EmboDirector: Directing 4D Human–Scene Interactions from a Single Image

Anonymous authors
Paper under double-blind review

Abstract

Full-body human–scene interaction data with consistent geometry, motion, and contact is scarce and costly to capture. Video models acquire rich interaction priors through large-scale pretraining, yet generated interactions can exhibit object-orientation drift that direct reconstruction would reproduce. We present *EmboDirector*, which reconstructs structured 4D interactions from videos generated using a single indoor image and an interaction description. A multimodal large language model (MLLM) Director corrects the initial recovered layout, assigns frame reliability to reduce the influence of video drift, and organizes estimated contact candidates into a spatiotemporal interaction graph. The graph specifies interaction relationships and their active intervals, guiding joint optimization of human and object trajectories alongside reliable visual evidence. The input image anchors layout corrections throughout. Comparisons with HOI generation and video-based reconstruction baselines show improved contact rate, mean penetration, and object alignment; an ablation examines graph-conditioned optimization. The reconstructed scenes support physics simulation and geometry-guided video synthesis.

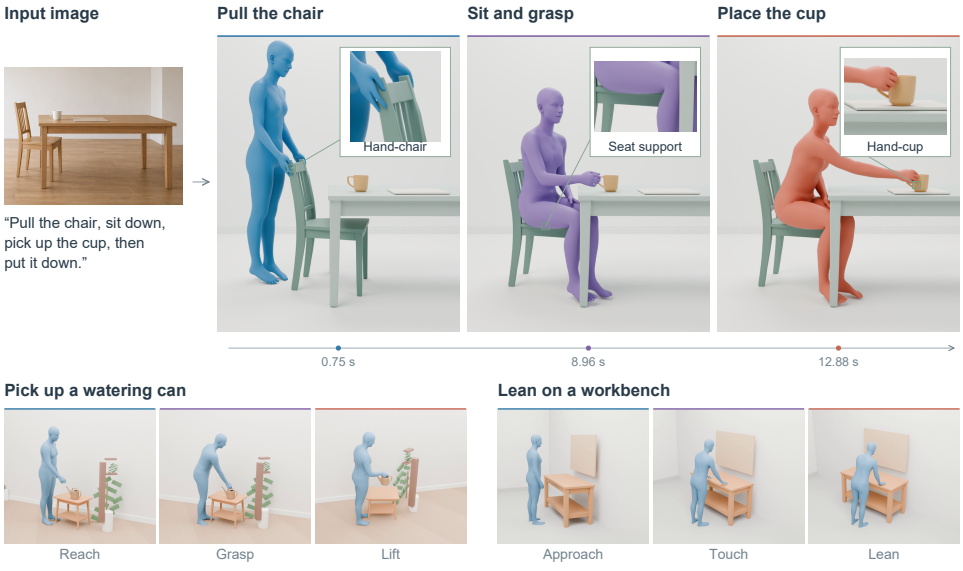


Figure 1: Top: chair pulling, seated cup grasping, and placement. Bottom: mesh sequences for picking up a watering can and leaning on a workbench.

1 Introduction

Full-body interactions with objects and scenes are central to embodied intelligence and simulation, but their data is hard to obtain. Motion capture is costly and limited to the objects, layouts, and

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.

actions available in a studio (Bhatnagar et al., 2022); web videos carry no metric geometry or contact labels. Recovering interactions from video can broaden coverage if reconstruction resolves contact drift, foot sliding, and interpenetration.

Video generators such as CogVideoX and Wan learn strong motion priors from large-scale data (Yang et al., 2025; WanTeam et al., 2025), making generated video a useful source of interaction motion. HOI sequences can nevertheless drift: a kettle or cup may change orientation inconsistently with its grasp and neighboring frames. This differs from an erroneous initial layout introduced by 3D recovery; fitting the drifting frames more closely would reproduce the video error. Reconstruction must both correct initialization and determine which temporal evidence to trust.

We propose *EmboDirector* (Figure 2), which reconstructs continuous 4D human–scene interactions from videos generated using a single indoor image and description. A sequence such as pulling a chair, sitting down, picking up a cup, and placing it back (Figure 1) requires coordinated motion as contacts change across multiple objects. We recover human and object geometry from the video and estimate occlusion-aware contact candidates.

An MLLM Director organizes these geometric candidates into a spatiotemporal interaction graph of relations and active intervals. The estimator determines contact locations; the graph determines when the corresponding constraints apply. The Director also corrects the initial layout against the input image and assigns frame reliability to reduce the influence of video drift. Joint optimization combines these decisions with contact, slip, and penetration constraints to refine human and object trajectories.

Our contributions are as follows:

- A framework for reconstructing multi-stage human–scene interactions from a single image and description, producing continuous human and object trajectories for interactions involving multiple objects in a shared scene.
- An MLLM Director that organizes estimated contact candidates into a spatiotemporal interaction graph, specifying interaction relations and their active intervals to guide reconstruction.
- Graph-conditioned joint refinement of human and object trajectories from the corrected layout, combining contact, slip, and penetration constraints with reliability-weighted video evidence.

2 Related Work

Interaction datasets such as GRAB, BEHAVE, and TRUMANS capture body–object relationships and dynamic scene interactions (Taheri et al., 2020; Bhatnagar et al., 2022; Jiang et al., 2024b). Scene-conditioned methods model body–surface contact and generate motion from scene geometry or language (Hassan et al., 2021b;a; Wang et al., 2022; Huang et al., 2023), with extensions to multi-stage instructions, navigation, and grasping (Jiang et al., 2024a; Yao et al., 2025). Moving objects require human and object trajectories to be modeled together. OMOMO and CHOIS use object motion and spatial controls to couple these trajectories (Li et al., 2023; 2024a); other diffusion-based methods encode contact or kinematic dependencies during generation (Xu et al., 2023; Peng et al., 2025; Zeng et al., 2025). HOIDiNi separates contact selection from full-body refinement (Ron et al., 2025), allowing contact requirements to guide the subsequent motion.

Generative visual priors offer another source of interaction motion. GenZI, AvatarGO, and AnchorHOI lift image or video priors into 3D or 4D interactions (Li & Dai, 2024; Cao et al., 2024; Dai & Xu, 2026); InterDreamer, DAViD, and ZeroHOI combine generative proposals with motion, affordance, or dynamics models (Xu et al., 2024; Kim et al., 2025; Lou et al., 2025). The closest scene-level methods are ZeroHSI, which reconstructs generated interactions conditioned on a 3D scene and initial state (Li et al., 2026a), and GenHSI, which uses a scene image and character references to construct VLM-planned 3D keyframes for video generation (Li et al., 2026b). We focus on reconstructing continuous trajectories from generated video, where plausible individual frames can still contain inconsistent object motion. The Director corrects the recovered initialization, weights unreliable frames, and organizes contact intervals so that geometric refinement can resolve these inconsistencies.

Related approaches also use structured reasoning for interaction generation. HOI-PAGE takes 3D object assets and text, including multi-object configurations, and uses part-affordance graphs for

grounding, video generation, and 4D fitting with contact and relative-motion constraints (Li & Dai, 2026). FantasyHSI uses graph-based agents for navigation, action decomposition, and trajectory feedback (Mu et al., 2025). EmboDirector starts from an indoor scene image, retaining its background and support geometry as references for correction. Its interval-based graph organizes estimated contacts during reconstruction, with geometric residuals guiding subsequent event revisions.

Contact estimation supplies the geometric correspondences needed by this reasoning process. LEMON predicts contact, affordance, and spatial relations (Yang et al., 2024c); CONTHO and PICO refine body–object geometry through contact, while EasyHOI uses foundation-model priors for hand–object recovery (Nam et al., 2024; Cseke et al., 2025; Liu et al., 2025). PROX and JOSH incorporate scene geometry into body fitting (Hassan et al., 2019; Liu et al., 2026), CRISP recovers static support surfaces with contact-guided planar primitives (Wang et al., 2026), and RePHO and HSImul3R incorporate control or simulation (Huang et al., 2026; Cao et al., 2026). We retain a geometric contact estimator and use the Director to determine which candidate relationships should persist, and when. The resulting constraints act jointly with image-anchored layout correction and reliability-weighted video fitting. Appendix D discusses broader visual priors.

