and occlusion boundaries, we render full meshes with a shared depth buffer:

$$C_t^{\mathrm{geo}}(\mathbf{u}) = \mathbf{a}_{s_t(\mathbf{u})}(I_0)\rho_{e_t(\mathbf{u})}(D_t(\mathbf{u})), \qquad \mathbf{u} \in \Omega_t. \tag{11}$$

Here, Ω_t is the visible hand–object region, D_t is camera depth, and e_t and s_t identify the visible entity and region. Colors $\mathbf{a}_s(I_0)$ are sampled from corresponding first-frame regions, with fixed palm and finger offsets. The depth-to-brightness map ρ_e uses a sequence-fixed normalization range, encoding depth while stable colors preserve region correspondence to the first frame.

A shared video VAE encoder $\mathcal{E}$ encodes both conditions. Their latents are concatenated and projected into control tokens by P_θ. A control branch copied from the first 12 DiT blocks injects features through zero-initialized projections:

$$q = P_\theta([\mathcal{E}(C^{\mathrm{trk}}); \mathcal{E}(C^{\mathrm{geo}})]_{\mathrm{ch}}), \qquad h_\ell \leftarrow h_\ell + \alpha(p)\mathcal{Z}_\ell(g_\ell). \tag{12}$$

Here, g_ℓ and h_ℓ are control and backbone features, and $\alpha(p)$ controls their influence during denoising.

3.3 TRAINING AND INFERENCE

We train the joint generator and fusion module with $\mathcal{L}_{\mathrm{gen}}$ (Eq. equation 6), then fix the generator and fused trajectories to train the refiner with $\mathcal{L}_{\mathrm{ref}}$ (Eq. equation 10). For video generation, only the condition projector and control branch are trained; the DiT, text encoder, and video VAE remain frozen. Training conditions are rendered from 3D sequences aligned with target videos.

For latent flow matching, we interpolate target latents z^* and Gaussian noise ϵ as $z_\sigma = (1-\sigma)z^* + \sigma\epsilon$, with target velocity $v^* = \epsilon - z^*$. Replacing the first temporal slice with the encoded first frame gives $\widetilde{z}_\sigma$, from which the model predicts $\widehat{v}$ and $\widehat{z}^* = \widetilde{z}_\sigma - \sigma\widehat{v}$. A latent hand–object mask $\mathcal{M}$ emphasizes interaction regions:

$$\mathcal{L}_{\mathrm{video}} = \mathbb{E}\big[N^{-1}\sum_{j=1}^{N}((1+\mathcal{M}_j)(\widehat{v}_j - v_j^*)^2 + \lambda_{\mathrm{rec}}(1+\mathcal{M}_j)^2(\widehat{z}_j^* - z_j^*)^2)\big]. \tag{13}$$

Here, N counts supervised latent elements; $\mathcal{M}$ is broadcast over channels and used only during training.

At inference, the 3D sequence generated and refined from (y, S_0, H, G) is rendered using the camera associated with I_0. The resulting conditions, I_0, and y guide denoising. Control strength $\alpha(p)$ remains 1 for the first half of denoising, then decays to 0.5 following a cosine schedule. We reinsert the first-frame latent before each velocity prediction and after each sampling update, then decode with the frozen video VAE.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets and Data Processing. We evaluate our method on two hand–object interaction video datasets: OakInk2 (Zhan et al., 2024) and TACO (Liu et al., 2024). Both datasets provide synchronized egocentric and third-person videos. As our method focuses on egocentric hand–object interaction video generation, we exclusively use videos captured by the head-mounted cameras for training and

evaluation. For OakInk2, we select video clips with complete hand, object, and camera annotations, and use a fixed set of 180 clips for testing. For TACO, we perform the train–test split at the raw-sequence level to prevent clips originating from the same sequence from appearing in both subsets. After filtering and temporal windowing, the resulting split contains 2,347 training clips and 345 testing clips. To remain consistent with prior 3D hand–object motion generation settings, we represent each 3D interaction sequence using 196 frames. For video generation, we temporally subsample the corresponding egocentric videos into 81-frame clips, where the first frame serves as the image condition and the remaining frames are used as generation targets. The Tracking and Geometry Videos are rendered from the temporally aligned 3D hand and object annotations.

