PACHO: PARALLEL AFFORDANCE AND CONTACT DIFFUSION FOR HUMAN-OBJECT INTERACTION SYNTHESIS

Anonymous authors
Paper under double-blind review

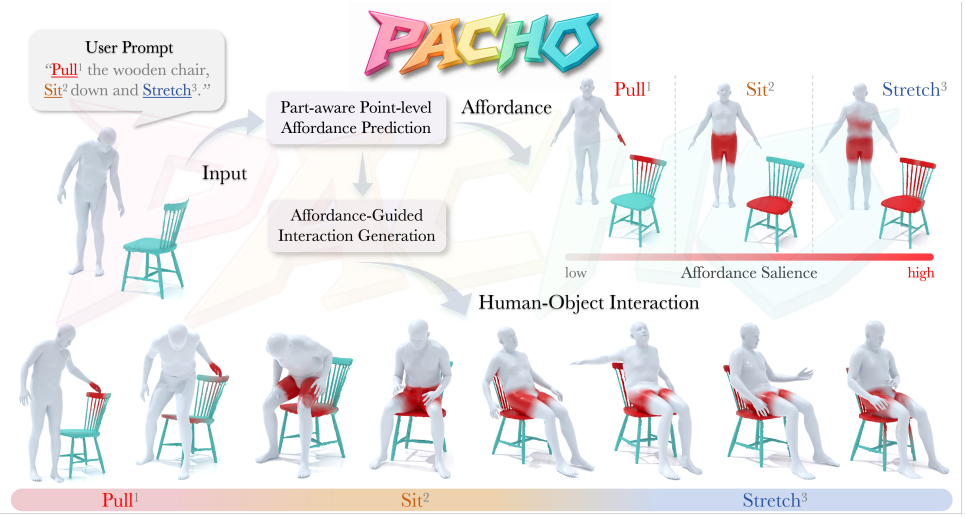


Figure 1: **Text-driven 3D HOI Synthesis with PACHO**. Given a text prompt, PACHO generates coherent interaction motions by explicitly modeling the dynamic physical "handshake". By predicting the human contact regions and functional object affordances (in warm colors), PACHO guides the motion synthesis to achieve physically plausible and context-aware interactions.

ABSTRACT

Text-driven 3D human-object interaction (HOI) synthesis requires precise spatiotemporal synchronization and stable physical contact between the human and object. Existing approaches struggle to capture the spatiotemporal co-evolution of localized human-contact and object-affordance regions, leading to loose physical coupling. To provide fine-grained spatial guidance, our key insight is to explicitly model the dynamic "handshake" between entities: the tightly coupled, time-varying human contact regions and functional object affordances. We propose **PACHO**, a novel diffusion framework that predicts object affordance and human contact maps in parallel to guide human–object motion denoising. Specifically, PACHO utilizes Dual Motion-Prior Learners to extract decoupled kinematic representations for both the human and the object. To bridge these independent streams, we introduce a Cross-Affordance Interaction module that performs adaptive, part-aware point-level contact-affordance reasoning to establish a contact-aware bidirectional physical coupling. During denoising, a joint motion-salience objective supervises the motion alongside the affordance and contact, encouraging local geometric consistency. Extensive experiments demonstrate that PACHO achieves state-of-the-art performance, with clear improvements in physical realism, interpenetration reduction and contact stability.

1 INTRODUCTION

3D human-object interaction (HOI) synthesis aims to generate coordinated human-object motions conditioned on a textual description and object geometry, and is a key step toward building interactive

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.

3D worlds. Unlike human-only motion generation, realistic HOI synthesis requires coherent object dynamics, accurate contact timing, and stable human-object contact. Many real-world interactions, such as sitting on chairs, lifting monitors, or manipulating tools, involve dynamic contact regions that shift across human body parts and functional regions of objects over time. Although recent studies have explored diffusion-based multimodal motion generation and fine-grained 3D mesh modeling Dai et al. (2024); Cai et al. (2026); Xu et al. (2026a), generating physically plausible and contact-consistent 3D HOI sequences remains a fundamental challenge.

Recent HOI generators mainly adopt joint, sequential, or decoupled human-object modeling strategies Petrov et al. (2025); He et al. (2025); Li et al. (2023); Wu et al. (2025a); Jung et al. (2026). While these approaches produce plausible global motions, they only partially capture the dynamic interaction structure underlying HOI. Contact-guided methods typically emphasize one-sided contact cues, affordance-based methods often rely on sparse points or discrete object regions, and decoupled models do not explicitly maintain paired human-contact and object-affordance fields over time. Consequently, how interacting human body parts correspond to functional object regions and how such correspondences evolve throughout an action remains insufficiently modeled.

To address this gap, we introduce a dynamic, affordance-and-contact-guided perspective for 3D HOI generation. Our key insight is that realistic HOI synthesis requires explicitly modeling the dynamic "handshake" between the human and the object. This perspective views contact and affordance as two mutually dependent interaction signals, and their co-evolution provides fine-grained spatial cues and serves as an intermediate structure connecting high-level textual intent with low-level geometric realization. Realizing this perspective requires addressing three coupled challenges. First, the model must preserve entity-specific motion priors before interaction coupling to maintain the distinct kinematic regularities of articulated humans and rigid objects. Second, it must establish bidirectional contact-affordance awareness by jointly inferring complementary interaction regions alongside the evolving motion context. Third, this semantic correspondence must be geometrically grounded throughout denoising via explicit point-level alignment, which is essential to prevent penetration, floating, and contact drift.

To address these challenges, we propose **PACHO**, a **P**arallel **A**ffordance and **C**ontact diffusion framework for text-driven 3D **H**uman-**O**bject interaction synthesis. PACHO first employs Dual Motion-Prior Learners to model human and object motions through two parallel denoising branches conditioned on a shared global context. This design preserves clean entity-specific kinematic priors while maintaining a common semantic grounding from the text and object geometry. To bridge the two branches, we introduce a Cross-Affordance Interaction module that performs adaptive, part-aware, point-level reasoning between human contact salience and object affordance salience. The predicted contact-affordance fields are fed back through reciprocal residual updates, with object affordance refining the human stream and human contact refining the object stream, thereby serving as an explicit interaction interface between the decoupled motion priors. Finally, PACHO adopts an Affordance-Guided Motion Diffusion objective that jointly supervises human-object motion, human contact maps, and object affordance maps during denoising. A salience-weighted contact alignment loss further encourages locally consistent contact while discouraging penetration and contact drift.

Our contributions are summarized as follows:

- **Dual-sided dynamic interaction representation.** We represent HOI using time-varying human contact and object affordance defined over human and object surfaces, providing a spatial interaction representation beyond single-sided, sparse, or part-level cues.

- **Contact-affordance-mediated reciprocal coupling.** We introduce Cross-Affordance Interaction, which uses object affordance to update the human motion stream and human contact to update the object motion stream, making the predicted interaction fields an explicit communication interface between decoupled motion priors.

- **Affordance-guided motion diffusion.** We jointly predict motion, human contact, and object affordance throughout denoising and impose salience-weighted geometric alignment, enabling the interaction representation and motion to evolve consistently over time.

2 Related Work

Text-driven human motion generation. Text-driven human motion generation has evolved from discrete-token and motion-language modeling to diffusion-based continuous synthesis Tevet et al. (2023); Hong et al. (2025); Meng et al. (2025). Token-based methods such as MotionGPT Jiang et al. (2023a) and MoMask Guo et al. (2024) learn compact motion vocabularies and exploit autoregressive language modeling for controllable motion synthesis. Diffusion-based methods, including MotionDiffuse Zhang et al. (2024) and MLD Chen et al. (2023), improve diversity and realism by modeling motion in latent spaces. More recent works, such as AToM Han et al. (2025), MoTiGA Chen et al. (2026), MARDM Meng et al. (2025), and MotionMaster Jiang et al. (2026), further enhance semantic alignment and whole-body fidelity. However, these methods primarily address human-only motion generation in free space and do not explicitly model dynamic human-object interactions. Directly applying these architectures to interaction tasks may lead to premature entanglement and physically inconsistent human-object motions, which motivates our decoupled dual motion-prior modeling.

Human-scene and human-object interaction generation. Context-aware motion generation extends text-to-motion from free-space human motion to 3D environments Wang et al. (2024a); Yu et al. (2025); Xu et al. (2024); Li & Dai (2026); Ghosh et al. (2026). Early human-scene methods Xiao et al. (2024); Cen et al. (2024); Wang et al. (2022) improve spatial grounding through object-centric representations. However, they mainly focus on static objects and are less suited for dynamic human-object interactions. Recent datasets Bhatnagar et al. (2022); Jiang et al. (2023b); Li et al. (2023); Xu et al. (2025a); Lu et al. (2025) enable full-body dynamic HOI modeling beyond dexterous hand-object manipulation Taheri et al. (2020); Fan et al. (2023). Recent text-driven HOI methods explore relation intervention, perceptive diffusion, contact guidance, and fine-grained semantic alignment Wu et al. (2025b); Song et al. (2024); Diller & Dai (2024); Peng et al. (2025); Yang et al. (2024). LIGHT Wang et al. (2026b) improves motion quality through pace-induced guidance based on asynchronous denoising. Recent works further explore physics-based control Xu et al. (2025b; 2026b), cooperative multi-agent interaction Lionar & Lee (2026), kinematic chain modeling Zeng et al. (2025), long-horizon interaction synthesis from high-level instructions Wu et al. (2025d), and open-set interaction synthesis Xu et al. (2024); Zhang et al. (2025a); Jia et al. (2025). Nevertheless, most existing HOI methods either entangle human and object features in a latent space or rely on sparse contact cues or high-level guidance, limiting their ability to model dynamic, bidirectional, fine-grained contact-affordance coupling. In contrast, PACHO learns decoupled kinematic priors and bridges them through adaptive part-aware point-level cross-affordance reasoning.

Affordance modeling for interaction synthesis. Affordance provides an intermediate representation that links geometry to action and has been widely studied in 3D vision. Recent methods infer object affordances from images, video, or language instructions Zhu et al. (2025); Shao et al. (2025); Chu et al. (2025); Wang et al. (2026a); Zhao et al. (2026). DAViD Kim et al. (2025) models dynamic affordance using video diffusion priors, while OpenHOI Zhang et al. (2025b) explores open-world affordance grounding for language-guided interaction synthesis. In motion generation, affordance-related cues Wang et al. (2024b); Diller & Dai (2024); Wu et al. (2026) have been used for static scene grounding, contact-based guidance, or interaction synthesis. CG-HOI Diller & Dai (2024) models dynamic body-anchored contact distances to coordinate human and object motion, but lacks explicit object affordance modeling to align human contact with task-relevant functional regions. HOI-Diff Peng et al. (2025) models affordance as sparse point-level contact cues for motion guidance, but is sensitive to inaccurate point prediction. HOI-PAGE Li & Dai (2026) further introduces part-level affordance guidance for novel objects, but remains coarse within each object part and depends on segmentation quality. Despite these advances, existing methods typically rely on sparse points, discrete object parts, or indirect dynamic priors. In contrast, PACHO dynamically co-predicts dense human contact and object affordance maps and performs part-aware point-level bidirectional interaction modeling, enabling finer geometric consistency and more stable human-object contact.