3 METHOD

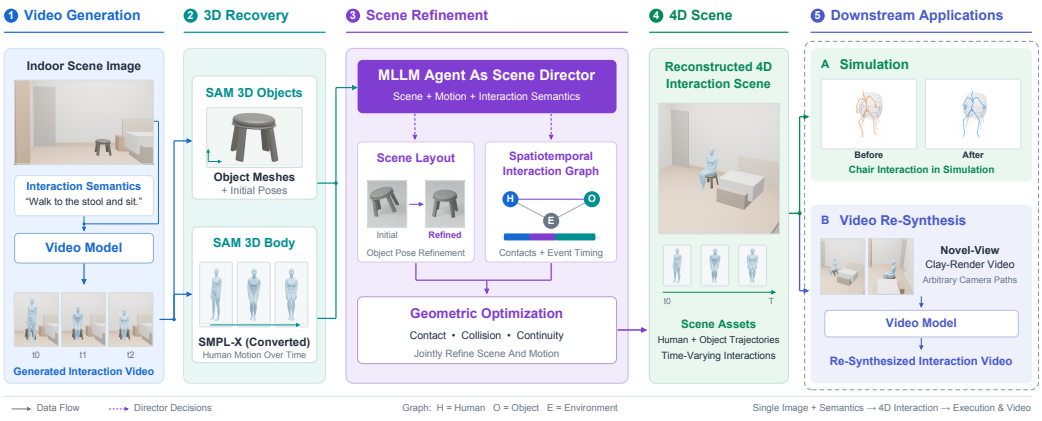


Figure 2: EmboDirector: video-based recovery, Director-guided layout correction and frame weighting, and spatiotemporal graph optimization for simulation and rendering.

3.1 PROBLEM FORMULATION

Given an indoor image I and an interaction description c, we generate a video $\widehat{V}$ and reconstruct from it a 4D scene with human motion, object trajectories, and contact relationships, correcting implausible interactions in the process. The video serves as temporal evidence rather than ground truth: the image anchors the visible scene, and the description specifies the intended action. We assume a single person, rigid objects, and a fixed background, and represent the scene as

$$\mathcal{S} = \{\mathcal{E},\ (M_i, s_i, X_{i,1:T})_{i=1}^{N},\ \beta, q_{1:T},\ \mathcal{G}\}, \tag{1}$$

where $\mathcal{E}$ denotes the background and floor, M_i is the canonical mesh of object i, $s_i > 0$ is its scale, and $X_{i,t} = (R_{i,t}, p_{i,t}) \in \mathrm{SE}(3)$ is its pose at time t, so that a mesh point $y \in M_i$ has world coordinates $s_i R_{i,t} y + p_{i,t}$. The body is $H_t = \mathrm{SMPLX}(\beta, q_t)$, with shape β shared across time and per-frame pose and translation q_t. The interaction graph $\mathcal{G}$ (Section 3.4) describes the interaction events. The camera calibration $\mathcal{C} = (K, C_{1:T})$, consisting of intrinsics and extrinsics, is estimated during recovery and then held fixed; it is used for projection into the video but is not part of the optimized state. Static objects keep a single pose throughout the sequence, while movable objects follow trajectories that alternate between stationary and active intervals. All temporal quantities are computed from the sequence timestamps.

3.2 Video Generation and 3D Recovery

We generate $\widehat{V} = G_{\text{vid}}(I, c)$ with MiniMax H3 (MiniMax, 2026), obtaining the action order and approximate timing. We prompt for a stable viewpoint and represent residual camera motion with a shared trajectory. The input image anchors the layout throughout refinement, since generated frames may contain inconsistent contact or object motion.

SAM 3D Objects (Chen et al., 2025) initializes object meshes and poses from the image and selected frames. Each object retains a persistent identity and canonical mesh. SAM 3D Body (Yang et al., 2026) recovers per-frame human geometry; we convert its MHR output to SMPL-X (Pavlakos et al., 2019) by fitting joints and surface observations with a shared body shape (Appendix B). This sequence supports contact estimation and optimization.

Camera projection, floor orientation, relative depth, and visible support align the body and objects in a common frame. A body-size prior resolves metric-scale ambiguity; the recovered floor and initial scene define the world frame. Visibility weights down-weight uncertain observations. The aligned initialization $\widehat{\mathcal{S}}$, with human states $\widehat{q}_t$ and object poses $\widehat{X}_{i,t}$, remains the reference for measuring refinement-induced changes.

3.3 Occlusion-Aware Contact Estimation

Estimating contact requires both knowing what the action is and checking that the two surfaces actually meet. Before the Director reasons about the scene, we initialize human–object correspondences with an occlusion-aware estimator (Figure 3). InteractVLM (Dwivedi et al., 2025) provides contact seeds on the body regions likely to participate in the interaction.

We recover correspondences for both human-occludes-object and object-occludes-human configurations. In the former, we intersect the occluded part of the rendered object silhouette with projected body contact seeds. In the latter, we use a self-occlusion-corrected mask. Layered rendering resolves the surface pairs, which are back-projected to 3D. Candidates are filtered by 3D separation and opposing surface normals; Appendix B.1 details layer selection and self-occlusion correction. These correspondences provide the contact locations. The Director organizes them into interaction relations and active intervals using scene and temporal context (Section 3.4). Appendix B.1 also compares the estimator with InteractVLM.

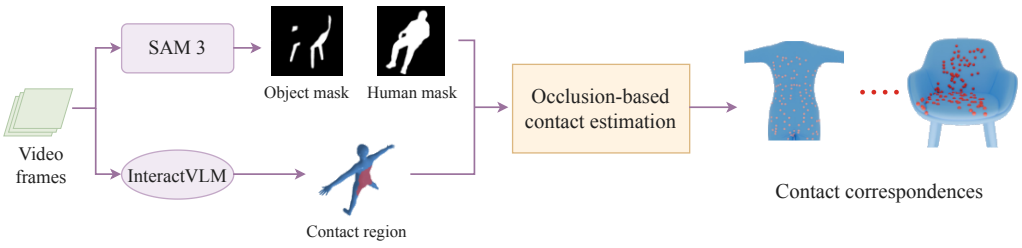


Figure 3: Occlusion-aware contact candidates from segmentation masks and InteractVLM seeds, supplied to the Director for temporal graph construction.

3.4 MLLM-Guided Scene Direction

The Director uses GPT-6 Astra (OpenAI, 2026) within an author-developed harness. Given the image, instruction, sampled video frames, contact candidates, and entity-labeled recovery renders, it produces layout corrections, frame reliability weights, and a spatiotemporal interaction graph (Appendix A). These outputs address initialization errors, unreliable video evidence, and interaction timing, respectively.