Evaluation Metrics. For 3D interaction generation, following IMoS (Ghosh et al., 2023), we report Top-1 and Top-3 action classification accuracy and motion FID using a dataset-specific action classifier. Accuracy measures action-semantic consistency, while FID compares classifier-feature distributions between real and generated motions (Heusel et al., 2017). Physical plausibility is evaluated using hand–object interpenetration volume (IV, cm^3) and maximum interpenetration depth (ID, cm) (Li et al., 2025). We also report the frame-level contact ratio at a 5 mm threshold (CR@5, %) over frames with object translation or articulation (Song et al., 2026; Zhang et al., 2025). CR@5 is interpreted relative to the ground-truth ratio: closer values are preferred when penetration remains low. Physical metrics are averaged over both hands.

For video generation, we evaluate frame-level fidelity against temporally aligned ground truth using PSNR (Huynh-Thu & Ghanbari, 2008), SSIM (Wang et al., 2004), and LPIPS (Zhang et al., 2018). We additionally report image FID (Heusel et al., 2017), computed from Inception features of real and generated frames, to assess distributional similarity, and mean frame–text CLIP similarity (Radford et al., 2021) to measure semantic alignment. Frame-level scores are averaged over generated frames, excluding the conditioned first frame; FID is computed over the corresponding sets of real and generated frames.

4.2 Baseline Comparisons

Table 1: Comparison of 3D hand–object interaction generation. * denotes our dataset-adapted implementations.

Dataset	Method	Acc (%) ↑		FID ↓	Interpenetration ↓		CR@5 (%) →
		Top-1	Top-3		IV (cm^3)	ID (cm)	
OakInk2	GT	63.89	77.78	–	1.740	1.061	89.59
	Text2HOI	52.77	63.88	0.421	4.390	2.159	62.26
	LatentHOI	54.50	72.78	0.366	5.918	1.420	72.09
	DiffH2O	56.11	72.22	0.382	**3.369**	1.456	73.81
	Ours	**58.89**	**76.67**	**0.216**	3.899	**1.094**	**82.77**
TACO	GT	68.99	90.43	–	2.307	0.264	76.76
	DiffH2O*	**68.41**	90.14	0.236	3.592	0.347	**80.42**
	Dex2HOI*	66.38	86.96	0.273	**3.246**	0.352	84.30
	Ours	**68.41**	**91.01**	**0.183**	3.612	**0.335**	81.82

3D interaction generation. We compare against Text2HOI (Cha et al., 2024), DiffH2O (Christen et al., 2024), LatentHOI (Li et al., 2025), and Dex2HOI (Pratikaki et al., 2026). These baselines cover complementary approaches: contact-guided motion generation, staged grasping and manipulation, latent grasp priors, and hand-relative representations for bimanual interaction. DiffH2O and Dex2HOI are adapted to TACO, as indicated by asterisks in Table 1.

Video generation. We compare against Cosmos3-Nano (NVIDIA, 2026), Wan2.2-I2V-A14B (Team Wan et al., 2025), MiniMax H3 (MiniMax, 2026), and PAM (Gao et al., 2026). The first three provide general-purpose video-generation baselines for evaluating visual quality and interaction consistency. PAM provides a task-specific comparison through depth, segmentation, and hand-keypoint conditioning, allowing us to assess our tracking and geometry representation against established multi-condition HOI video control.

Main Results on 3D Interaction Generation. As shown in Table 1, our method achieved the lowest motion FID and maximum penetration depth among generated methods on both datasets. On OakInk2, it obtained the highest Top-1 and Top-3 accuracy at 58.89% and 76.67%, respectively, and reduced motion FID by 40.9% relative to LatentHOI. Its contact ratio was also closest to ground truth, with a gap of 6.82 percentage points. On TACO, our method matched DiffH2O in Top-1 accuracy, achieved the highest Top-3 accuracy at 91.01%, and reduced motion FID by 22.7% relative to DiffH2O. These results support improved motion-distribution alignment and strong action-semantic consistency, although the physical advantages were not uniform: DiffH2O achieved a lower penetration volume on OakInk2 and a contact ratio closer to ground truth on TACO, while Dex2HOI obtained the lowest penetration volume on TACO.