3 Method

PACHO explicitly models dynamic human contact and object affordance salience, by learning decoupled human and object motion priors, coupling them through part-aware contact-affordance reasoning, and enforcing local geometric consistency during denoising. The overall pipeline is illustrated in Figure 2. Sec. 3.2 introduces the Dual Motion-Prior Learners for entity-specific

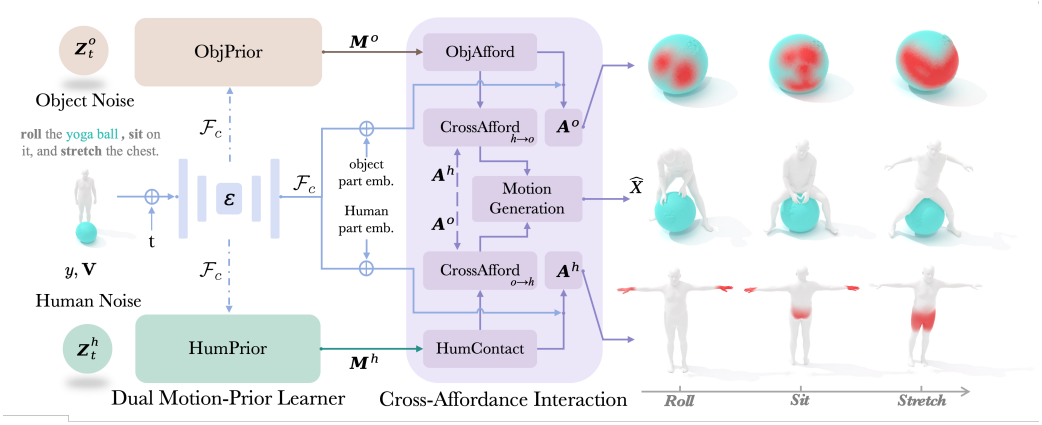


Figure 2: **Pipeline of PACHO.** PACHO comprises two core modules and an Affordance-Guided Motion Diffusion objective. (1) **Dual Motion-Prior Learners** extract decoupled human and object motion priors, $\mathbf{M}^h$ and $\mathbf{M}^o$, from a shared condition $\mathcal{F}_c$ encoded from the text prompt y, object geometry $\mathbf{V}$, and diffusion timestep t. (2) **Cross-Affordance Interaction** predicts part-aware human contact salience $\mathbf{A}^h$ and object affordance salience $\mathbf{A}^o$, and uses them to establish bidirectional coupling between the two motion streams. (3) **Affordance-Guided Motion Diffusion** denoises human–object motion using jointly predicted human contact and object affordance salience, providing fine-grained geometric guidance for physically plausible and contact-consistent HOI synthesis.

kinematic modeling; Sec. 3.3 presents the Cross-Affordance Interaction module for residual contact-aware coupling; and Sec. 3.4 describes the Affordance-Guided Motion Diffusion objective for contact-consistent generation.

3.1 Problem Setup

We denote a human-object interaction sequence of length L as $\mathbf{X} = \{\mathbf{X}^h, \mathbf{X}^o\}$. The human motion is represented as $\mathbf{X}^h \in \mathbb{R}^{L \times D_h}$, where D_h denotes the dimension of the human state. At frame l, the human state $\mathbf{X}^h_l = (\mathbf{t}_l, \mathbf{r}_l)$ consists of the global joints positions $\mathbf{t}_l \in \mathbb{R}^{24\times 3}$ and global joints orientation $\mathbf{r}_l \in \mathbb{R}^{22\times 6}$ in the 6D rotation representation, allowing full-body mesh reconstruction via the SMPL-H Romero et al. (2017). The object motion is represented as $\mathbf{X}^o \in \mathbb{R}^{L\times D_o}$, where D_o denotes the dimension of the object state. At frame l, the object state $\mathbf{X}^o_l = (\mathbf{c}_l, \mathbf{q}_l)$ consists of the global centroid $\mathbf{c}_l \in \mathbb{R}^3$ and the relative 6D rotation $\mathbf{q}_l \in \mathbb{R}^6$ with respect to the initial pose. The initial spatial geometry is represented by the human and object mesh vertices $\mathbf{V}^h$ and $\mathbf{V}^o$. Given the text instruction y and the initial geometry $\mathbf{V} = [\mathbf{V}^h, \mathbf{V}^o]$, our goal is to model the conditional distribution $P_\theta(\mathbf{X} \mid y, \mathbf{V})$ and generate a plausible interaction sequence $\hat{\mathbf{X}}$.

3.2 Dual Motion-Prior Learners

Human and object motions exhibit inherently different kinematic properties Xu et al. (2025b); Peng et al. (2025); Yu et al. (2026). Before introducing contact and affordance-aware interaction, we first learn decoupled motion priors for human and object separately. Specifically, our Dual Motion-Prior Learners adopt a shared context condition encoder $\mathcal{E}$ with a dual-branch Diffusion Transformer, $\mathrm{HumPrior}(\cdot)$ and $\mathrm{ObjPrior}(\cdot)$, to learn human and object motion distributions, respectively. The context encoder $\mathcal{E}$ incorporates semantic, geometric, and diffusion-step cues, while the two diffusion branches model entity-specific motion dynamics within a shared conditional space.

The shared context encoder $\mathcal{E}$ takes the text instruction y, diffusion timestep t, and human and object geometries $\mathbf{V} = [\mathbf{V}^h, \mathbf{V}^o]$ as input. $\mathbf{V}^h$ and $\mathbf{V}^o$ are equal-sized point clouds obtained via farthest-point sampling Qi et al. (2017). The encoder $\mathcal{E}$ produces a compact global condition $\mathcal{F}_c$:

$$\mathcal{F}_c = \mathcal{E}(y, \mathbf{V}, t), \quad \mathbf{M}^h = \mathrm{HumPrior}(\mathbf{Z}^h_t \mid \mathcal{F}_c), \quad \mathbf{M}^o = \mathrm{ObjPrior}(\mathbf{Z}^o_t \mid \mathcal{F}_c). \tag{1}$$

Here, $\mathbf{Z}^h_t$ and $\mathbf{Z}^o_t$ are the noisy human and object motion tokens in latent space. The resulting priors $\mathbf{M} = (\mathbf{M}^h, \mathbf{M}^o)$ encode decoupled motion features: $\mathbf{M}^h$ captures plausible human kinematic

patterns, while $\mathbf{M}^o$ captures plausible object motion dynamics. Since these priors are learned before explicit interaction modeling, they provide entity-specific motion representations.

3.3 Cross-Affordance Interaction

Given the decoupled human and object motion priors $\mathbf{M}^h$ and $\mathbf{M}^o$, we introduce Cross-Affordance Interaction. Rather than directly coupling the two motion streams, we first instantiate two explicit interaction fields: human contact salience $\mathbf{A}^h$ and object affordance salience $\mathbf{A}^o$. These fields serve as the "handshake" between the streams. The human branch identifies contact-relevant body regions, while the object branch identifies affordance-relevant functional regions.

We first use SMPL-H body part labels Romero et al. (2017) to partition the human body into semantic regions, and adopt P3-SAM Ma et al. (2025) to decompose the object into part-level regions. Let $\mathbf{L}^h \in \mathbb{N}^{n_p'}$ and $\mathbf{L}^o \in \mathbb{N}^{n_p'}$ denote the part labels associated with the sampled human and object vertices, respectively. We then encode these part labels together with the sampled geometry to obtain part-aware human and object embeddings:

$$\mathbf{P}^h = \mathrm{PartEmb}^h(\mathbf{V}^h, \mathbf{L}^h), \qquad \mathbf{P}^o = \mathrm{PartEmb}^o(\mathbf{V}^o, \mathbf{L}^o), \tag{2}$$

Given the decoupled motion priors $\mathbf{M}^h$ and $\mathbf{M}^o$, the shared global condition $\mathcal{F}_c$, and the part embeddings $\mathbf{P}^h$ and $\mathbf{P}^o$, we predict dynamic contact-affordance salience over the entire interaction sequence. To this end, the motion-prior features, global condition, and part embeddings are first projected into a shared latent space and expanded to the point-wise temporal dimension. The human contact predictor and object affordance predictor then produce part-aware point-level salience maps:

$$\mathbf{A}^h = \mathrm{HumContact}\big([\mathbf{M}^h, \mathcal{F}_c, \mathbf{P}^h]\big), \qquad \mathbf{A}^o = \mathrm{ObjAfford}\big([\mathbf{M}^o, \mathcal{F}_c, \mathbf{P}^o]\big), \tag{3}$$

where $[\cdot]$ denotes feature concatenation, and $\mathbf{A}^h, \mathbf{A}^o \in [0,1]^{L \times n_p'}$ denote the predicted human contact salience and object affordance salience maps, respectively. At frame l, $\mathbf{A}^h_l(v)$ measures the contact salience of a sampled human vertex $v \in \mathbf{V}^h$, while $\mathbf{A}^o_l(u)$ measures the affordance salience of a sampled object vertex $u \in \mathbf{V}^o$.

The predicted salience fields are then used to establish bidirectional contact-aware coupling between the decoupled motion streams. Object affordance salience provides functional-region cues for refining the human motion prior, while human contact salience provides reciprocal body-region cues for refining the object motion prior under interaction constraints. Specifically, we encode the predicted object affordance salience and human contact salience into interaction tokens using lightweight point encoders. CrossAfford then uses these interaction tokens as contact-affordance keys and values, with the corresponding motion-prior tokens as queries:

$$\mathbf{M}^h \leftarrow \mathbf{M}^h + \mathrm{CrossAfford}_{o \to h}\big(\mathbf{M}^h, \mathbf{A}^o\big), \qquad \mathbf{M}^o \leftarrow \mathbf{M}^o + \mathrm{CrossAfford}_{h \to o}\big(\mathbf{M}^o, \mathbf{A}^h\big). \tag{4}$$

After cross-affordance coupling, $\mathbf{M}^h$ and $\mathbf{M}^o$ encode interaction-aware coupled features and are decoded into clean-motion estimates at diffusion timestep t. The human decoder predicts $\tilde{\mathbf{X}}^{h,(t)}_{1:L} = (\tilde{\mathbf{T}}^{(t)}_{1:L}, \tilde{\mathbf{R}}^{(t)}_{1:L})$, including global joints translation and global joints orientation. The object decoder predicts $\tilde{\mathbf{X}}^{o,(t)}_{1:L} = (\tilde{\mathbf{C}}^{(t)}_{1:L}, \tilde{\mathbf{Q}}^{(t)}_{1:L})$, including the object centroid and rotation.

3.4 Affordance-Guided Motion Diffusion

In the final stage, PACHO jointly predicts human motion, object motion, the human contact map, and the object affordance map. Let the clean human-object motion sequence be $\mathbf{X}_{1:L} = \{\mathbf{X}^h_{1:L}, \mathbf{X}^o_{1:L}\}$. At diffusion step t, we denote the noisy motion sequence as $\mathbf{Z}_t = \{\mathbf{Z}^h_t, \mathbf{Z}^o_t\}$. Our denoising network ϵ_θ combines the Dual Motion-Prior Learners with the Cross-Affordance Interaction module to perform conditional denoising. Following Zhang et al. (2024); Tevet et al. (2023); Tseng et al. (2023); Yuan et al. (2023), we use the $\mathbf{x}_0$-prediction parameterization: instead of predicting the injected noise, the network directly estimates the clean human-object motion sequence at each reverse diffusion step:

$$\big\{\tilde{\mathbf{X}}^{(t)}_{1:L}, \mathbf{A}^{h,(t)}, \mathbf{A}^{o,(t)}\big\} = \epsilon_\theta\big(\mathbf{Z}_t, t, y, \mathbf{V}\big), \tag{5}$$

where $\tilde{\mathbf{X}}^{(t)}_{1:L}$ denotes the predicted $\mathbf{x}_0$ estimate at diffusion step t, and $\mathbf{A}^{h,(t)}$ and $\mathbf{A}^{o,(t)}$ denote the corresponding human contact and object affordance salience predictions, respectively. The final generated motion sequence after iterative reverse denoising is denoted as $\hat{\mathbf{X}}_{1:L}$.