Initial-layout correction. Monocular recovery can misplace an object or estimate the wrong orientation before motion begins. The Director compares the recovered layout with the input image and proposes corrections to initial poses (Figure 4). Silhouettes, relative depth, support, and collision checks constrain these changes, with a penalty on deviation from recovery. Accepted geometry

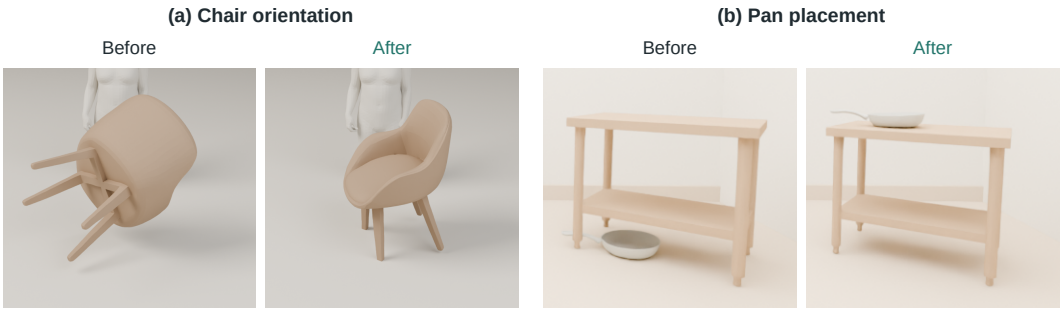


Figure 4: Qualitative scene-layout refinement: chair orientation (left) and pan support placement (right).

determines the contact anchors. Action-induced motion, such as chair pulling, remains a separate trajectory.

Frame reliability. Generated HOI can contain orientation drift even when individual frames look plausible: a kettle or cup may turn inconsistently with the grasp and surrounding motion. The Director identifies these inconsistencies and assigns frame weights $\alpha_t \in [0, 1]$. Lower weights reduce the corresponding video-fitting terms, allowing reconstruction to follow coherent interaction relationships across drifting frames. The intended action, grasp, and neighboring frames distinguish drift from legitimate object rotation.

Spatiotemporal organization. Contact locations are supplied by the estimator in Section 3.3. The Director organizes candidate body–surface correspondences into relations and active intervals, rather than predicting new contact coordinates. The resulting graph $\mathcal{G} = (\mathcal{V}, \mathcal{A})$ contains human, object, and environment nodes, with events

$$e = (u, r, v, a, m, [\tau_s, \tau_e], w_e), \tag{2}$$

where u, v are source and target entities, r a body region, a a candidate-supported surface patch, m the relation, $[\tau_s, \tau_e]$ its interval, and w_e its confidence. Relations include approach, touch, grasp, support, and release. Event confidence w_e concerns the interaction assignment; frame reliability α_t concerns the visual evidence.

For sitting, pelvis/thigh–seat and foot–floor support are separate events. Cup manipulation requires grasp, transport, and release intervals. An active grasp can remain constrained when a frame receives low reliability, preserving the relationship without fitting its erroneous appearance. The harness checks entity and patch identifiers, intervals, and conflicting events, then builds correspondences and temporal activations $g_e(t) \in [0, 1]$. These stay fixed within each solve and are rebuilt after graph revision.

3.5 Graph-Conditioned Geometric Optimization

Starting from the corrected layout, we jointly refine human poses and movable-object trajectories, $\Theta = \{q_{1:T}, (X_{i,1:T})_{i\in\mathcal{M}}\}$, where $\mathcal{M}$ indexes movable objects. The Director's graph and frame weights remain fixed within each solve. We minimize

$$\min_{\Theta} \ \lambda_I \mathcal{L}_I + \lambda_V \mathcal{L}_V + \lambda_c \mathcal{L}_{\text{contact}} + \lambda_s \mathcal{L}_{\text{slip}} + \lambda_p \mathcal{L}_{\text{penetration}} + \lambda_t \mathcal{L}_{\text{temporal}} + \lambda_r \mathcal{L}_{\text{prior}}. \tag{3}$$

Each term is normalized over its valid samples, with weights accounting for differing units. Meshes and scales, static geometry, cameras, and body shape remain fixed.

The input-image term $\mathcal{L}_I$ anchors silhouettes, landmarks, and relative placement. Video evidence enters through $\mathcal{L}_V = Z_V^{-1} \sum_t \alpha_t \ell_{V,t}$, where Z_V normalizes valid weighted samples. Each $\ell_{V,t}$ combines visibility-weighted body reprojection, object tracking, and deviation from recovered motion. Visibility handles occlusion, while the Director's α_t reduces the influence of visible but drifting frames. Robust penalties limit residual outliers, and $\mathcal{L}_{\text{prior}}$ regularizes articulation and intended motion.

The graph complements this visual evidence by maintaining relationships over their active intervals, including frames with low reliability. For a human–object event e, let $h_{e,k,t}$ be a selected human surface point and $a_{e,k}$ its corresponding anchor in scaled object-local coordinates. Their world-space separation is penalized by

$$\mathcal{L}_{\text{contact}} = \frac{1}{Z_c} \sum_{e,k,t} w_e g_e(t)\, \rho(\|h_{e,k,t} - (R_{v,t} a_{e,k} + p_{v,t})\|)\,, \tag{4}$$

where ρ is a robust penalty and Z_c normalizes active samples. The estimator supplies correspondences within the graph-designated patch; event confidence w_e and activation $g_e(t)$ determine their contribution. A coherent patch keeps each grasp or support event attached to the same surface as the entities move. Environment contacts use the same expression with a static target transform.

Proximity alone allows a body point to slide along its target. To maintain sticking relations, we express the point in the target frame, $z_{e,k,t} = R_{v,t}^{\top}(h_{e,k,t} - p_{v,t})$, and penalize tangential relative velocity:

$$\mathcal{L}_{\text{slip}} = \frac{1}{Z_s} \sum_{e \in \mathcal{A}_{\text{stick}},k,t} w_e \bar{g}_e(t) \left\| P_{e,k} \frac{z_{e,k,t+1} - z_{e,k,t}}{\Delta t_t} \right\|^2 , \tag{5}$$

Here Δt_t is the elapsed time, $P_{e,k}$ projects onto the tangent plane, and $\bar{g}_e(t) = \min(g_e(t), g_e(t+1))$ requires both frames to be active. Target-relative motion allows a grasped object to travel with the hand. The constraint applies to sticking relations and is inactive during approach, release, or permitted sliding. Smooth activation ramps avoid abrupt transitions.

Contact and slip constraints are combined with collision and temporal regularization. For a collision field $d_B(x)$ that is negative inside target B, $\mathcal{L}_{\text{penetration}}$ uses $\max(0, \epsilon - d_B(x))^2$ on sampled surface points, with tolerance ϵ and separate handling of intended contacts. Collision proxies replace meshes without a reliable signed interior. The temporal term penalizes changes in root velocity, joint angular velocity, and object velocity using timestamps. Stationary intervals share an object pose; active intervals allow motion. Together with the motion prior, these terms balance stable contact and fidelity to the action.

Refinement first adjusts body placement and object trajectories, then jointly refines pose and contact. Missed contacts, penetration, or excessive deviation inform subsequent Director review; revised events rebuild the affected anchors and constraints before the next solve. The final states, timestamps, graph, and assets form the shared representation used for rendering and simulation.

4 Experiments

4.1 Experimental Setup

For HOI generation, we evaluate 12 actions sampled from FullBodyManipulation (OMOMO) (Li et al., 2023), comparing with CHOIS (Li et al., 2024a) and HOIDiNi (Ron et al., 2025) (Table 1). For reconstruction, we evaluate 24 generated videos of human interactions with scene objects, each 15 seconds long, against CRISP (Wang et al., 2026) and RePHO (Huang et al., 2026) (Table 3). We also ablate spatiotemporal interaction graph optimization (Table 4). All video proposals use MiniMax H3 (MiniMax, 2026). The generation and reconstruction comparisons differ in inputs, outputs, and protocols; their numerical results are not directly comparable across tables.

Table 1: Comparison with HOI generation methods. ↓/↑: lower/higher is better; best in bold.

Method	Pene.Mean ↓	Pene.Max ↓	Contact Rate ↑	R_{prec} ↑
CHOIS	0.0233	0.1212	0.6625	**0.65**
HOIDiNi	0.0118	**0.0231**	0.8327	0.40
Ours	**0.0109**	0.0448	**0.9945**	0.45

Table 2: User study: percentage of trials in which participants preferred our result over all compared methods. Q1: text alignment; Q2: physical plausibility; Q3: motion naturalness; Q4: temporal consistency; Q5: overall preference.