Table 2: Comparison of video generation on OakInk2 and TACO. PAM results are reported only for OakInk2.

Dataset	Method	FID ↓	SSIM ↑	LPIPS ↓	PSNR ↑	CLIP ↑
OakInk2	PAM	30.972	**0.665**	0.371	17.377	0.901
	Wan2.2-I2V-A14B	60.573	0.572	0.544	14.397	0.831
	Cosmos3-Nano	59.508	0.589	0.517	14.961	0.866
	MiniMax H3	53.761	0.613	0.469	15.724	0.890
	Ours	**28.148**	0.656	**0.339**	**18.237**	**0.908**
TACO	Wan2.2-I2V-A14B	40.143	0.563	0.549	14.026	0.833
	Cosmos3-Nano	38.242	0.586	0.522	14.460	0.855
	MiniMax H3	37.026	0.591	0.528	14.604	0.853
	Ours	**11.889**	**0.674**	**0.235**	**20.100**	**0.917**

Main Results on Video Generation. Table 2 compares video generation on OakInk2 and TACO. On OakInk2, our method achieved the best FID, LPIPS, PSNR, and CLIP scores among the compared methods. Relative to PAM, the strongest baseline across these metrics, it reduced FID by 9.1% and improved PSNR from 17.377 to 18.237, although PAM retained a slightly higher SSIM (0.665 versus 0.656). On TACO, our method achieved the best results across all five metrics among the reported methods, reducing FID by 67.9% relative to MiniMax H3 and LPIPS by 55.0% relative to Cosmos3-Nano, the strongest baselines on the respective metrics.

Figure 2 complements these quantitative results with qualitative comparisons. The baselines exhibit changes in object shape or identity during manipulation: for example, MiniMax H3 replaces the wooden board with a plate-like object in panel (b), while PAM alters the cup shape in panel (c). In these examples, our method better preserves the manipulated objects' shape and appearance across frames. Together with the condition ablation, these observations support the complementary roles of tracking and geometry cues in controlled HOI video synthesis.

4.3 Ablation Study

3D Interaction Generation. Table 3 compares the baseline, fusion-only configuration, and full model. The fusion-only configuration improved Top-1 and Top-3 accuracy and reduced motion FID, but increased both penetration depth and volume. This indicates that improved motion-distribution alignment does not necessarily imply better local geometry. The full model achieved the highest classification accuracy and lowest FID, reducing FID by 22.2% relative to the baseline. Compared