To encourage geometric consistency around contact-affordance regions, we introduce a salience-weighted contact alignment loss. At diffusion step t, the loss is imposed on the predicted clean motion estimate, encouraging high-salience human contact vertices to remain close to the object surface and high-salience object affordance vertices to align with the human surface:

$$\mathcal{L}_{\text{sca}}^{(t)} = \frac{1}{L}\sum_{l=1}^{L}\left[\frac{\sum_i A_{l,i}^{h,*}\, d\Big(\tilde{v}_{l,i}^{h,(t)}, \tilde{\mathcal{S}}_l^{o,(t)}\Big)}{\sum_i A_{l,i}^{h,*} + \epsilon} + \frac{\sum_j A_{l,j}^{o,*}\, d\Big(\tilde{u}_{l,j}^{o,(t)}, \tilde{\mathcal{S}}_l^{h,(t)}\Big)}{\sum_j A_{l,j}^{o,*} + \epsilon}\right]. \tag{6}$$

where $A_l^{h,*}$ and $A_l^{o,*}$ denote the pseudo-label targets for human contact and object affordance salience, respectively. The vertices $\tilde{v}_{l,i}^{h,(t)}$ and $\tilde{u}_{l,j}^{o,(t)}$, together with the reconstructed human and object surfaces $\tilde{\mathcal{S}}_l^{h,(t)}$ and $\tilde{\mathcal{S}}_l^{o,(t)}$, are derived from the predicted clean motion estimate at diffusion step t. $d(\cdot, \mathcal{S})$ denotes the point-to-surface distance. The loss is weighted by a noise-dependent coefficient $\lambda_{\text{sca}}(t) = \lambda_{\text{sca}}^{\text{base}}\bar{\alpha}_t^{\gamma}$, where $\lambda_{\text{sca}}^{\text{base}}$ is the base weight, γ controls the decay rate, and $\epsilon > 0$ avoids division by zero. This design reduces the influence of geometric supervision when the estimate is still highly noisy, while enforcing stronger contact alignment as the reverse process approaches cleaner motion states. The full training objective is provided in Appendix B.

3.5 Implementation Details

We employ an 8-layer Transformer architecture for PACHO, divided into 4 prior-isolation blocks for the Dual Motion-Prior Learners and 4 cross-interaction blocks for the Cross-Affordance Interaction. The number of downsampled points $n'p$ is set to 256. All embedding dimensions are set to 512. A 6-layer point-wise MLP and CLIP ViT-B/32 are used as the shared context encoder $\mathcal{E}$. PACHO is trained using the AdamW optimizer with a learning rate of 2×10^{-4}. The diffusion process uses $T = 1000$ timesteps and directly predicts the clean sample $\mathbf{x}0$. The salience-weighted contact alignment loss $\mathcal{L}\text{sca}^{(t)}$ is configured with base weight $\lambda_{\text{sca}}^{\text{base}} = 0.1$, decay exponent $\gamma = 2.0$, and numerical constant $\epsilon = 10^{-6}$. Inference is performed using a 1000-step sampler. PACHO is implemented in PyTorch and takes approximately 20 hours to converge on a single NVIDIA RTX 4090 GPU.

4 Experiment

4.1 Experimental Setup

Dataset. We evaluate PACHO on OMOMO Li et al. (2023) and InterAct Xu et al. (2025a) benchmarks. OMOMO provides 10 hours of paired human and object motion sequences involving 15 everyday objects, with human motions originally represented by SMPL-X Pavlakos et al. (2019) full-body parameters. InterAct is a large-scale 3D human-object interaction dataset with rich text annotations, SMPL-H Romero et al. (2017) human motion representations, and dynamic object states. It covers diverse everyday object categories and complex multi-stage behaviors. For a fair comparison, we follow the standard LIGHT Wang et al. (2026b) split protocol. Since neither OMOMO nor InterAct provides dense contact-affordance annotations, we construct contact-affordance pseudo-labels from ground-truth meshes, with details provided in Appendix E.

Evaluation Metrics. We evaluate PACHO from three aspects: (1) *Kinematic Quality*, measured by Fréchet Inception Distance (**FID**), which compares generated and ground-truth human–object interaction distributions in a learned feature space; (2) *Semantic Alignment*, measured by **R-Precision**, which evaluates text–interaction retrieval accuracy. Both metrics use a pretrained full-HOI evaluator following LIGHT Wang et al. (2026b); and (3) *Physical Plausibility*, evaluated using Optimal Contact Distance (**OCD**), Penetration Score (**Pene**), and Foot Skating Ratio (**FSR**). Following ChainHOI Zeng et al. (2025), OCD reports the minimum contact distance across object-matched, semantically consistent reference contact annotations. Pene quantifies human–object interpenetration, while FSR measures the percentage of ground-contact frames exhibiting foot sliding.

Baselines. We compare PACHO with representative baselines, including MDM Tevet et al. (2023), InterDiff Xu et al. (2023), CHOIS Li et al. (2024), HOI-Diff Peng et al. (2025), Text2HOI Cha et al. (2024), ChainHOI Zeng et al. (2025), HOI-Dyn Wu et al. (2025a) and LIGHT Wang et al. (2026b). MDM serves as a foundational text-to-motion diffusion model, which we fine-tune on

Table 1: **Comparison of PACHO with baselines on the InterAct Xu et al. (2025a) and OMOMO Li et al. (2023) datasets.** ⋆ denotes fine-tuning; the best and second-best results are shown in **bold** and underlined, respectively.

Method	FID ↓	R-Precision ↑			OCD ↓	Pene ↓	FSR ↓
		Top1	Top2	Top3			
Evaluated on the InterAct Dataset							
HOI-Diff Peng et al. (2025)	$0.689^{\pm 0.031}$	$0.413^{\pm 0.010}$	$0.624^{\pm 0.009}$	$0.740^{\pm 0.011}$	$0.412^{\pm 0.008}$	$\underline{0.103}^{\pm 0.010}$	$0.072^{\pm 0.160}$
CHOIS Li et al. (2024)	$0.572^{\pm 0.008}$	$0.439^{\pm 0.005}$	$\underline{0.660}^{\pm 0.002}$	$0.766^{\pm 0.003}$	$0.315^{\pm 0.006}$	$0.131^{\pm 0.013}$	$0.119^{\pm 0.234}$
InterDiff Xu et al. (2023)	$0.215^{\pm 0.001}$	$\mathbf{0.501}^{\pm 0.009}$	$\mathbf{0.722}^{\pm 0.003}$	$\mathbf{0.824}^{\pm 0.007}$	$0.248^{\pm 0.005}$	$0.116^{\pm 0.011}$	$0.092^{\pm 0.178}$
Text2HOI Cha et al. (2024)	$0.331^{\pm 0.003}$	$0.428^{\pm 0.003}$	$0.637^{\pm 0.003}$	$0.745^{\pm 0.000}$	$0.284^{\pm 0.004}$	$0.105^{\pm 0.003}$	$\mathbf{0.055}^{\pm 0.123}$
LIGHT Wang et al. (2026b)	$\underline{0.148}^{\pm 0.014}$	$0.421^{\pm 0.014}$	$0.637^{\pm 0.016}$	$0.754^{\pm 0.016}$	$\underline{0.182}^{\pm 0.006}$	$0.132^{\pm 0.001}$	$0.078^{\pm 0.197}$
Our PACHO	$\mathbf{0.085}^{\pm 0.011}$	$\underline{0.455}^{\pm 0.013}$	$0.658^{\pm 0.014}$	$\underline{0.768}^{\pm 0.012}$	$\mathbf{0.065}^{\pm 0.011}$	$\mathbf{0.062}^{\pm 0.012}$	$\underline{0.058}^{\pm 0.015}$
Evaluated on the OMOMO Dataset							
MDM⋆ Tevet et al. (2023)	$0.209^{\pm 0.004}$	$0.135^{\pm 0.003}$	$0.208^{\pm 0.008}$	$0.303^{\pm 0.010}$	$0.560^{\pm 0.003}$	$0.026^{\pm 0.007}$	$0.134^{\pm 0.001}$
InterDiff Xu et al. (2023)	$0.234^{\pm 0.008}$	$0.129^{\pm 0.011}$	$0.241^{\pm 0.010}$	$0.324^{\pm 0.005}$	$0.472^{\pm 0.002}$	$\underline{0.024}^{\pm 0.002}$	$0.139^{\pm 0.001}$
ChainHOI Zeng et al. (2025)	$0.122^{\pm 0.003}$	$0.254^{\pm 0.006}$	$0.472^{\pm 0.009}$	$0.561^{\pm 0.006}$	$0.263^{\pm 0.002}$	$0.023^{\pm 0.001}$	$0.089^{\pm 0.009}$
HOI-Dyn Wu et al. (2025a)	$\underline{0.105}^{\pm 0.005}$	$\underline{0.291}^{\pm 0.006}$	$\underline{0.524}^{\pm 0.008}$	$\underline{0.686}^{\pm 0.007}$	$\underline{0.178}^{\pm 0.003}$	$0.028^{\pm 0.002}$	$\mathbf{0.011}^{\pm 0.002}$
LIGHT Wang et al. (2026b)	$0.099^{\pm 0.018}$	$\underline{0.302}^{\pm 0.004}$	$\underline{0.538}^{\pm 0.015}$	$\underline{0.707}^{\pm 0.011}$	$\underline{0.165}^{\pm 0.004}$	$0.100^{\pm 0.001}$	$\underline{0.013}^{\pm 0.045}$
Our PACHO	$\mathbf{0.082}^{\pm 0.003}$	$\mathbf{0.325}^{\pm 0.004}$	$\mathbf{0.560}^{\pm 0.005}$	$\mathbf{0.725}^{\pm 0.004}$	$\mathbf{0.075}^{\pm 0.001}$	$\mathbf{0.012}^{\pm 0.001}$	$0.020^{\pm 0.001}$

interaction datasets for comparison. InterDiff synthesizes human-object interactions by modeling mutual dependencies between human and object motions. CHOIS performs controllable HOI synthesis conditioned on object trajectory. HOI-Diff adopts a dual-branch diffusion architecture for text-driven 3D interaction sequences. Text2HOI is included as a text-guided human-object interaction baseline. ChainHOI models interaction generation through kinematic chain reasoning. HOI-Dyn formulates HOI generation as a driver-responder system and explicitly models object responses to human motion. LIGHT improves HOI animation quality using classifier-free guidance.

4.2 Comparison with State-of-the-Art

Table 1 compares PACHO with baselines in terms of kinematic quality, semantic alignment, and physical plausibility. Overall, PACHO achieves the best FID, OCD, and Pene on InterAct, and obtains the best performance across all metrics except FSR on OMOMO, demonstrating strong distributional realism and contact consistency. Although InterDiff achieves the strongest R-Precision on InterAct, PACHO remains competitive in semantic alignment and further achieves the best R-Precision across all Top-k metrics on OMOMO. More importantly, PACHO consistently improves contact-related metrics, achieving the best OCD and Pene on both datasets, which indicates its effectiveness in reducing contact distance and human-object interpenetration. While PACHO does not obtain the lowest FSR, it remains competitive on both benchmarks. These results show that Dual Motion-Prior Learners and Cross-Affordance Interaction provide effective spatial guidance for generating realistic, semantically aligned, and more physically plausible HOI sequences.

4.3 Visual Comparisons

We provide qualitative evaluations of PACHO against LIGHT Wang et al. (2026b) across complex, multi-stage interaction scenarios in Figure 3, demonstrating improvements in three key areas: (1) **Kinematic realism.** PACHO generates fluid, affordance-adaptive motions, smoothly executing necessary postural adjustments like bending or lowering. Conversely, LIGHT frequently exhibits unnatural joint rigidity and lacks sufficient preparatory articulation (highlighted in yellow). (2) **Geometric consistency.** Guided by part-aware point-level affordances, PACHO significantly mitigates unphysical mesh intersections. For instance, during the wooden chair interaction, PACHO yields precise hand-object alignment, whereas LIGHT suffers from severe interpenetration (red boxes). (3) **Contact stability.** PACHO ensures sustained, stable contact throughout sequential actions such as lifting, rotating, and placing. In contrast, LIGHT is prone to contact breakage (green boxes), resulting in floating artifacts where objects lack proper physical support. Overall, these visualizations validate that our decoupled motion priors and adaptive cross-affordance coupling effectively suppress geometric artifacts and enhance the physical plausibility of synthesized interactions.