Method	Q1	Q2	Q3	Q4	Q5
Ours (%)	87.13	85.14	85.42	85.03	92.25

Table 3: Video-based reconstruction comparison in nominal 3D units. CRISP models static scenes; moving-object metrics are unavailable (—), so it is omitted from (b).

(a) Image consistency

Method	Human mask IoU ↑	Human mask Dice ↑	Object box IoU ↑	Center error (px) ↓
CRISP	0.520	0.678	—	—
RePHO	0.553	0.708	0.4040	141.17
Ours	**0.621**	**0.745**	**0.479**	**38.56**

(b) Human–object geometry

Method	Wrist proximity rate ↑	Wrist–object distance ↓	Pene.Mean ↓	Pene.Max ↓
RePHO	0.647	0.0741	1.487×10^{-4}	0.00866
Ours	**0.855**	**0.0354**	$\mathbf{2.140\times10^{-5}}$	**0.00309**

Metrics. We measure image consistency with human-mask IoU and Dice, object-box IoU, and box-center error. Human–object geometry uses wrist proximity rate ($d \leq 0.02$ in nominal scene units), mean wrist–object distance, and mean/maximum penetration. HOI generation additionally uses contact rate and text–motion retrieval precision R_{prec}. Image agreement is assessed alongside geometry because generated frames can contain interaction errors. Appendix C defines all metrics.

4.2 Human–Object Interaction Generation

EmboDirector achieves the highest contact rate (0.9945, versus 0.8327 for HOIDiNi and 0.6625 for CHOIS) and the lowest mean penetration (0.0109). Maximum penetration is higher than HOIDiNi's (0.0448 versus 0.0231), indicating a trade-off between average contact geometry and worst-case intersection.

CHOIS obtains the highest R_{prec} (0.65), followed by EmboDirector (0.45) and HOIDiNi (0.40). This retrieval metric measures text–motion alignment; our image-and-text-conditioned video generation differs from direct text-to-motion synthesis.

To complement the retrieval metric, we asked 30+ participants to compare 10 interaction sequences in randomized method order and select the best result for each of five questions. EmboDirector receives 87.13% preference for text alignment, 92.25% overall, and over 85% on every question (Table 2).

4.3 Reconstruction Quality

Table 3(a) shows improved image agreement. Object-box IoU increases from 0.4040 for RePHO to 0.479, while center error decreases from 141.17 to 38.56 px. Human-mask IoU is 0.621, compared with 0.553 for RePHO and 0.520 for CRISP. Figure 5 compares projected bodies and object localization in two interaction examples. CRISP's static representation cannot track moving objects.

Figure 4 shows a tipped chair reoriented upright and a pan moved from beneath a table onto its surface. In the supplementary demo, the optimized object motion remains coherent despite drift in later generated frames.

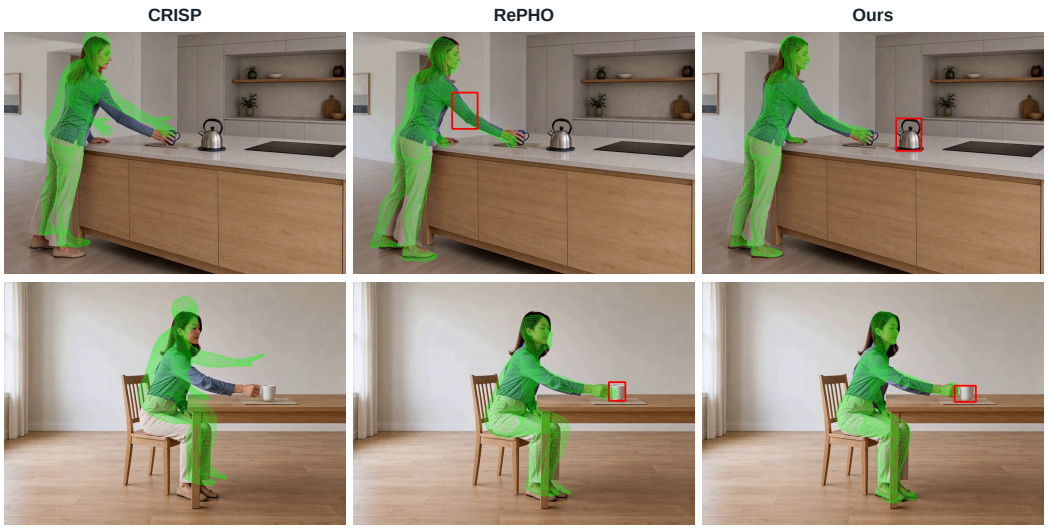


Figure 5: Qualitative comparison with CRISP and RePHO. Projected human meshes (green) and object localization (red boxes) are shown for two interaction examples.

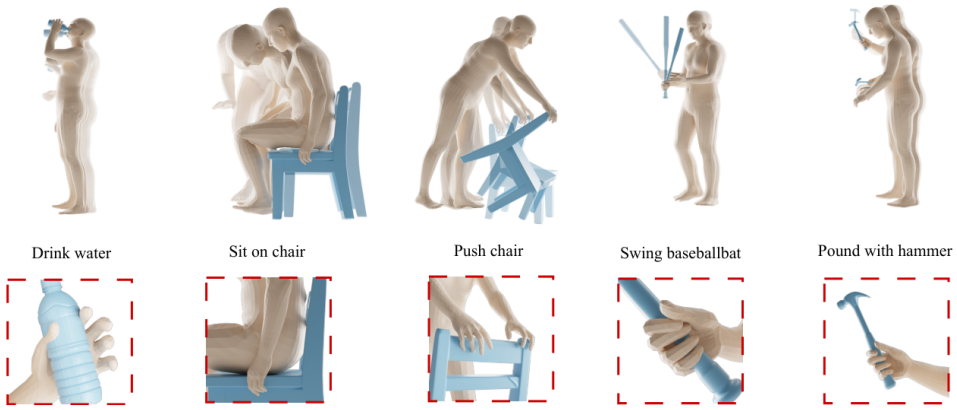


Figure 6: Reconstructions for five actions. Temporal overlays show the motion; insets zoom in on the contact regions.

Table 3(b) shows wrist proximity increasing from 0.647 to 0.855 and mean wrist–object distance decreasing from 0.0741 to 0.0354. Mean and maximum penetration decrease by about $7\times$ and $2.8\times$, respectively. These results reflect the full reconstruction, combining graph-conditioned alignment with collision penalties.

Figure 6 shows drinking, sitting, chair pushing, bat swinging, and hammer use, with close-ups of hand and pelvis contact. Appendix B.1 compares the geometric contact candidates supplied to the Director.

4.4 Spatiotemporal Graph Ablation

To isolate graph construction, we replace the Director graph with geometric rules, sharing the Director-corrected scene, contact candidates, hand configurations, initialization, losses, and 150-iteration budget. Rules use distance hysteresis and a relative-velocity threshold; the graph remains fixed during optimization. Table 4 reports pre-smoothing sequence means within contact intervals independently annotated by humans and shared by both variants; Floor Pene.Max averages sequence maxima in nominal units. This protocol differs from the main comparisons.

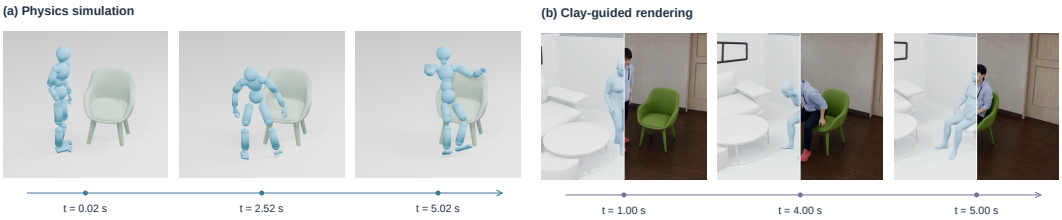


Figure 7: Chair interaction in simulation (left) and clay-guided synthesis (right; clay/appearance pairs). Each panel shows three stages with sequence-relative timestamps.