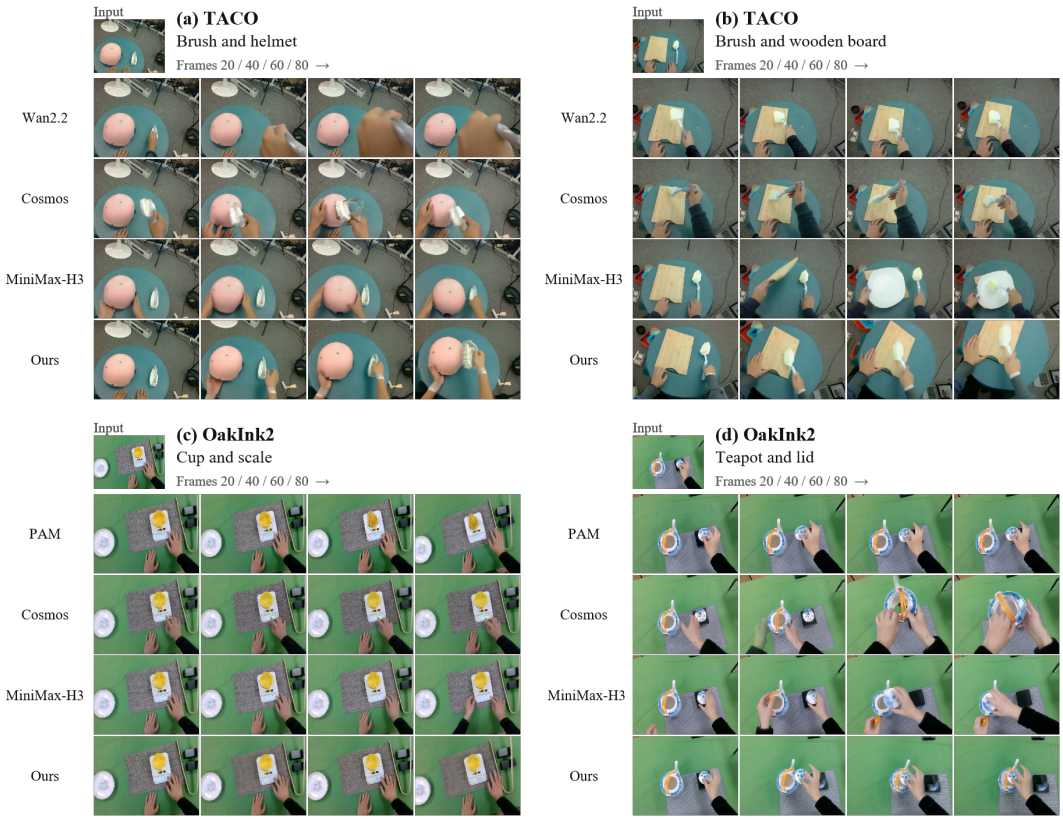


Figure 2: Qualitative comparisons of HOI video generation on TACO (a–b) and OakInk2 (c–d). Each panel shows the input image and generated frames at indices 20, 40, 60, and 80. Baseline results are shown in the upper rows, and our results in the bottom row. The examples illustrate differences in object-shape preservation and appearance consistency during manipulation.

Table 3: Ablation of 3D interaction generation on TACO.

Configuration	Top-1 (%) ↑	Top-3 (%) ↑	FID ↓	ID (cm) ↓	IV (cm^3) ↓	CR@5 (%) →
Baseline	64.64	85.51	0.235	0.482	3.647	**78.32**
Fusion only	68.12	89.86	0.201	0.536	3.812	82.78
Full model	**68.41**	**91.01**	**0.183**	**0.335**	**3.612**	81.82

with fusion alone, it also reduced maximum penetration depth by 37.5% and lowered penetration volume from 3.812 to 3.612. However, the baseline retained the contact ratio closest to ground truth. Overall, the complete configuration improved motion quality and penetration metrics, while contact-frequency agreement remained imperfect.

Table 4: Ablation of the tracking and geometry conditions on TACO.

Condition	FID ↓	SSIM ↑	LPIPS ↓	PSNR ↑	CLIP ↑
Baseline	83.614	0.470	0.887	13.650	0.594
Geometry only	23.798	0.634	0.359	17.469	0.797
Tracking only	21.390	0.638	0.337	17.988	0.852
Full model	**11.889**	**0.674**	**0.235**	**20.100**	**0.917**

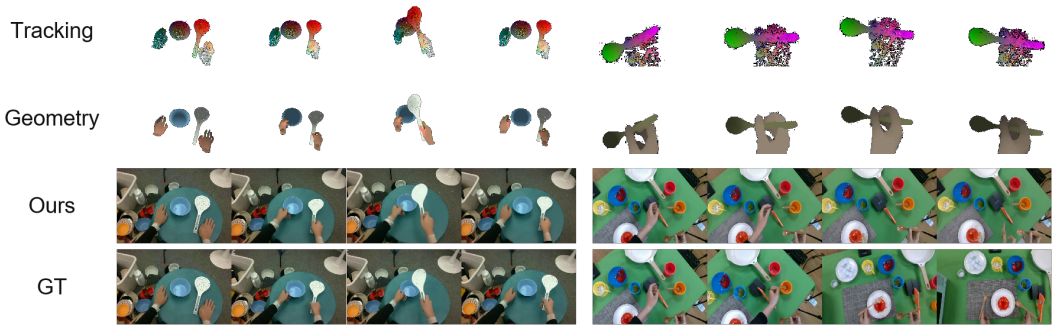


Figure 3: Qualitative results of joint inference. Columns show successive frames for two examples. Rows show the Tracking Video, Geometry Video, generated RGB video (Ours), and ground-truth video (GT). Both visual conditions are rendered from predicted 3D interactions rather than ground-truth future annotations. The generated videos qualitatively follow the conditioned motion while preserving object shape and first-frame appearance. GT is shown for reference.