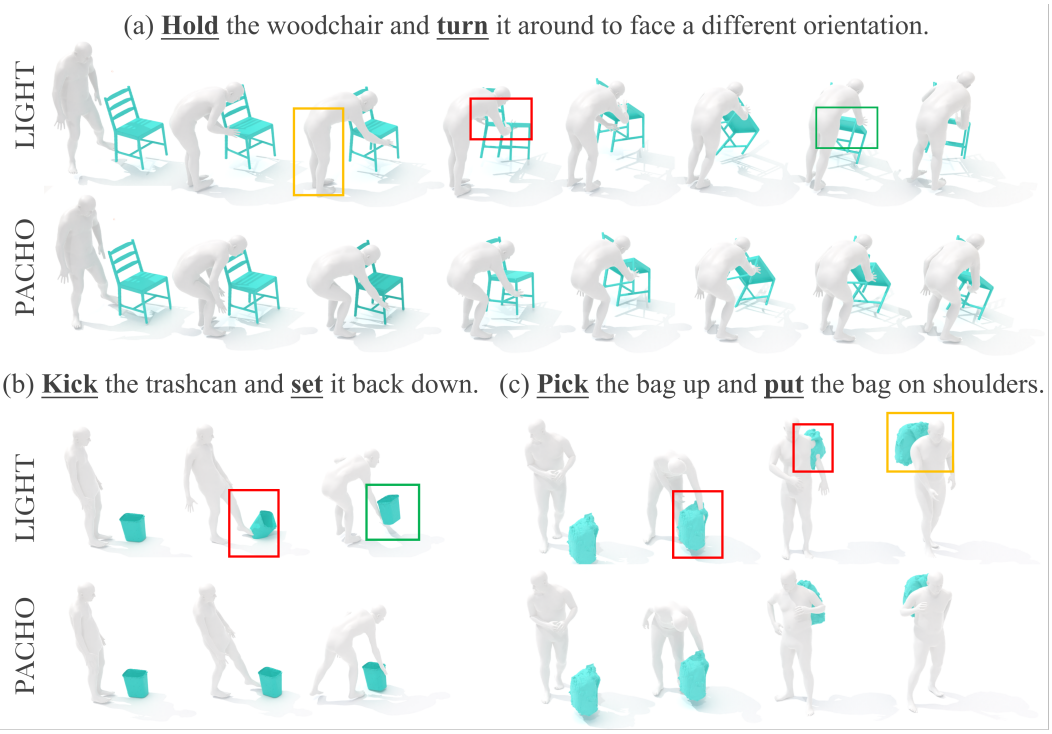


Figure 3: **Visual comparisons.** We compare PACHO with the state-of-the-art baseline LIGHT Wang et al. (2026b). The yellow boxes indicate unnatural joint rigidity, the red boxes highlight severe interpenetration, and the green boxes denote contact breakage during the interaction.

Table 2: **Ablation studies on the InterAct dataset.** Subtable (a) evaluates the contribution of key components. Subtable (b) compares different coupling strategies between the human and object.

Components	FID ↓	R-Top3 ↑	OCD ↓	Pene ↓
w/o Dual Motion-Prior.	0.125	0.720	0.085	0.095
w/o Cross-Affordance	0.090	0.735	0.145	0.125
w/o Contact-Affordance Co-Pred.	0.101	0.712	0.128	0.118
Full PACHO	**0.085**	**0.768**	**0.065**	**0.062**

(a) **Core components.** Dual Motion-Prior Learners improve kinematic realism, while Cross-Affordance Interaction and Contact-Affordance Co-Prediction are crucial for contact-aware coupling and physical plausibility.

Coupling Strategy	FID ↓	R-Top3 ↑	OCD ↓	Pene ↓
No Cross Coupling	0.090	0.735	0.145	0.125
Object → Human	0.094	0.725	0.088	0.086
Human → Object	0.092	0.735	0.085	0.083
Human ↔ Object	**0.085**	**0.768**	**0.065**	**0.062**

(b) **Cross-affordance coupling.** Bidirectional coupling between human contact salience and object affordance salience yields more stable contact than one-way coupling.

4.4 Ablation Study

To validate the effectiveness of the core designs in PACHO, we conduct comprehensive ablation studies, summarized in Table 2. Additional ablation studies are provided in Appendix C.

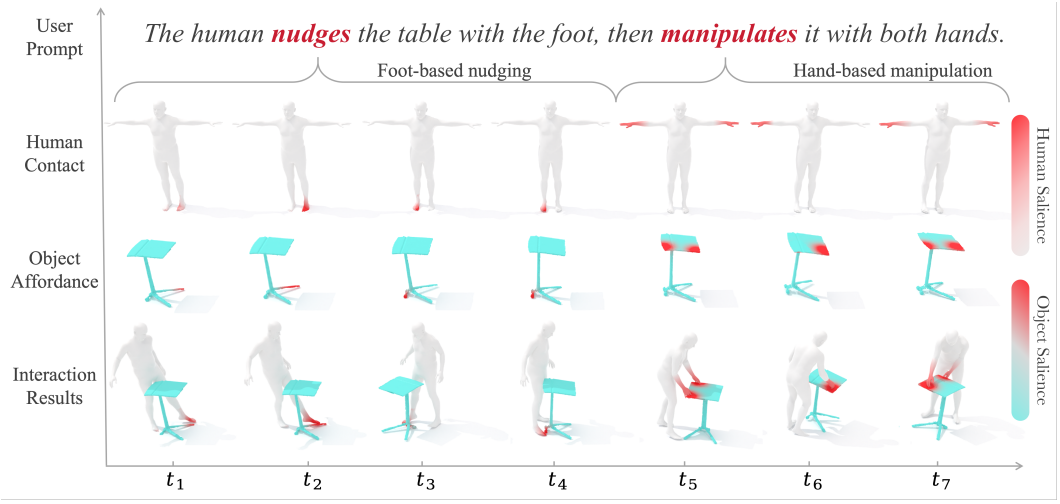
Core Components. Table 2(a) evaluates the contribution of the main components in PACHO. Removing the Dual Motion-Prior Learners degrades FID and penetration-related metrics, indicating that decoupled human and object priors help preserve entity-specific kinematic patterns before interaction modeling. Removing Cross-Affordance Interaction substantially increases OCD and Pene, showing the importance of explicit contact-aware coupling. Removing contact-affordance co-prediction further degrades semantic alignment and physical plausibility, confirming that parallel prediction of human contact salience and object affordance salience provides useful spatial guidance. These results show that the main components of PACHO are complementary.

Cross-Affordance Coupling. Table 2(b) analyzes the effect of different coupling directions. Without cross coupling, PACHO produces weaker contact consistency, as reflected by much higher OCD and Pene. Using only object → human or human → object coupling improves physical plausibility compared with no coupling, but remains inferior to bidirectional coupling. This indicates that human

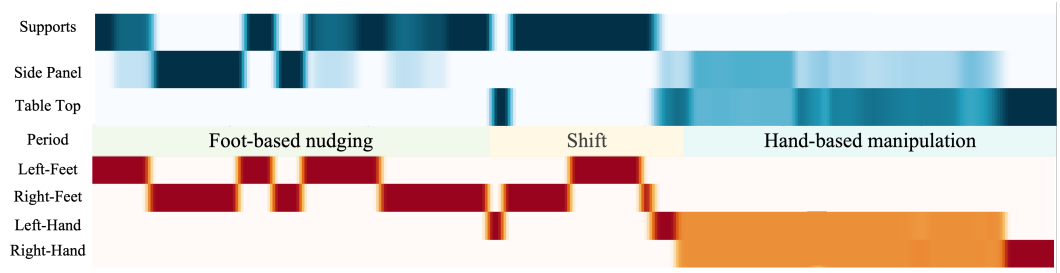
contact salience and object affordance salience provide complementary interaction cues: object-side salience helps guide human motion toward plausible contact regions, while human-side structural cues help refine object motion under interaction constraints. Human ↔ Object Cross-Affordance Interaction therefore yields the most stable and more physically plausible human-object motions.

4.5 Dynamic Contact-Affordance Study

Figure 4 illustrates the spatial and temporal dynamic contact-affordance evolution predicted by PACHO for a multi-stage interaction. The predicted salience follows the semantic progression of the instruction, shifting from foot-table-base contact during foot-based nudging to hand-tabletop contact during hand-based manipulation, as shown in Figure 4a. Figure 4b further confirms this coordinated evolution across human and object parts, indicating that PACHO captures time-varying, stage-dependent contact-affordance cues throughout the full interaction sequence.



(a) Spatial dynamic contact-affordance evolution.



(b) Temporal dynamic contact-affordance evolution.

Figure 4: **Dynamic contact-affordance evolution.** (a) shows time-varying human contact and object affordance salience for a multi-stage instruction. (b) shows the part-wise temporal salience evolution over object and human parts, with interaction stages annotated in the middle.

5 Discussion and Conclusion

In this paper, we present PACHO, a contact-affordance-guided framework for text-driven 3D human-object interaction synthesis. By modeling the dynamic "handshake" between human contact and object affordances, PACHO provides fine-grained spatial guidance. PACHO learns decoupled human and object motion priors, couples them through contact-affordance reasoning, and denoises human–object motion using predicted contact and affordance salience. Despite the effectiveness of PACHO, several limitations remain. PACHO may struggle with rapid contact-affordance transitions, where temporally unstable salience predictions can cause abrupt pose changes or contact drift, and it does not explicitly model full physical dynamics. In addition, PACHO mainly focuses on single human-object interactions, leaving multi-object and multi-agent scenarios for future exploration.

AI USE STATEMENT

We used Qwen3 for offline text-to-prior parsing to support pseudo-label construction. The parsed priors were further constrained by geometric filtering, and the reliability of the parser was assessed against manual annotations. We also used generative AI tools for language polishing and to improve the clarity and readability of the manuscript. We did not use generative AI tools for research ideation, methodology design, or result analysis. All AI-assisted content was carefully reviewed and revised by the authors, who take full responsibility for the final manuscript.

REPRODUCIBILITY STATEMENT

We provide architecture and training details in Sec. 3.5 and Appendices B and D, evaluation protocols in Appendix A, pseudo-label construction and validation in Appendix E, and additional ablations and generalization analyses in Appendices C, F, G, and H. We will release the generated contact-affordance labels and associated processing scripts subject to the licenses of the original datasets.

REFERENCES

Bharat Lal Bhatnagar, Xianghui Xie, Ilya A. Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. BEHAVE: Dataset and method for tracking human object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15935–15946, June 2022.

Songjin Cai, Linjie Zhong, Ling Guo, and Changxing Ding. ViHOI: Human-object interaction synthesis with visual priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 30686–30695, June 2026.

Zhi Cen, Huaijin Pi, Sida Peng, Zehong Shen, Minghui Yang, Shuai Zhu, Hujun Bao, and Xiaowei Zhou. Generating human motion in 3D scenes from text descriptions. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1855–1866. IEEE, 2024.

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

Xiaodong Chen, Qian Bao, Xudong Liu, Jianping Fang, Jintao Fang, Yongdong Zhang, Tao Mei, and Wu Liu. Multi-level causal LLM-based text-to-motion generation with human alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9342–9351, 2026.

Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. Executing your commands via motion diffusion in latent space. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18000–18010. IEEE, 2023.

Hengshuo Chu, Xiang Deng, Qi Lv, Xiaoyang Chen, Yinchuan Li, Jianye HAO, and Liqiang Nie. 3D-affordanceLLM: Harnessing large language models for open-vocabulary affordance detection in 3D worlds. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=GThTiuXgDC.

Sisi Dai, Wenhao Li, Haowen Sun, Haibin Huang, Chongyang Ma, Hui Huang, Kai Xu, and Ruizhen Hu. Interfusion: Text-driven generation of 3D human-object interaction. In *European Conference on Computer Vision*, pp. 18–35. Springer, 2024.

Christian Diller and Angela Dai. CG-HOI: Contact-guided 3D human-object interaction generation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19888–19901. IEEE, 2024.

Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J Black, and Otmar Hilliges. ARCTIC: A dataset for dexterous bimanual hand-object manipulation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12943–12954. IEEE, 2023.