Table 4: Graph-construction ablation.

Graph	Human IoU ↑	Object-box IoU ↑	Contact Rate ↑	Foot skating ↓	Floor Pene.Max ↓
Director	0.6326	0.8741	**0.8875**	0.0941	**0.0203**
Geometric rules	**0.6400**	**0.8862**	0.1978	**0.0867**	0.0225

The Director graph raises the contact rate from 0.1978 to 0.8875 (+69.0 percentage points), evaluated on shared human-annotated contact intervals. While both IoUs stay within 0.013 of the rule-based variant, the additional contact thus does not come at the expense of image agreement. Foot–floor behavior is comparable: the Director graph increases foot skating by 8.6% but lowers the mean maximum floor penetration by 9.7%.

Graph construction determines which contacts to enforce and when; graph-conditioned optimization enforces them. Removing the graph-derived contact and slip terms while keeping all other terms lowers the wrist proximity rate from 0.9715 to 0.2823. These results support both components in maintaining contact.

4.5 Simulation and Rendering

Figure 7 illustrates physics-based tracking by a frozen InterMimic controller (Xu et al., 2025) (Appendix B) and novel-view synthesis. Simulation panels show physics-engine states. For synthesis, MiniMax H3 (MiniMax, 2026) combines clay renders $V_{\text{clay}} = \mathcal{R}(\mathcal{S}, C'_{1:T})$, appearance references, and the description. Supplementary videos show both outputs.

5 Limitations

EmboDirector assumes a single person, rigid objects, and a fixed background. Severe occlusion and incomplete geometry can make contact ambiguous. Errors in the Director's layout, frame weighting, or timing decisions can propagate into refinement. Simulation execution also depends on the controller and physical model. Extending the representation to articulated and deformable objects would broaden its scope.

6 Conclusion

EmboDirector reconstructs 4D human–scene interactions from a single image and description, using generated video as a motion proposal. The Director corrects recovered layouts against the input image, down-weights drifting frames, and organizes contacts into a spatiotemporal graph. Joint optimization preserves these relations while refining human and object trajectories. Comparisons show stronger contact maintenance, lower mean penetration, and better object alignment; ablations support the complementary roles of graph construction and optimization. The reconstructed interactions guide physics-based tracking and novel-view synthesis.

Reproducibility Statement

Section 3 defines the scene representation, layout correction, frame reliability, interaction graph, and optimization objective. Appendix A describes the Director interface and how semantic events are bound to geometric constraints. Appendix B details body conversion, contact correspondence recovery, and trajectory refinement, and Appendix C defines the reported metrics. Videos are generated with MiniMax H3 (MiniMax, 2026), and the Director uses GPT-6 Astra (OpenAI, 2026) within an author-developed harness that constructs and revises the interaction graph.

Ethics Statement

Our outputs depict synthetic human interactions. Responsible use requires appropriate permissions for source images, human assets, and evaluation data, and clear labeling of synthetic depictions of identifiable people.

AI Use Statement

We used generative AI tools to assist with research ideation and execution, including discussing and refining aspects of the methodology and experimental design. We also used AI tools for literature retrieval and discovery, reference checking, manuscript organization, language revision, drafting and revising portions of the method descriptions and evaluation protocols, and figure preparation. MiniMax H3 was used to generate the interaction-video proposals evaluated in this work, as described in Section 3.2, and GPT-6 Astra was used as the MLLM Director within the proposed method, as described in Section 3.4. All AI-assisted research decisions, generated data, manuscript content, citations, experimental protocols, and reported claims were reviewed and validated by the authors. The authors take full responsibility for the contents of this submission.

References

B. L. Bhatnagar, X. Xie, I. A. Petrov, C. Sminchisescu, C. Theobalt, and G. Pons-Moll. BEHAVE: Dataset and method for tracking human object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15935–15946, 2022.

Yukang Cao, Liang Pan, Kai Han, Kwan-Yee K. Wong, and Ziwei Liu. AvatarGO: Zero-shot 4D human-object interaction generation and animation. *arXiv preprint arXiv:2410.07164*, 2024.

Yukang Cao, Haozhe Xie, Fangzhou Hong, Long Zhuo, Zhaoxi Chen, Liang Pan, and Ziwei Liu. HSImul3R: Physics-in-the-loop reconstruction of simulation-ready human-scene interactions. *arXiv preprint arXiv:2603.15612*, 2026. URL https://arxiv.org/abs/2603.15612.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, et al. SAM 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025.

Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, Chao Weng, and Ying Shan. VideoCrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.19512*, 2023.

X. Chen, F. J. Chu, P. Gleize, K. J. Liang, A. Sax, H. Tang, W. Wang, M. Guo, T. Hardin, X. Li, et al. SAM 3D: 3Dfy anything in images. *arXiv preprint arXiv:2511.16624*, 2025.

Alpár Cseke, Shashank Tripathi, Sai Kumar Dwivedi, Arjun S. Lakshmipathy, Agniv Chatterjee, Michael J. Black, and Dimitrios Tzionas. PICO: Reconstructing 3D people in contact with objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1783–1794, 2025.

Sisi Dai and Kai Xu. AnchorHOI: Zero-shot generation of 4D human-object interaction via anchor-based prior distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026.

Karthik Dharmarajan, Wenlong Huang, Jiajun Wu, Li Fei-Fei, and Ruohan Zhang. Dream2Flow: Bridging video generation and open-world manipulation with 3D object flow. *arXiv preprint arXiv:2512.24766*, 2025.

S. K. Dwivedi, D. Antić, S. Tripathi, O. Taheri, C. Schmid, M. J. Black, and D. Tzionas. InteractVLM: 3D interaction reasoning from 2D foundational models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2025.

Mingqi Gao, Yunqi Miao, and Jungong Han. SAM-Body4D: Training-free 4D human body mesh recovery from videos. *arXiv preprint arXiv:2512.08406*, 2025.

Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J. Black. Resolving 3D human pose ambiguities with 3D scene constraints. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019. URL https://arxiv.org/abs/1908.06963.

Mohamed Hassan, Duygu Ceylan, Ruben Villegas, Jun Saito, Jimei Yang, Yi Zhou, and Michael J. Black. Stochastic scene-aware motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021a.

Mohamed Hassan, Partha Ghosh, Joachim Tesch, Dimitrios Tzionas, and Michael J. Black. Populating 3D scenes by learning human-scene interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021b. URL https://posa.is.tue.mpg.de/.

Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P. Kingma, Ben Poole, Mohammad Norouzi, David J. Fleet, and Tim Salimans. Imagen Video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022.

Dingbang Huang, Etienne Vouga, Qixing Huang, and Georgios Pavlakos. Recovering physically plausible human-object interactions from monocular videos. *arXiv preprint arXiv:2606.05359*, 2026. URL https://arxiv.org/abs/2606.05359.

Siyuan Huang, Zan Wang, Puhao Li, Baoxiong Jia, Tengyu Liu, Yixin Zhu, Wei Liang, and Song-Chun Zhu. Diffusion-based generation, optimization, and planning in 3D scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. URL https://arxiv.org/abs/2301.06015.

N. Jiang, Z. He, Z. Wang, H. Li, Y. Chen, S. Huang, and Y. Zhu. Autonomous character-scene interaction synthesis from text instruction. In *SIGGRAPH Asia 2024 Conference Papers*, pp. 1–11, 2024a.

Nan Jiang, Zhiyuan Zhang, Hongjie Li, Xiaoxuan Ma, Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, and Siyuan Huang. Scaling up dynamic human-scene interaction modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024b. URL https://arxiv.org/abs/2403.08629.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), 2023.