Tracking and Geometry Conditions. Table 4 compares the baseline without either visual condition, each condition individually, and the full model. Both individual conditions improved all five metrics over the baseline, with tracking alone consistently outperforming geometry alone. Combining the two conditions yielded the best results across all metrics: FID decreased from 83.614 to 11.889, LPIPS decreased from 0.887 to 0.235, and PSNR increased from 13.650 to 20.100. The full model also improved over tracking alone, indicating that the geometry condition provides useful information beyond surface-point motion correspondences. These results support the complementary roles of tracking and geometry: tracking conveys cross-frame motion associations, while geometry supplements visible surface structure and region correspondence.

4.4 Joint Inference.

We combine the 3D interaction generator and the video generator for joint inference, as illustrated in Figure 3. The predicted 3D interaction sequence is rendered into Tracking and Geometry Videos, which guide video synthesis together with the observed first RGB frame and text. The qualitative results are consistent with the intended behavior: the generated videos follow the predicted 3D trajectories and maintain consistent object geometry during interaction. These examples suggest that the generated 3D sequence provides an effective structural reference for video synthesis, supporting geometric consistency without requiring a ground-truth future trajectory or an exact target hand pose.

5 Discussion and Conclusion

We presented GEOMOTION-HOI, a framework connecting future 3D hand–object interaction generation with geometry-guided egocentric video synthesis. Its central contribution is to generate future interaction geometry as a control signal rather than require it as an input. By coordinating motion, contact, and object-relative relationships, the framework produces structured trajectories that guide video synthesis through complementary tracking and geometry conditions, while the observed first frame anchors appearance. This design connects 3D interaction generation with its visual realization, without requiring a predefined future trajectory or an exact target hand pose.

The current framework is limited to interactions involving at most two objects. Its responsiveness to camera motion and viewpoint changes is also limited, which may reduce generation consistency under substantial camera movement.

AI Use Statement

We used OpenAI's ChatGPT to aid and polish manuscript writing, including drafting, and language editing. The authors reviewed and revised the AI-assisted content and take full responsibility for the final manuscript.

REFERENCES

Rick Akkerman, Haiwen Feng, Michael J. Black, Dimitrios Tzionas, and Victoria Fernández Abrevaya. InterDyn: Controllable interactive dynamics with video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12467–12479, 2025. doi: 10.1109/CVPR52734.2025.01163.

Junuk Cha, Jihyeon Kim, Jae Shin Yoon, and Seungryul Baek. Text2HOI: Text-guided 3D motion generation for hand-object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1577–1585, 2024. doi: 10.1109/CVPR52733.2024.00156.

Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. VideoCrafter2: Overcoming data limitations for high-quality video diffusion models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7310–7320, 2024. doi: 10.1109/CVPR52733.2024.00698.

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, 2026.

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. 145:1–145:11. Association for Computing Machinery, 2024. doi: 10.1145/3680528.3687563.

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.

Anindita Ghosh, Rishabh Dabral, Vladislav Golyanik, Christian Theobalt, and Philipp Slusallek. IMoS: Intent-driven full-body motion synthesis for human-object interactions. *Computer Graphics Forum*, 42(2):1–12, 2023. doi: 10.1111/cgf.14739.

Zekai Gu, Rui Yan, Jiahao Lu, Peng Li, Zhiyang Dou, Chenyang Si, Zhen Dong, Qifeng Liu, Cheng Lin, Ziwei Liu, Wenping Wang, and Yuan Liu. Diffusion as shader: 3D-aware video diffusion for versatile video generation control. In *Proceedings of the ACM SIGGRAPH 2025 Conference Papers*, pp. 62:1–62:12. Association for Computing Machinery, 2025. doi: 10.1145/3721238.3730607.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David Fleet. Video diffusion models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 8633–8646. Curran Associates, Inc., 2022. doi: 10.52202/068431-0628. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/39235c56aef13fb05a6adc95eb9d8d66-Paper-Conference.pdf.

Quan Huynh-Thu and Mohammed Ghanbari. Scope of validity of PSNR in image/video quality assessment. *Electronics Letters*, 44(13):800–801, 2008. doi: 10.1049/el:20080522.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17416–17425, 2025. doi: 10.1109/CVPR52734.2025.01623.