Anindita Ghosh, Vladislav Golyanik, Taku Komura, Philipp Slusallek, Christian Theobalt, and Rishabh Dabral. SceMoS: Scene-aware 3D human motion synthesis by planning with geometry-grounded tokens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16443–16453, June 2026.

Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, and Li Cheng. Momask: Generative masked modeling of 3D human motions. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1900–1910. IEEE, 2024.

Haonan Han, Xiangzuo Wu, Huan Liao, Zunnan Xu, Zhongyuan Hu, Ronghui Li, Yachao Zhang, and Xiu Li. Atom: Aligning text-to-motion model at event-level with gpt-4vision reward. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22746–22755. IEEE, 2025.

Wenkun He, Yun Liu, Ruitao Liu, and Li Yi. Syncdiff: Synchronized motion diffusion for multi-body human-object interaction synthesis. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11731–11743. IEEE, 2025.

Seokhyeon Hong, Chaelin Kim, Serin Yoon, Junghyun Nam, Sihun Cha, and Junyong Noh. Salad: Skeleton-aware latent diffusion for text-driven motion generation and editing. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7158–7168. IEEE, 2025.

Kai Jia, Tengyu Liu, Mingtao Pei, Yixin Zhu, and Siyuan Huang. PrimHOI: Compositional human-object interaction via reusable primitives. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11491–11501, 2025.

Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. *Advances in Neural Information Processing Systems*, 36:20067–20079, 2023a.

Nan Jiang, Tengyu Liu, Zhexuan Cao, Jieming Cui, Zhiyuan Zhang, Yixin Chen, He Wang, Yixin Zhu, and Siyuan Huang. Full-body articulated human-object interaction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9365–9376, October 2023b.

Nan Jiang, Yunhao Li, Lexi Pang, Zimo He, Siyuan Huang, and Yixin Zhu. Motionmaster: Generalizable text-driven motion generation and editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 30629–30639, 2026.

Hwanhee Jung, Seunggwan Lee, Jeongyoon Yoon, SeungHyeon Kim, Giljoo Nam, Qixing Huang, and Sangpil Kim. Decoupled generative modeling for human-object interaction synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2253–2263, 2026.

Hyeonwoo Kim, Sangwon Baik, and Hanbyul Joo. David: Modeling dynamic affordance of 3D objects using pre-trained video diffusion models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10330–10341. IEEE, 2025.

Jiaman Li, Jiajun Wu, and C Karen Liu. Object motion guided human motion synthesis. *ACM Transactions on Graphics (TOG)*, 42(6):1–11, 2023.

Jiaman Li, Alexander Clegg, Roozbeh Mottaghi, Jiajun Wu, Xavier Puig, and C Karen Liu. Controllable human-object interaction synthesis. In *European Conference on Computer Vision*, pp. 54–72. Springer, 2024.

Lei Li and Angela Dai. HOI-PAGE: Zero-shot human-object interaction generation with part affordance guidance. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=c7NEZ94WAq.

Stefan Lionar and Gim Hee Lee. Teamhoi: Learning a unified policy for cooperative human-object interactions with any team size. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 37121–37132, June 2026.

Jiaxin Lu, Chun-Hao Paul Huang, Uttaran Bhattacharya, Qixing Huang, and Yi Zhou. Humoto: A 4D dataset of mocap human object interactions. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10886–10897. IEEE, 2025.

Changfeng Ma, Yang Li, Xinhao Yan, Jiachen Xu, Yunhan Yang, Chunshi Wang, Zibo Zhao, Yanwen Guo, Zhuo Chen, and Chunchao Guo. P3-sam: Native 3D part segmentation. *arXiv preprint arXiv:2509.06784*, 2025.

Zichong Meng, Yiming Xie, Xiaogang Peng, Zeyu Han, and Huaizu Jiang. Rethinking diffusion for text-driven human motion generation: Redundant representations, evaluation, and masked autoregression. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27859–27871. IEEE, 2025.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA 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.

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 (CVPR) Workshops*, pp. 2903–2913, June 2025.

Ilya A Petrov, Riccardo Marin, Julian Chibane, and Gerard Pons-Moll. Tridi: Trilateral diffusion of 3D humans, objects, and interactions. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5523–5535. IEEE, 2025.

Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017.

Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: modeling and capturing hands and bodies together. *ACM Trans. Graph.*, 36(6), November 2017. ISSN 0730-0301. doi: 10.1145/3130800.3130883.

Yawen Shao, Wei Zhai, Yuhang Yang, Hongchen Luo, Yang Cao, and Zheng-Jun Zha. Great: Geometry-intention collaborative inference for open-vocabulary 3D object affordance grounding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17326–17336. IEEE, 2025.

Wenfeng Song, Xinyu Zhang, Shuai Li, Yang Gao, Aimin Hao, Xia Hou, Chenglizhao Chen, Ning Li, and Hong Qin. HOIanimator: Generating text-prompt human-object animations using novel perceptive diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 811–820, 2024.

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*, pp. 581–600. Springer, 2020.

Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=SJ1kSyO2jwu.

Jonathan Tseng, Rodrigo Castellon, and Karen Liu. Edge: Editable dance generation from music. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 448–458, 2023.

Hanqing Wang, Mingyu Liu, Xiaoyu Chen, Chengwei Ma, Yiming Zhong, Wenti Yin, Yuhao Liu, Zhiqing Cui, Jiahao Yuan, Lu Dai, et al. Videoafford: Grounding 3D affordance from human-object-interaction videos via multimodal large language model. *arXiv preprint arXiv:2602.09638*, 2026a.

Tuo Wang, Zhenbo Song, Zhenyu Lou, Qiongjie Cui, Luoming Zhang, Cheng Cheng, Haofan Wang, Xu Tang, Huaxia Li, and Hong Zhou. Harmonizing stochasticity and determinism: Scene-responsive diverse human motion prediction. *Advances in Neural Information Processing Systems*, 37:39784–39811, 2024a.

Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. Humanise: Language-conditioned human motion generation in 3D scenes. *Advances in Neural Information Processing Systems*, 35:14959–14971, 2022.

Zan Wang, Yixin Chen, Baoxiong Jia, Puhao Li, Jinlu Zhang, Jingze Zhang, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. Move as you say, interact as you can: Language-guided human motion generation with scene affordance. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 433–444. IEEE, 2024b.

Ziyin Wang, Sirui Xu, Chuan Guo, Bing Zhou, Jiangshan Gong, Jian Wang, Yu-Xiong Wang, and Liang-Yan Gui. Unleashing guidance without classifiers for human-object interaction animation. In *ICLR*, 2026b.

Lin Wu, Zhixiang Chen, and Jianglin Lan. HOI-dyn: Learning interaction dynamics for human-object motion diffusion. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 90795–90825. Curran Associates, Inc., 2025a.

Qianyang Wu, Ye Shi, Xiaoshui Huang, Lan Xu, Jingyi Yu, and Jingya Wang. Thor: Text to human-object interaction diffusion via relation intervention. In *2025 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–7, 2025b.

Xiaofei Wu, Yi Zhang, Yumeng Liu, Yuexin Ma, Yujiao Shi, and Xuming He. AffordGrasp: Cross-modal diffusion for affordance-aware grasp synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15944–15953, June 2026.

Xiaoyang Wu, Daniel DeTone, Duncan Frost, Tianwei Shen, Chris Xie, Nan Yang, Jakob Engel, Richard Newcombe, Hengshuang Zhao, and Julian Straub. Sonata: Self-supervised learning of reliable point representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22193–22204, June 2025c.

Zhen Wu, Jiaman Li, Pei Xu, and C Karen Liu. Human-object interaction from human-level instructions. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11176–11186. IEEE, 2025d.

Zeqi Xiao, Tai Wang, Jingbo Wang, Jinkun Cao, Wenwei Zhang, Bo Dai, Dahua Lin, and Jiangmiao Pang. Unified human-scene interaction via prompted chain-of-contacts. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=1vCnDyQkjg`.

Miao Xu, Xiangyu Zhu, Zidu Wang, Xusheng Liang, Bao Li, Jinlin Wu, Zelin Zang, and Zhen Lei. RegenHOI: Unifying reconstruction and generation for 3D human-object interaction understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 42847–42857, 2026a.

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 (ICCV)*, pp. 14928–14940, October 2023.

Sirui Xu, Ziyin Wang, Yu-Xiong Wang, and Liang-Yan Gui. Interdreamer: Zero-shot text to 3D dynamic human-object interaction. *Advances in Neural Information Processing Systems*, 37: 52858–52890, 2024.

Sirui Xu, Dongting Li, Yucheng Zhang, Xiyan Xu, Qi Long, Ziyin Wang, Yunzhi Lu, Shuchang Dong, Hezi Jiang, Akshat Gupta, et al. Interact: Advancing large-scale versatile 3D human-object interaction generation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7048–7060. IEEE, 2025a.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12266–12277. IEEE, 2025b.

Sirui Xu, Samuel Schulter, Morteza Ziyadi, Xialin He, Xiaohan Fei, Yu-Xiong Wang, and Liang-Yan Gui. Interprior: Scaling generative control for physics-based human-object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23516–23527, 2026b.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Jie Yang, Xuesong Niu, Nan Jiang, Ruimao Zhang, and Siyuan Huang. F-HOI: Toward fine-grained semantic-aligned 3D human-object interactions. In *European Conference on Computer Vision*, pp. 91–110. Springer, 2024.

Ting Yu, Yi Lin, Jun Yu, Zhenyu Lou, and Qiongjie Cui. Vision-guided action: Enhancing 3D human motion prediction with gaze-informed affordance in 3D scenes. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12335–12346. IEEE, 2025.

Zhengdi Yu, Simone Foti, Linguang Zhang, Amy Zhao, Cem Keskin, Stefanos Zafeiriou, and Tolga Birdal. Geometric neural distance fields for learning human motion priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2232–2242, 2026.

Ye Yuan, Jiaming Song, Umar Iqbal, Arash Vahdat, and Jan Kautz. Physdiff: Physics-guided human motion diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16010–16021, October 2023.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12358–12369. IEEE, 2025.

Jinlu Zhang, Yixin Chen, Zan Wang, Jie Yang, Yizhou Wang, and Siyuan Huang. Interactanything: Zero-shot human object interaction synthesis via LLM feedback and object affordance parsing. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7015–7025. IEEE, 2025a.

Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *IEEE transactions on pattern analysis and machine intelligence*, 46(6):4115–4128, 2024.

Zhenhao Zhang, Ye Shi, Lingxiao Yang, Suting Ni, Qi Ye, and Jingya Wang. OpenHOI: Open-world hand-object interaction synthesis with multimodal large language model. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 166582–166612, 2025b.

Weiyu Zhao, Ru Li, Jiaqi Liu, Sizhe Zhao, Qinglin Liu, and Shengping Zhang. Queryme: Query-driven open-vocabulary 3D object affordances grounding from multimodal evidence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2603–2612, 2026.

He Zhu, Quyu Kong, Kechun Xu, Xunlong Xia, Bing Deng, Jieping Ye, Rong Xiong, and Yue Wang. Grounding 3D object affordance with language instructions, visual observations and interactions. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17337–17346. IEEE, 2025.

APPENDIX

A Evaluation Metrics

We provide the complete definitions of the evaluation metrics used in our experiments. Unless otherwise specified, all geometric metrics are computed on the decoded human–object interaction sequences in world coordinates.

Interaction Evaluator. Following LIGHT Wang et al. (2026b), we evaluate complete human–object interaction sequences rather than human motion alone. The evaluator's motion input comprises global human joint positions, object rotations and translations, and dynamic object Basis Point Set (BPS) features computed using a fixed basis point set. Its pretrained motion and text encoders map interaction sequences and text descriptions into a shared feature space.

FID. We compute Fréchet Inception Distance (FID) using the same pretrained interaction evaluator adopted for R-Precision. The evaluator maps complete human–object interaction motions into a 256-dimensional latent space. FID measures the distribution discrepancy between the generated interaction motions and the ground-truth motions in this learned representation space. Lower FID indicates that the generated motion distribution is closer to the ground-truth distribution.