Hyeonwoo Kim, Sangwon Baik, and Hanbyul Joo. DAViD: Modeling dynamic affordance of 3D objects using pre-trained video diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025. URL https://arxiv.org/abs/2501.08333.

Kling Team, Jialu Chen, Yuanzheng Ci, Xiangyu Du, Zipeng Feng, et al. Kling-Omni technical report. *arXiv preprint arXiv:2512.16776*, 2025.

Zeqiang Lai, Yunfei Zhao, Haolin Liu, Zibo Zhao, Qingxiang Lin, et al. Hunyuan3D 2.5: Towards high-fidelity 3D assets generation with ultimate details. *arXiv preprint arXiv:2506.16504*, 2025.

Vincent Leroy, Yohann Cabon, and Jerome Revaud. Grounding image matching in 3D with MASt3R. In *European Conference on Computer Vision*, 2024.

Hongjie Li, Hong-Xing Yu, Jiaman Li, and Jiajun Wu. ZeroHSI: Zero-Shot 4D Human-Scene Interaction by Video Generation. In *International Conference on 3D Vision (3DV)*, 2026a. URL https://awfuact.github.io/zerohsi/.

J. Li, A. Clegg, R. Mottaghi, J. Wu, X. Puig, and C. K. Liu. Controllable Human-Object Interaction Synthesis. In *European Conference on Computer Vision*, pp. 54–72. Springer, 2024a.

Jiaman Li, Jiajun Wu, and C. Karen Liu. Object motion guided human motion synthesis. *ACM Transactions on Graphics*, 42(6), 2023. URL https://arxiv.org/abs/2309.16237.

Lei Li and Angela Dai. GenZI: Zero-shot 3D human-scene interaction generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20465–20474, 2024. URL https://arxiv.org/abs/2311.17737.

Lei Li and Angela Dai. HOI-PAGE: Zero-shot human-object interaction generation with part affordance guidance. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2506.07209.

Xinyang Li, Chengjie Yi, Jiawei Lai, Mingbao Lin, Yansong Qu, Shengchuan Zhang, and Liujuan Cao. SynergyAmodal: Deocclude anything with text control. *arXiv preprint arXiv:2504.19506*, 2025.

Zekun Li, Rui Zhou, Rahul Sajnani, Xiaoyan Cong, Daniel Ritchie, and Srinath Sridhar. GenHSI: Controllable generation of human-scene interaction videos. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2026b. URL https://ivl.cs.brown.edu/research/genhsi.html.

Zhe Li, Zerong Zheng, Lizhen Wang, and Yebin Liu. Animatable gaussians: Learning pose-dependent gaussian maps for high-fidelity human avatar modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19711–19722, 2024b.

Haotong Lin, Sili Chen, Junhao Liew, Donny Y. Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth Anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025.

Yumeng Liu, Xiaoxiao Long, Zemin Yang, Yuan Liu, Marc Habermann, Christian Theobalt, Yuexin Ma, and Wenping Wang. EasyHOI: Unleashing the power of large models for reconstructing hand-object interactions in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7037–7047, 2025.

Zhizheng Liu, Joe Lin, Wayne Wu, and Bolei Zhou. Joint optimization for 4D human-scene reconstruction in the wild. In *International Conference on Learning Representations*, 2026. URL https://vail.cs.ucla.edu/JOSH/.

Yuke Lou, Yiming Wang, Zhen Wu, Rui Zhao, Wenjia Wang, Mingyi Shi, and Taku Komura. Zero-shot human-object interaction synthesis with multimodal priors. *arXiv preprint arXiv:2503.20118*, 2025. URL https://arxiv.org/abs/2503.20118.

Yanzhe Lyu, Chen Geng, Karthik Dharmarajan, Yunzhi Zhang, Hadi AlZayer, Shangzhe Wu, and Jiajun Wu. Choreographing a world of dynamic objects. *arXiv preprint arXiv:2601.04194*, 2026.

MiniMax. MiniMax H3: An open model breaking the boundaries between tasks and modalities, 2026. URL https://www.minimax.io/blog/minimax-h3. Model release, July 31, 2026.

Lingzhou Mu, Qiang Wang, Fan Jiang, Mengchao Wang, Yaqi Fan, Mu Xu, and Kai Zhang. FantasyHSI: Video-generation-centric 4D human synthesis in any scene through a graph-based multi-agent framework. *arXiv preprint arXiv:2509.01232*, 2025. URL https://arxiv.org/abs/2509.01232.

Hyeongjin Nam, Daniel Sungho Jung, Gyeongsik Moon, and Kyoung Mu Lee. Joint reconstruction of 3D human and object via contact-based refinement transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10218–10227, 2024.

OpenAI. GPT-6 Astra model. OpenAI API documentation, 2026. URL https://developers.openai.com/api/docs/models/gpt-6-astra. Accessed September 26, 2026.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10975–10985, 2019. URL https://smpl-x.is.tue.mpg.de/.

Xiaogang Peng, Yiming Xie, Zizhao Wu, Varun Jampani, Deqing Sun, and Huaizu Jiang. HOI-Diff: Text-driven synthesis of 3D human-object interactions using diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 2878–2888, 2025.

Jathushan Rajasegaran, Georgios Pavlakos, Angjoo Kanazawa, and Jitendra Malik. Tracking people by predicting 3D appearance, location and pose. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2740–2749, 2022.

R. Ron, G. Tevet, H. Sawdayee, and A. H. Bermano. HOIDiNi: Human-Object Interaction through Diffusion Noise Optimization. *arXiv preprint arXiv:2506.15625*, 2025.

SpatialVerse Research Team, Manycore Tech Inc. InteriorGS: A 3D gaussian splatting dataset of semantically labeled indoor scenes, 2025. URL https://huggingface.co/datasets/spatialverse/InteriorGS.

Omid Taheri, Nima Ghorbani, Michael J. Black, and Dimitrios Tzionas. GRAB: A dataset of whole-body human grasping of objects. In *European Conference on Computer Vision*, 2020. URL https://grab.is.tue.mpg.de/.

Ruicheng Wang, Sicheng Xu, Cassie Dai, Jianfeng Xiang, Yu Deng, Xin Tong, and Jiaolong Yang. MoGe: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5261–5271, 2025a.

Ruicheng Wang, Sicheng Xu, Yue Dong, Yu Deng, Jianfeng Xiang, Zelong Lv, Guangzhong Sun, Xin Tong, and Jiaolong Yang. MoGe-2: Accurate monocular geometry with metric scale and sharp details. In *Advances in Neural Information Processing Systems*, 2025b.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. DUSt3R: Geometric 3D vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20697–20709, 2024.

Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. HUMANISE: Language-conditioned human motion generation in 3D scenes. In *Advances in Neural Information Processing Systems*, 2022.

Zihan Wang, Jiashun Wang, Jeff Tan, Yiwen Zhao, Jessica Hodgins, Shubham Tulsiani, and Deva Ramanan. Contact-guided Real2Sim from Monocular Video with Planar Scene Primitives. In *International Conference on Learning Representations*, 2026. URL https://crisp-real2sim.github.io/CRISP-Real2Sim/.

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

Sirui Xu, Zhengyuan Li, Yu-Xiong Wang, and Liang-Yan Gui. InterDiff: Generating 3D human-object interactions with physics-informed diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14928–14940, 2023. URL https://arxiv.org/abs/2308.16905.

Sirui Xu, Ziyin Wang, Yu-Xiong Wang, and Liang-Yan Gui. InterDreamer: Zero-shot text to 3D dynamic human-object interaction. In *Advances in Neural Information Processing Systems*, pp. 52858–52890, 2024.

Sirui Xu, Hung Yu Ling, Yu-Xiong Wang, and Liang-Yan Gui. InterMimic: Towards universal whole-body control for physics-based human-object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12266–12277, 2025. URL https://yxw.cs.illinois.edu/files/InterMimic_CVPR2025.pdf.

Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth Anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10371–10381, 2024a.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth Anything V2. In *Advances in Neural Information Processing Systems*, 2024b.

Xitong Yang, Devansh Kukreja, Don Pinkus, Anushka Sagar, Taosha Fan, Jinhyung Park, Soyong Shin, Jinkun Cao, Jiawei Liu, Nicolas Ugrinovic, Matt Feiszli, Jitendra Malik, Piotr Dollár, and Kris Kitani. SAM 3D Body: Robust full-body human mesh recovery. *arXiv preprint arXiv:2602.15989*, 2026. URL https://github.com/facebookresearch/sam-3d-body.

Yuhang Yang, Wei Zhai, Hongchen Luo, Yang Cao, and Zheng-Jun Zha. LEMON: Learning 3D human-object interaction relation from 2D images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16284–16295, 2024c.

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. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/ce31378e9f41d8907e97dab172b6c559-Abstract-Conference.html.

Wei Yao, Yunlian Sun, Hongwen Zhang, Yebin Liu, and Jinhui Tang. HOSIG: Full-body human-object-scene interaction generation with hierarchical scene perception. *arXiv preprint arXiv:2506.01579*, 2025.

Ling-An Zeng, Guohong Huang, Yi-Lin Wei, Shengbo Gu, Yu-Ming Tang, Jingke Meng, and Wei-Shi Zheng. ChainHOI: Joint-based kinematic chain modeling for human-object interaction generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12358–12369, 2025.

A Director Interface

The harness supplies object identifiers, mesh-local frames, camera parameters, timestamps, a body-region map, and contact candidates from the occlusion-aware estimator. Entity-labeled renders associate visible evidence with recovered geometry. The Director returns corrections to initial object poses, frame reliability weights α_t, and interval-based interaction events. The input image anchors layout correction; generated frames supply motion evidence whose reliability is assessed separately. Table 5 lists the event fields and associated outputs.

Table 5: Director event fields and associated outputs.

Field	Meaning and geometric binding
Source / region	Entity u and body region r, resolved to geometry.
Target / surface	Entity v and candidate-supported mesh-local patch a.
Relation	Type m: approach, touch, grasp, support, or release.
Time interval	Ordered times $[\tau_s, \tau_e]$ within the sequence.
Event confidence	Weight w_e for the interaction assignment.
Evidence	Frames and visible cues supporting the event and its timing.
Contact behavior	Whether the event activates sticking constraints.
Layout correction	Initial object-pose adjustment relative to recovery.
Frame reliability	Weight α_t controlling video evidence at frame t.

Contact coordinates originate in the estimator and its mesh projection. The Director refers to the resulting body regions and candidate surface patches when organizing events; the harness resolves these identifiers to the correspondences and anchors used by the optimizer. A grasp event determines the target object, relation, and active interval. If the accepted initial layout changes, the harness recomputes the corresponding transforms and anchors.

Frame reliability and event confidence express different judgments. A frame can contain an implausible change in cup or kettle orientation while the grasp relationship remains valid. Reducing α_t weakens agreement with that frame without removing the active grasp event. The event's w_e instead controls confidence in the region–surface assignment. Visibility weights account for occlusion, so a clearly visible but drifting object can still receive low temporal reliability. The action, grasp, and surrounding motion provide context for distinguishing legitimate rotation from drift.

A layout proposal identifies the object and a rigid transform relative to its recovered initial pose; scale is handled separately. Bounds and checks on silhouettes, relative depth, support, and collisions keep corrections tied to the input scene. These adjustments correct initialization, while time-varying trajectories describe motion during the action.

For sitting, the graph separates pelvis/thigh–seat support from foot–floor support. For cup manipulation, candidate correspondences are organized into approach, grasp, transport, and release phases; transport maintains the grasp relation rather than introducing a new contact-point prediction. The harness checks identifiers, nonempty patches, ordered intervals, and contradictory assignments before building temporal activations and constraints.

Within each solve, static geometry, frame reliability, event intervals, and mesh-local target correspondences remain fixed. Human poses and movable-object trajectories are refined from the corrected initialization under weighted visual evidence and graph-derived constraints. A subsequent Director review uses geometric residuals to revise events; the harness rebuilds affected anchors and constraints before refinement resumes. This separates the Director's layout, evidence, and temporal decisions from geometric correspondence handling and trajectory optimization.

B Representation and Optimization Details

The body conversion maps the native MHR output of SAM 3D Body to the SMPL-X representation used by the optimizer. Let $\widehat{J}_t$ and $\widehat{U}_t$ denote the recovered joints and sampled surface observations.

The fitting objective combines joint alignment, surface agreement, and body priors:

$$\min_{\beta, q_{1:T}} \sum_t \left[\alpha_J \left\| W_J(J(\beta, q_t) - \widehat{J}_t) \right\|^2 + \alpha_U D(U(\beta, q_t), \widehat{U}_t) \right] + \alpha_\beta \|\beta\|^2 + \alpha_q \mathcal{L}_{\text{pose}}. \quad (6)$$

Here W_J encodes the joint correspondences between the two models and D measures surface agreement; both account for differences in topology and anatomy. Body shape is shared over the sequence, and the pose includes the global transform. Hands are fitted only to reliable observations, so that occluded fingers do not introduce unsupported contact constraints.

B.1 Occlusion-Aware Contact Correspondences

The estimator in Section 3.3 converts image-space occlusion cues into candidate human–object correspondences. InteractVLM supplies body contact seeds, which we project onto the image as a contact mask $\mathcal{M}_{\text{hc}}$. For a human-occludes-object configuration, we intersect this mask with the difference between the rendered object silhouette and the observed object mask. The resulting pixels identify candidate locations where the human may meet an occluded object surface. Both occlusion directions are handled before geometric filtering.

A camera ray can intersect several surface layers of either mesh. We therefore use depth peeling to render multiple human and object layers, ordered from front to back. When the human occludes the object, we associate the last human layer with the first object layer; when the object occludes the human, we associate the first human layer with the last object layer. The selected layer depths and camera calibration determine the 3D points obtained by back-projection.

Human self-occlusion requires additional care because both depth orderings can occur in the same image region. Subtracting the entire observed human foreground would remove some valid correspondences behind the object. Let $\mathcal{M}_{\text{h}}^{\text{obs}}$ be the human foreground mask, $\mathcal{M}_{\text{o}}^{\text{rend}}$ the rendered object mask, and $\mathcal{M}_{\text{hcc}}$ the self-occlusion mask of the projected human contact region, obtained from multi-layer rendering. With Ω denoting the image domain, the corrected back-side mask is

$$\mathcal{M}_{\text{back}} = \mathcal{M}_{\text{hc}} \cap \mathcal{M}_{\text{o}}^{\text{rend}} \cap \left[\Omega \setminus \left(\mathcal{M}_{\text{h}}^{\text{obs}} \setminus \mathcal{M}_{\text{hcc}} \right) \right]. \quad (7)$$

The self-occlusion mask preserves contact-layer pixels that would otherwise be removed with the human foreground. We back-project the corrected masks using the corresponding peeled layers to obtain candidate pairs (p_h, p_o).

Occlusion alone is insufficient to establish contact. For unit surface normals $\mathbf{n}_h$ and $\mathbf{n}_o$, we retain a pair only if

$$\langle \mathbf{n}_h, -\mathbf{n}_o \rangle > \cos 60^\circ, \qquad \|p_h - p_o\|_2 < 0.1\,\text{m}. \quad (8)$$

Thus the normals must oppose one another within 60°, and the points must be less than 10 cm apart in the body-scale-aligned reconstruction (Section 3.2). This cutoff filters candidate surface pairs before optimization; the wrist-proximity metric in Appendix C instead thresholds the reconstructed wrist–object distance. For a pair recovered on object v at frame t, the scaled object-local anchor is $a = R_{v,t}^\top (p_o - p_{v,t})$, consistent with Eq. (4); its world position follows the optimized object pose. The retained body regions and object anchors provide the geometric input to the Director, which organizes their interaction roles and active intervals as described in Section 3.4. The graph then determines when the corresponding contact and sticking constraints act during trajectory refinement.