Yun Liu, Haolin Yang, Xu Si, Ling Liu, Zipeng Li, Yuxiang Zhang, Yebin Liu, and Li Yi. TACO: Benchmarking generalizable bimanual Tool-ACtion-Object understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21740–21751, 2024.

MiniMax. MiniMax H3: An open model breaking the boundaries between tasks and modalities. MiniMax Research, 2026. URL https://www.minimax.io/blog/minimax-h3. Accessed September 20, 2026.

NVIDIA. Cosmos 3: Omnimodal world models for physical AI. *arXiv preprint arXiv:2606.02800*, 2026. URL https://research.nvidia.com/labs/cosmos-lab/cosmos3/technical-report.pdf.

NVIDIA, Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, Daniel Dworakowski, Jiaojiao Fan, Michele Fenzi, Francesco Ferroni, Sanja Fidler, Dieter Fox, Songwei Ge, Yunhao Ge, Jinwei Gu, Siddharth Gururani, Ethan He, Jiahui Huang, Jacob Huffman, Pooya Jannaty, Jingyi Jin, Seung Wook Kim, Gergely Klár, Grace Lam, Shiyi Lan, Laura Leal-Taixe, Anqi Li, Zhaoshuo Li, Chen-Hsuan Lin, Tsung-Yi Lin, Huan Ling, Ming-Yu Liu, Xian Liu, Alice Luo, Qianli Ma, Hanzi Mao, Kaichun Mo, Arsalan Mousavian, Seungjun Nah, Sriharsha Niverty, David Page, Despoina Paschalidou, Zeeshan Patel, Lindsey Pavao, Morteza Ramezanali, Fitsum Reda, Xiaowei Ren, Vasanth Rao Naik Sabavat, Ed Schmerling, Stella Shi, Bartosz Stefaniak, Shitao Tang, Lyne Tchapmi, Przemek Tredak, Wei-Cheng Tseng, Jibin Varghese, Hao Wang, Haoxiang Wang, Heng Wang, Ting-Chun Wang, Fangyin Wei, Xinyue Wei, Jay Zhangjie Wu, Jiashu Xu, Wei Yang, Lin Yen-Chen, Xiaohui Zeng, Yu Zeng, Jing Zhang, Qinsheng Zhang, Yuxuan Zhang, Qingqing Zhao, and Artur Zolkowski. Cosmos world foundation model platform for physical AI. *arXiv preprint arXiv:2501.03575*, 2025. URL https://arxiv.org/abs/2501.03575.

Chrysa Pratikaki, Pablo Ruiz-Ponce, Jiankang Deng, Stefanos Zafeiriou, and Rolandos Alexandros Potamias. Dex2HOI: Dexterous bimanual two-object interaction generation, 2026. URL https://arxiv.org/abs/2605.30444.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021.

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, 2026. URL https://arxiv.org/abs/2607.01768.

Sruthi Sudhakar, Ruoshi Liu, Basile Van Hoorick, Carl Vondrick, and Richard Zemel. Controlling the world by sleight of hand. In *Computer Vision – ECCV 2024*, pp. 414–430. Springer Nature Switzerland, 2025.

Team Wan, Ang Wang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. URL https://arxiv.org/abs/2503.20314.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004. doi: 10.1109/TIP.2003.819861.

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, Xu Bin, Xiaotao Gu, Yuxiao Dong, and Jie Tang. CogVideoX: Text-to-video diffusion models with an expert transformer. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 83048–83077, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/ce31378e9f41d8907e97dab172b6c559-Paper-Conference.pdf.

Xinyu Zhan, Lixin Yang, Yifei Zhao, Kangrui Mao, Hanlin Xu, Zenan Lin, Kailin Li, and Cewu Lu. OAKINK2: A dataset of bimanual hands-object manipulation in complex task completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 445–456, 2024.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018.

Wanyue Zhang, Rishabh Dabral, Vladislav Golyanik, Vasileios Choutas, Eduardo Alvarado, Thabo Beeler, Marc Habermann, and Christian Theobalt. BimArt: A unified approach for the synthesis of 3D bimanual interaction with articulated objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27694–27705, 2025.