R-Precision. R-Precision evaluates the semantic consistency between the generated human–object interaction and its corresponding text description. The pretrained evaluator embeds both the text description and the complete HOI motion into a shared 256-dimensional feature space. For each generated motion, retrieval is performed among 256 in-batch text candidates according to their embedding similarity. We report Top-1, Top-2, and Top-3 R-Precision, where a higher value indicates better text–motion semantic alignment.

Optimal Contact Distance (OCD). Following ChainHOI Zeng et al. (2025), we evaluate contact consistency using a multi-reference contact-distance metric. For a generated interaction $\hat{X}$ conditioned on text y and object o, let $\mathcal{R}(y, o)$ denote the set of object-matched, semantically consistent ground-truth reference interactions, and let $C^{(r)}$ denote the contact annotations of reference r:

$$\mathrm{OCD}(\hat{X} \mid y, o) = \min_{r \in \mathcal{R}(y,o)} \mathrm{CD}_{\mathrm{ChainHOI}}\big(\hat{X}; C^{(r)}\big), \tag{7}$$

where $\mathrm{CD}_{\mathrm{ChainHOI}}(\hat{X}; C^{(r)})$ denotes the contact distance computed for the generated interaction using reference annotations $C^{(r)}$ under ChainHOI's evaluation protocol. The minimum is taken across reference annotations for the same generated interaction, accommodating alternative contact patterns consistent with the interaction semantics rather than relying on a single reference. Lower OCD indicates better contact consistency under this multi-reference protocol.

Penetration Score (Pene). Following LIGHT Wang et al. (2026b), we measure human–object interpenetration by querying the signed distance field (SDF) of the generated human body at all object vertices. For a generated interaction $\hat{X}$ with L frames and N_o object vertices, let $\hat{\mathbf{o}}_{l,j}$ denote object vertex j at frame l, and let $s_l^h(\mathbf{x})$ denote the human-body SDF, with negative values inside the body:

$$\mathrm{Pene}(\hat{X}) = \frac{1}{L} \sum_{l=1}^{L} \sum_{j=1}^{N_o} \max\big\{-s_l^h(\hat{\mathbf{o}}_{l,j})\,,\, 0\big\}\,. \tag{8}$$

Non-penetrating vertices contribute zero. We sum the penetration depth over all object vertices in each frame and average the result over the sequence. Pene therefore measures the frame-averaged aggregate penetration depth. Lower values indicate less severe human–object interpenetration.

Foot Skating Ratio (FSR). FSR measures foot sliding during ground contact. Let $\mathcal{G}$ denote the set of frames identified as grounded using a foot-height threshold of $5\,\mathrm{cm}$. Let $\mathcal{S} \subseteq \mathcal{G}$ denote grounded frames whose inter-frame horizontal foot displacement exceeds $2.5\,\mathrm{cm}$:

$$\mathrm{FSR}(\hat{X}) = \frac{|\mathcal{S}|}{|\mathcal{G}|}, \qquad |\mathcal{G}| > 0. \tag{9}$$

FSR is a dimensionless proportion in $[0, 1]$. Its denominator contains grounded frames rather than all sequence frames. Lower values indicate less foot sliding during ground contact.

B Training Details

PACHO is trained with three complementary objectives: motion reconstruction, contact-affordance salience supervision, and salience-weighted contact alignment. At each diffusion step t, the final training objective is:

$$\mathcal{L}^{(t)} = \mathcal{L}^{(t)}_{\text{mot}} + \lambda_{\text{aff}}\mathcal{L}^{(t)}_{\text{aff}} + \lambda_{\text{sca}}(t)\mathcal{L}^{(t)}_{\text{sca}}. \tag{10}$$

Salience-Weighted Contact Alignment. The salience-weighted contact alignment loss $\mathcal{L}^{(t)}_{\text{sca}}$ is defined in Sec. 3.4. It penalizes the point-to-surface distance between contact-relevant human vertices and object surfaces, and vice versa. Its weight is scheduled as

$$\lambda_{\text{sca}}(t) = \lambda^{\text{base}}_{\text{sca}}\bar{\alpha}_t^{\gamma}, \tag{11}$$

where $\lambda^{\text{base}}_{\text{sca}}$ controls the overall strength of the alignment loss and γ controls how rapidly geometric guidance decays at noisy diffusion steps.

Motion Reconstruction Loss. The motion reconstruction loss supervises the clean human and object motion estimates predicted by the denoiser:

$$\mathcal{L}^{(t)}_{\text{mot}} = \left\| \tilde{\mathbf{X}}^{h,(t)}_{1:L} - \mathbf{X}^{h}_{1:L} \right\|_1 + \left\| \tilde{\mathbf{X}}^{o,(t)}_{1:L} - \mathbf{X}^{o}_{1:L} \right\|_1 . \tag{12}$$

In practice, this loss is applied to the translation, rotation, and pose components of the human and object motion states.

Contact-Affordance Salience Supervision. The predicted human contact salience and object affordance salience are supervised by pseudo labels $A^{h,*}$ and $A^{o,*}$ over the full sequence:

$$\mathcal{L}^{(t)}_{\text{aff}} = \frac{1}{L}\sum_{l=1}^{L}\left[\frac{1}{n'_p}\sum_i \text{BCE}\Big(A^{h,(t)}_{l,i}, A^{h,*}_{l,i}\Big) + \frac{1}{n'_p}\sum_j \text{BCE}\Big(A^{o,(t)}_{l,j}, A^{o,*}_{l,j}\Big)\right], \tag{13}$$

where n'_p denotes the number of sampled points for both the human and the object. We set $\lambda_{\text{aff}} = 0.5$ in our experiments.

C Additional Experimental Results

Part-aware Salience Design. Table 3(a) studies different designs for human contact salience and object affordance salience. Global salience provides only coarse interaction cues and performs poorly on contact-sensitive metrics, since it cannot distinguish fine-grained contact-relevant regions. Point-level salience improves localization by providing geometric guidance, while part-conditioned salience further benefits from structured human and object part embeddings. The full part-aware point-level salience design achieves the best R-Precision, OCD, and Pene, suggesting that conditioning point-wise salience prediction on part embeddings enables more accurate contact-affordance reasoning.

Co-Prediction Target. Table 3(b) evaluates the impact of different co-prediction targets. Motion-only prediction lacks explicit spatial guidance and therefore suffers from larger contact distance and penetration artifacts. Introducing either human contact salience or object affordance salience improves physical consistency, but single-side guidance cannot fully capture the coupled nature of human-object interactions. In contrast, parallel contact-affordance co-prediction achieves the best performance across all metrics, showing that co-predicting human contact salience and object affordance salience provides stronger guidance for physically plausible interaction synthesis.

Affordance Label Generation. Table 3(c) studies the effect of different pseudo-label generation strategies. Geometry-only labels rely solely on spatial proximity and therefore introduce noisy supervision when body parts are close to the object but not functionally involved. Adding Text2Prior

Table 3: **Additional ablation studies on the InterAct Xu et al. (2025a) dataset.** We analyze the salience representation, co-prediction target, affordance label generation, and salience-weighted contact alignment designs of PACHO. The results validate the importance of part-aware point-level salience, parallel contact-affordance co-prediction, text-guided pseudo-label construction, and scheduled contact alignment for physically plausible HOI synthesis.

Salience Design	R-Top3 ↑	OCD ↓	Pene ↓
Global Salience	0.645	0.155	0.125
Point-level Salience	0.755	0.082	0.088
Part-conditioned Salience	0.761	0.073	0.074
Part-aware Point-level Salience	**0.768**	**0.065**	**0.062**

(a) **Part-aware salience design.** Conditioning point-level human contact salience and object affordance salience on part embeddings provides finer interaction localization than global or part-agnostic alternatives.

Co-Prediction Target	FID ↓	R-Top3 ↑	OCD ↓	Pene ↓
Motion Only	0.101	0.712	0.128	0.118
Motion + Human Contact	0.093	0.742	0.087	0.083
Motion + Object Affordance	0.094	0.738	0.090	0.086
Parallel Contact-Affordance	**0.085**	**0.768**	**0.065**	**0.062**

(b) **Co-Prediction target.** Jointly co-predicting human contact salience and object affordance salience provides stronger spatial guidance than single-side supervision.

Affordance Label	FID ↓	R-Top3 ↑	OCD ↓	Pene ↓
Geometry Only	0.099	0.724	0.118	0.109
+ Text2Prior	0.094	0.742	0.093	0.088
+ Temporal Smoothing	0.089	0.756	0.076	0.073
+ Part Filtering / Full	**0.085**	**0.768**	**0.065**	**0.062**

(c) **Affordance label generation.** Text-conditioned priors reduce noisy proximity labels, while temporal smoothing and part-level filtering further improve contact-affordance supervision.

Contact Alignment	FID ↓	R-Top3 ↑	OCD ↓	Pene ↓
w/o Alignment Loss	0.092	0.752	0.092	0.084
Uniform Alignment	0.089	0.758	0.079	0.076
Salience-Weighted Alignment	0.086	0.764	0.069	0.067
+ Timestep Schedule / Full	**0.085**	**0.768**	**0.065**	**0.062**

(d) **Salience-weighted contact alignment.** Salience-weighted alignment focuses geometric supervision on contact-relevant regions, while timestep scheduling reduces unreliable guidance at noisy diffusion steps.

improves both semantic alignment and contact-related metrics by filtering contact candidates according to the action description. Temporal smoothing further reduces frame-wise label noise, while part-level filtering suppresses isolated activations and produces the best overall performance. These results show that text-guided and temporally consistent pseudo-label construction is important for reliable contact-affordance supervision.

Robustness to Part Segmentation Errors. Human body segmentation is reliably derived from the predefined SMPL-H topology, whereas object part segmentation relies on frozen P3-SAM inference. To evaluate the sensitivity of PACHO to object segmentation errors, we perturb the P3-SAM outputs at a fixed severity of 20% using three representative error types: boundary reassignment, missing-part merging, and incorrect part assignment.

Table 4: **Robustness to object part segmentation errors on the InterAct dataset.**

Segmentation Setting	R-Top3 ↑	FID ↓	OCD ↓	Pene ↓	FSR ↓
Clean P3-SAM	0.768	0.085	0.065	0.062	0.058
Boundary Noise	0.762	0.091	0.069	0.065	0.060
Missing Parts	0.752	0.102	0.075	0.071	0.065
Wrong Assignment	0.741	0.118	0.084	0.079	0.073

As shown in Table 4, PACHO remains reasonably robust to different object part segmentation errors. Boundary noise causes only minor degradation because the overall part structure is largely preserved. Missing parts lead to a larger performance drop due to the loss of local semantic cues, while incorrect part assignment produces the strongest degradation by introducing misleading part-level guidance. Nevertheless, all perturbed variants remain stable without catastrophic performance degradation, indicating that the proposed part-aware contact-affordance modeling is not overly sensitive to moderate segmentation errors.

Inference Efficiency and Fast Sampling. To evaluate the inference efficiency of PACHO, we compare the default 1000-step DDPM sampler with accelerated DDIM sampling using 100 and 50 steps on the InterAct dataset. Since PACHO adopts the $\mathbf{x}_0$-prediction parameterization, DDIM

sampling can be directly applied at inference time without retraining. All variants use the same pretrained checkpoint, conditional inputs, random seeds, and evaluation protocol.

Table 5: **Inference efficiency under different sampling configurations on the InterAct dataset.**

Sampler	FID ↓	R-Top1 ↑	R-Top2 ↑	R-Top3 ↑	FSR ↓	Time ↓
DDPM (1000)	0.085	0.455	0.658	0.768	0.058	38.33 s
DDIM (100)	0.083	0.451	0.653	0.760	0.057	3.99 s
DDIM (50)	0.082	0.449	0.651	0.764	0.055	1.97 s

As shown in Table 5, DDIM substantially reduces inference cost while preserving generation quality. DDIM-100 reduces the sampling time from 38.33 s to 3.99 s, corresponding to a $9.6\times$ speedup. DDIM-50 further reduces the inference time to 1.97 s, achieving a $19.5\times$ speedup with only minor changes in semantic alignment and slightly improved FID and FSR.