Figure 8 compares the resulting contact regions with InteractVLM on drinking, hammer use, and chair interaction.

B.2 Temporal Refinement and Execution

Contact activation is smoothed at the start and end of each event. For a normalized ramp coordinate $\xi \in [0, 1]$, the cubic $s(\xi) = 3\xi^2 - 2\xi^3$ gives a smooth transition to constant activation within the interval, with the ramp duration specified in physical time. Since relation types determine which losses apply, smoothing an approach or release transition does not impose sticking contact during that phase. Adjacent events are checked for incompatible target assignments.

Translational velocities are computed as $(p_{t+1} - p_t)/\Delta t_t$, and accelerations as differences of consecutive velocities over the corresponding elapsed time. Rotational increments are computed from

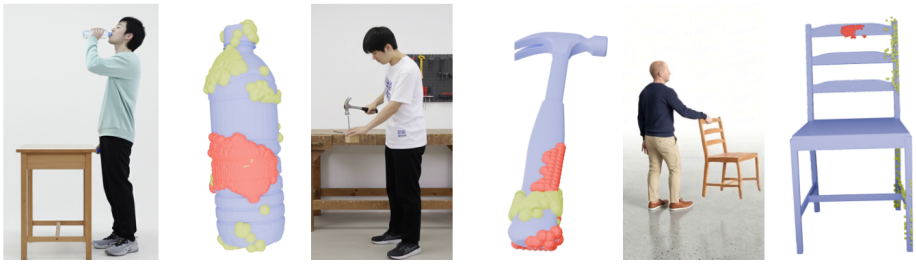


Figure 8: Contact predictions from our estimator (red) and InteractVLM (yellow) for drinking, hammer use, and chair interaction.

the Lie-group logarithm of $R_t^\top R_{t+1}$ in a consistent tangent frame. The same finite differences are applied to body and object variables. Stationary object intervals share a single pose, which builds stationarity into the trajectory representation instead of relying only on a large velocity penalty.

Contact fitting uses distances to the target surface, whereas penetration uses a signed interior defined by the collision geometry. For incomplete or open recovered meshes, a collision proxy supplies this interior. Intended contacts are allowed a small separation tolerance. Keeping surface correspondences and collision geometry separate lets the optimizer fit local contact while still controlling unintended intersections.

For simulation, we map body pose, root motion, and object trajectories into the simulator's frame and scale. SMPL-X poses become native joint references through a fixed skeleton correspondence, checked using forward kinematics and joint limits. The humanoid and objects start from the first reconstructed frame. A frozen InterMimic controller (Xu et al., 2025) reads current and upcoming reference frames and outputs joint PD targets; the physics engine advances the state. The root and dynamic objects respond to forces and are not reset to their recovered trajectories. The displayed rollout is rendered from these simulated states. Body correspondences, collision geometry, physical parameters, and controller settings determine how the reference is executed.

For video synthesis, we render all surfaces with uniform, untextured materials along a camera path $C'_{1:T}$. The clay sequence preserves the refined geometry, motion, occlusions, and contacts, while appearance references provide visual style. MiniMax H3 (MiniMax, 2026) conditions on both, together with the interaction description. Figure 7 pairs clay and synthesized frames from a viewpoint different from the input and generated video. Changing viewpoints requires only rerendering the refined scene. The simulation examples cover a small number of sequences as qualitative demonstrations; supplementary videos show the executed rollouts and synthesized sequences.

C Evaluation Metrics

We report image agreement, human–object geometry, HOI generation quality, and user preference. Tables 1 and 3 use separate evaluation settings, with 3D quantities expressed in the coordinates of each setting, so comparisons are meaningful only within a table.

For binary human masks A and B, intersection over union is $|A \cap B|/|A \cup B|$ and Dice is $2|A \cap B|/(|A| + |B|)$. Object-box IoU applies the same ratio to image-space bounding boxes, and center error is the pixel distance between box centers.

The wrist proximity rate in Table 3 is the fraction of frames in which the wrist–object distance satisfies $d \leq 0.02$ in nominal scene units. The mean wrist–object distance reports the magnitude of the remaining separation, which the thresholded rate does not capture.

For a signed target field $d_B(x)$, the penetration depth of a point is $\max(0, -d_B(x))$. Pene.Mean and Pene.Max are the mean and maximum depth between the body mesh and scene or object geometry: the mean reflects typical intersection severity, and the maximum the worst case. Both are expressed in the nominal coordinates of the respective evaluation.

The HOI evaluation does not require ground-truth trajectories. In addition to Pene.Mean and Pene.Max, it reports the Contact Rate, the fraction of frames in which the hand lies within the contact threshold of the target object, and R_{prec} (top-3), which measures whether a text–motion retrieval model matches the generated motion to its description.

D VISUAL PRIORS AND SCENE REPRESENTATIONS

Video generation provides temporal hypotheses from sparse conditioning. Imagen Video uses cascaded diffusion for text-conditioned synthesis (Ho et al., 2022), and VideoCrafter1 provides text-to-video and image-to-video models (Chen et al., 2023). CogVideoX and Wan scale up generative video modeling (Yang et al., 2025; WanTeam et al., 2025), and Kling-Omni supports generation and editing from multimodal inputs (Kling Team et al., 2025). We use MiniMax H3 (MiniMax, 2026). The Director operates on the generated frames and the recovered geometry, where interaction relationships that are only implicit in the video can be inspected directly.

Recovering a common scene frame builds on recent advances in visual geometry. Depth Anything and Depth Anything V2 estimate monocular depth (Yang et al., 2024a;b), and Depth Anything 3 extends spatial prediction to arbitrary sets of views (Lin et al., 2025). MoGe predicts affine-invariant point maps, and MoGe-2 estimates metric geometry (Wang et al., 2025a;b). DUSt3R regresses pairwise point maps, and MASt3R builds matching on top of this representation (Wang et al., 2024; Leroy et al., 2024). These works separate image alignment, relative geometry, and metric scale, all of which must share a consistent frame before contact can be optimized.

SAM 3 segments and tracks concept-prompted instances (Carion et al., 2025), and SynergyAmodal completes occluded appearance and shape in images (Li et al., 2025). PHALP associates people across frames through predicted 3D appearance, location, and pose (Rajasegaran et al., 2022), and SAM-Body4D improves temporal mesh recovery with occlusion-aware masklets (Gao et al., 2025). In EmboDirector, SAM 3D Objects and SAM 3D Body supply the geometric initialization (Chen et al., 2025; Yang et al., 2026), and SMPL-X provides the articulated body model (Pavlakos et al., 2019). The interaction graph binds semantic contact decisions to explicit surfaces on these recovered assets.

Appearance representations play a different role from contact geometry. 3D Gaussian Splatting enables real-time view synthesis, and Animatable Gaussians models pose-dependent human appearance (Kerbl et al., 2023; Li et al., 2024b). InteriorGS provides indoor Gaussian scenes with semantic and spatial annotations (SpatialVerse Research Team, Manycore Tech Inc., 2025), and Hunyuan3D 2.5 generates textured 3D assets (Lai et al., 2025). We use explicit meshes and transforms instead, so that interacting surfaces are directly available to the optimizer and the renderer. Among video-to-action systems, Dream2Flow uses reconstructed 3D object flow to guide manipulation, and CHORD distills multi-object dynamics from video priors (Dharmarajan et al., 2025; Lyu et al., 2026). Our output retains human and object trajectories together with their contact relationships, and InterMimic serves as the downstream simulation controller (Xu et al., 2025).