D Additional Implementation Details

Fair Comparison Protocol. To ensure a fair comparison, all methods are evaluated using the same dataset splits, text annotations, and evaluation scripts. When supported by the corresponding baseline, we also provide the same initial human-object states and object geometries. For methods originally designed for a different but related setting, such as human-only motion generation or non-text-conditioned HOI synthesis, we adapt only their input-output interfaces to the target benchmark while keeping their core architectures unchanged. Specifically, MDM Tevet et al. (2023) is fine-tuned on the target HOI datasets using the same training split and text annotations. For HOI-specific baselines, including InterDiff Xu et al. (2023), CHOIS Li et al. (2024), HOI-Diff Peng et al. (2025), Text2HOI Cha et al. (2024), ChainHOI Zeng et al. (2025), HOI-Dyn Wu et al. (2025a) and LIGHT Wang et al. (2026b), we follow their official implementations and retrain or evaluate them on the same benchmark splits according to their released protocols. All generated sequences are normalized to the same motion representation and sequence length before evaluation. During inference, each method receives the same text prompt and initial human-object configuration whenever applicable, and no method-specific post-processing is applied except for the standard mesh reconstruction required by each model. All metrics, including FID, R-Precision, OCD, Pene, and FSR, are computed using the same evaluation scripts to avoid implementation-induced bias.

Data Split and Test-Set Leakage. We strictly follow the standard train/validation/test split and ensure that no test-set information is used during training. Specifically, the dataset is first partitioned into training, validation, and test sets, after which pseudo-labels are generated exclusively for the training sequences. Ground-truth human and object meshes used for constructing contact and affordance supervision are therefore restricted to the training split. No ground-truth motion, pseudo-label, or annotation from the validation or test sequences is involved in the construction of the training supervision. At inference time, PACHO does not access any ground-truth motion or pseudo-labels, and generates the interaction solely from the text description and the initial human–object geometry provided as task conditions.

Part Parsing and Preprocessing. For representation consistency across benchmarks, we convert OMOMO's original SMPL-X representation to SMPL-H and use SMPL-H throughout our experiments, following Wang et al. (2026b). Human body part labels were obtained from the predefined SMPL-H segmentation Romero et al. (2017). Object part decomposition was performed using the automatic P3-SAM segmentation pipeline Ma et al. (2025). For each object mesh, we first cleaned the mesh by merging vertices and processing its topology. We then uniformly sampled 100,000 surface points and normalized them into a unit-centered bounding-box coordinate space. Next, we selected 400 point prompts via farthest point sampling and fed them, together with Sonata point features Wu et al. (2025c), into P3-SAM. For each prompt, P3-SAM predicted three candidate masks with corresponding IoU scores, and we retained the mask with the highest predicted IoU. The selected masks were binarized at a threshold of 0.5, clustered by mask IoU (> 0.9), and further merged according to 3D bounding-box IoU (> 0.5). We additionally recovered uncovered regions by assigning masks with sufficient overlap with the uncovered points. Finally, we applied the official

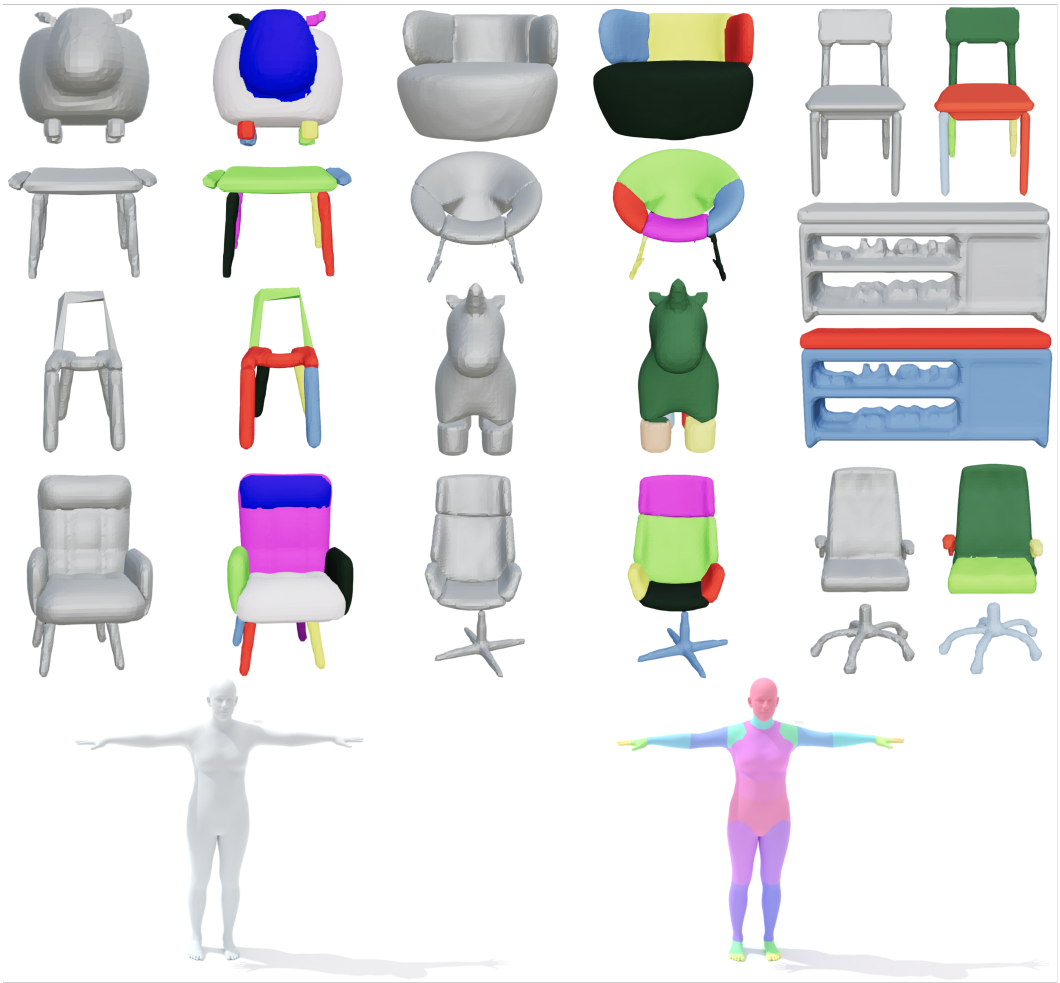


Figure 5: **Human and object part segmentation results.** We visualize representative part segmentation results used for part-aware contact-affordance reasoning. For each object, the gray mesh shows the original geometry and the colored mesh shows the corresponding part decomposition obtained by the part parser. The bottom row shows the predefined SMPL-H human body segmentation, where different colors denote different semantic body parts. These segmentations provide part-level structural priors for predicting human contact salience and object affordance salience.

area-based post-processing strategy with a threshold of 0.95, which merges small residual regions after the cumulative covered area exceeds 95% of the mesh surface. The resulting face labels on the cleaned mesh were projected back to the original mesh using rounded face-geometry matching at 8 decimal places, with centroid-based nearest-face assignment used only for unmatched faces. The segmentation results are shown in Figure 5.

E Contact-Affordance Label Annotation

We describe the automatic procedure used to generate affordance pseudo-labels for dense supervision of human contact and object affordance prediction. This procedure is used only during training and is not required at inference time. Given a human-object interaction sequence, we reconstruct the human mesh with SMPL-H and transform sampled object surface points to each frame, so that human vertices and object points are represented in a shared coordinate system.

Text-Conditioned Human Prior. Pure geometric proximity may produce noisy labels, since body parts can be close to an object without being functionally involved in the interaction. We therefore

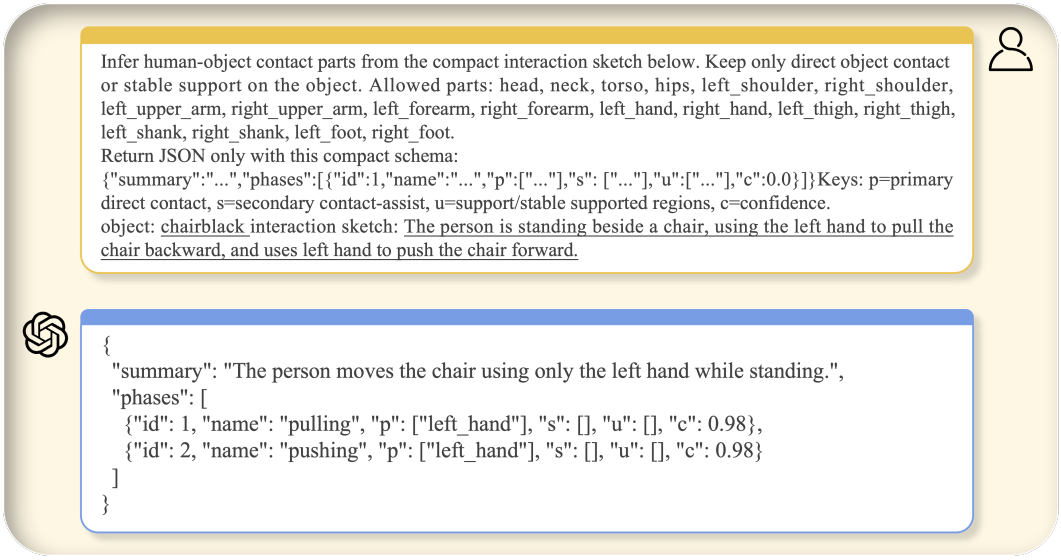


Figure 6: **Text-to-prior parsing for human contact regions.** Given an action description and a predefined set of human body parts, Qwen3 Yang et al. (2025) parses the interaction into structured contact phases and predicts primary contact parts, secondary contact parts, support parts, and confidence scores in JSON format. The parsed body-part priors are then mapped to SMPL-H vertices to construct the text-conditioned human contact prior used for affordance pseudo-label generation.

use the action text to infer a coarse prior over plausible human contact regions. Specifically, we implement an offline text-to-prior parser with Qwen3 Yang et al. (2025), which maps the action description to a vertex-level semantic prior $\mathbf{P}$ and a candidate contact mask $\mathbf{M}$:

$$(\mathbf{P}, \mathbf{M}) = \text{Text2Prior}(y), \tag{14}$$

where P_i indicates the semantic relevance of human vertex i to the described interaction, and M_i removes vertices outside the candidate contact regions. The parser extracts primary contact parts, secondary contact parts, and support parts from the action text, which are then mapped to body-region labels and converted into vertex-level priors. The prompt used in Text2Prior is shown in Figure 6.

Proximity-Based Soft Labeling. We convert point-to-surface proximity into soft salience scores using a thresholded Gaussian kernel:

$$g_{\tau,\sigma}(d) = \begin{cases} 1, & d \le \tau, \\ \exp\left(-\frac{d^2}{2\sigma^2}\right), & d > \tau, \end{cases} \tag{15}$$

where τ is the contact threshold and σ controls the spatial decay.

Human Contact Salience. For each human vertex $h_{t,i}$, we measure its nearest distance to the object surface points at frame t, and modulate the proximity score with the text-conditioned prior:

$$A^H_{t,i} = g_{\tau_H,\sigma_H}\left(\min_j \|h_{t,i} - o_{t,j}\|_2\right) \cdot M_i \cdot (0.25 + 0.75P_i). \tag{16}$$

This formulation preserves geometrically close vertices while emphasizing body regions that are semantically relevant to the action description.

Object Affordance Salience. Object affordance labels are generated from high-confidence human contact vertices rather than from the entire human body. Let $\mathcal{C}_t = \{i \mid A^H_{t,i} \ge \max(\theta_{\min}, \rho \max_k A^H_{t,k})\}$ denote the selected human contact sources at frame t. For each object point $o_{t,j}$, its affordance score is computed according to the nearest selected human contact source:

$$A^O_{t,j} = g_{\tau_O,\sigma_O}\left(\min_{i \in \mathcal{C}_t} \|o_{t,j} - h_{t,i}\|_2\right). \tag{17}$$

Table 6: **Sensitivity to pseudo-label construction on the InterAct Xu et al. (2025a) dataset.** Geometry-only labels rely solely on human-object proximity. Shuffled Text Prior replaces the parsed text-conditioned prior of each sequence with that from a randomly selected action-mismatched sequence, simulating semantic parser errors. Full Pipeline uses the correct LLM-based text prior, temporal smoothing, and part-level filtering.

Pseudo-label Construction	FID ↓	R-Prec. Top3 ↑	OCD ↓	Pene ↓
Geometry Only	0.099	0.724	0.118	0.109
Shuffled Text Prior	0.107	0.703	0.134	0.123
Full Pipeline	**0.085**	**0.768**	**0.065**	**0.062**

Both human contact and object affordance maps are temporally smoothed and filtered at the part level to suppress isolated noisy activations. When object part labels are available, we retain the most activated object parts at each frame. Static masks are obtained by taking the temporal maximum of the dynamic salience maps and applying a fixed threshold. For training point sets with different sampling resolutions, affordance values are transferred by nearest-neighbor interpolation.

Release of Affordance and Contact Labels. To facilitate reproducibility, we will release the generated human contact salience labels and object affordance salience labels used in our training pipeline, subject to the licenses and redistribution terms of the original datasets. These labels are generated offline from ground-truth human-object meshes using our text-conditioned prior parser, geometric proximity computation, temporal smoothing, and part-level filtering. The released annotations will include frame-wise human contact salience, object affordance salience, sampled point correspondences, object part labels when available, and the scripts required to map the labels back to the corresponding dataset sequences. Since the labels are used only for training-time supervision and are not required during inference, releasing them will allow future work to reproduce our supervision pipeline without relying on additional LLM calls. We will also provide the prompt template, JSON output schema, body-part mapping rules, parser configuration, and fallback rules.

Sensitivity to Pseudo-label Construction. Since our contact-affordance supervision is generated from pseudo labels rather than native dataset annotations, we further evaluate the effect of pseudo-label quality. The geometry-only variant constructs labels solely from human-object proximity, while the shuffled-prior variant replaces each sequence's parsed text prior with an action-mismatched prior from another sequence. As shown in Table 6, the full pipeline achieves the best performance across all metrics, indicating that text-conditioned priors, temporal smoothing, and part-level filtering provide more reliable supervision. The shuffled-prior variant performs worse than geometry-only labeling, showing that mismatched semantic priors can suppress true contact regions and activate incorrect body parts. This confirms that PACHO benefits from meaningful contact-affordance supervision rather than arbitrary additional pseudo labels.

Salience-Weighted Contact Alignment. Table 3(d) evaluates the salience-weighted contact alignment design. Removing the alignment loss leads to larger contact distance and penetration artifacts, indicating that salience supervision alone is insufficient to enforce local geometric consistency. Uniform alignment improves contact quality but treats all potential contact regions equally. In contrast, salience-weighted alignment focuses geometric supervision on contact-relevant human vertices and affordance-relevant object points. Adding the timestep-dependent schedule further improves stability by weakening unreliable geometric guidance at highly noisy diffusion steps and strengthening contact alignment near cleaner denoising states.

Quantitative Evaluation of the Text-to-Prior Parser. To evaluate the reliability of the Qwen3-based text parser, we manually annotate 150 diverse action descriptions using a unified human-part vocabulary. Since the primary, secondary, and support contact priors jointly contribute to the construction of the candidate contact mask, we take their union as the predicted contact set and compare it against the human-annotated ground truth using standard multi-label classification metrics.

As shown in Table 7, the parser achieves an Exact Match score of 0.760 and a Jaccard score of 0.834. It also obtains strong micro-level performance, with a Micro Precision of 0.790, Micro Recall of 0.950, and Micro F1 of 0.863, indicating that the parser successfully identifies most contact-relevant

Table 7: **Quantitative evaluation of the Qwen3-based text-to-prior parser on 150 sampled action descriptions.** The union of the predicted primary, secondary, and support contact parts is evaluated against manually annotated contact-part labels.

Exact Match	Jaccard	Micro P	Micro R	Micro F1	Macro P	Macro R	Macro F1
0.760	0.834	0.790	0.950	0.863	0.580	0.680	0.610

body parts for common interactions. The lower macro-level scores mainly reflect the increased difficulty of recognizing less frequent body parts. In our pseudo-label construction pipeline, these text-derived priors are further filtered by human–object geometric proximity, temporal smoothing, and part-level constraints, reducing the influence of residual parsing errors.

Robustness to Pseudo-Label Noise. To evaluate the robustness of PACHO to residual errors in the pseudo-labels, we simulate two common types of annotation noise during training: false negatives (FN), where positive-label vertices are randomly removed, and false positives (FP), where spurious activations are introduced into negative regions. The perturbations are applied to both human-contact and object-affordance labels across all pseudo-label-dependent objectives. We evaluate noise rates of 10% and 30%.

Table 8: **Sensitivity to pseudo-label noise on the InterAct dataset Xu et al. (2025a).** FN denotes false-negative perturbations that remove positive-label vertices, while FP denotes false-positive perturbations that introduce spurious activations into negative regions.

Noise Type	Noise Rate	R-Prec. Top1 ↑	R-Prec. Top2 ↑	R-Prec. Top3 ↑	FID ↓	OCD ↓	Pene ↓
Clean	0%	0.455	0.658	0.768	0.085	0.065	0.062
FN	10%	0.443	0.645	0.755	0.089	0.076	0.072
FP	10%	0.431	0.635	0.745	0.093	0.088	0.080
FN	30%	0.415	0.612	0.720	0.102	0.108	0.102
FP	30%	0.395	0.585	0.700	0.241	0.147	0.135

As shown in Table 8, moderate pseudo-label noise causes only limited degradation in generation performance. At a 10% noise rate, R-Precision decreases slightly while FID, OCD, and Pene remain close to the clean setting. Performance degradation becomes more pronounced at the 30% noise level. In particular, FP perturbations are consistently more harmful than FN perturbations: removing positive labels mainly reduces the density of effective supervision, whereas introducing false-positive labels creates incorrect correspondences between non-interacting human and object regions and can directly misguide the model. These results indicate that PACHO remains reasonably robust to moderate pseudo-label errors and does not require perfectly accurate pseudo supervision.

F Affordance Comparison

To better clarify the design choice of PACHO, we compare three representative affordance representations for human-object interaction synthesis: sparse contact-point affordance, part-level affordance, and our part-aware point-level dynamic affordance. These representations differ in spatial granularity, temporal adaptivity, and how they are coupled with motion generation.

Sparse Point-Level Affordance. As shown in Figure 7(a), sparse point-level affordance represents the interaction region with only a few predicted contact points on the object surface. HOI-Diff Peng et al. (2025) introduces an affordance prediction diffusion model to estimate such contact points from the text prompt, and incorporates them into classifier guidance to encourage closer human-object contact during motion generation. While this design provides an explicit contact cue and can correct errors from the motion generation branch, it is sensitive to inaccurate affordance prediction. When the predicted contact point is misplaced or significantly shifted, the guidance may pull the generated interaction toward an incorrect object region, thereby degrading geometric plausibility. For example, in Figure 7(a), the predicted interaction point lies near the center of the monitor; using

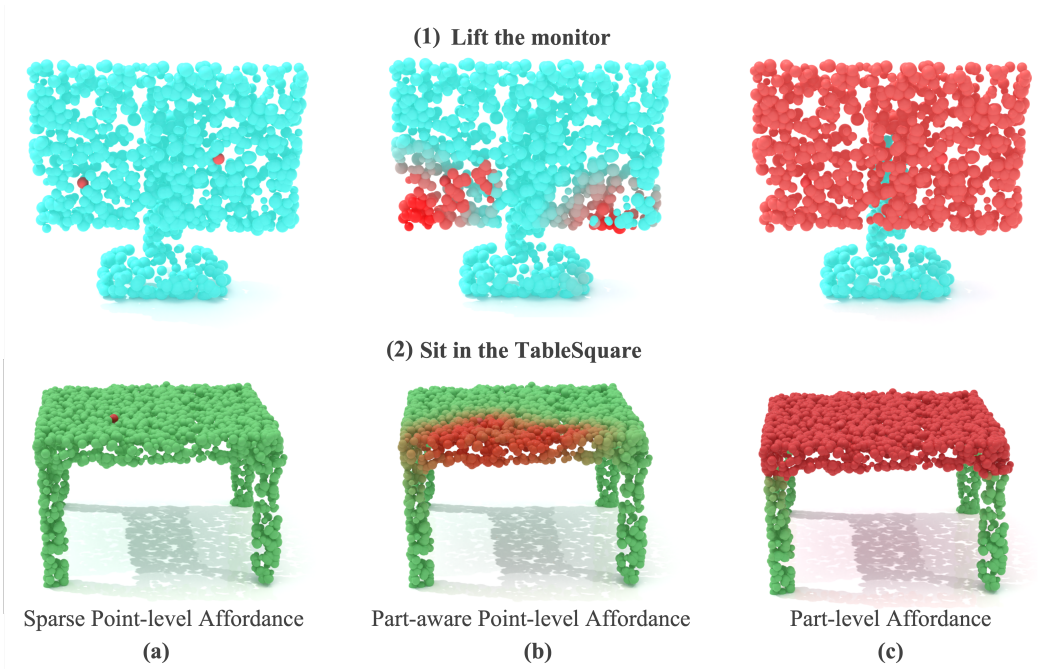


Figure 7: **Visualization of affordance granularity.** We compare three affordance representations under two interaction prompts: *Lift the monitor* and *Sit in the TableSquare*. Red indicates high affordance salience. Since affordance is dynamically predicted over the interaction sequence in our work, we visualize a representative frame at the key contact stage for comparison.

this point as a contact constraint may force the human body to approach an implausible region and lead to penetration or unnatural contact. Therefore, sparse point-level affordance is less effective in describing which human body parts and which object functional regions should interact across different stages of a complex action.

Part-Level Affordance. As shown in Figure 7(c), part-level affordance predicts object parts that are likely to participate in the interaction. Compared with sparse point-level affordance, this representation provides a broader interaction region, which can offer more flexible guidance and potentially improve the diversity of generated samples. HOI-PAGE Li & Dai (2026) advances affordance reasoning from sparse contact points to part-level interaction structure by introducing Part Affordance Graphs, which encode relations between human body parts and object parts for zero-shot 4D HOI synthesis. This part-level representation improves semantic interpretability and enables reasoning about how different body parts engage with functional object regions. However, part-level affordance is still defined over discrete semantic regions and therefore cannot provide dense point-wise guidance within each part. Its effectiveness also depends heavily on the quality and granularity of object part segmentation. When the segmented part is overly large or inaccurate, the guidance becomes too coarse and may reduce contact precision. For example, in Figure 7(c), the entire monitor screen is treated as an affordance part, whereas lifting the monitor typically requires contact with only a small local region such as the edge or frame. Using the whole part as the interaction condition may therefore lead to ambiguous or inaccurate contact prediction.

Part-Aware Point-Level Affordance. As shown in Figure 7(b), part-aware point-level affordance inherits the advantages of both sparse point-level affordance and part-level affordance. Sparse point-level affordance offers fine spatial cues but is sensitive to inaccurate point prediction, while part-level affordance provides semantic structure but remains coarse within each object part. PACHO further refines affordance modeling by combining part-level semantics with point-level spatial precision. Instead of predicting only sparse contact points or assigning a single affordance label to an entire object part, PACHO dynamically co-predicts human contact salience and object affordance salience over sampled human and object points during diffusion. The point-wise salience predictions are

conditioned on human and object part embeddings, making them part-aware while preserving fine-grained geometric localization. These dynamic salience fields are then used by the Cross-Affordance Interaction module to establish bidirectional contact-aware coupling between the decoupled human and object motion streams. This design enables PACHO to capture time-varying contact-affordance patterns, such as the transition from foot-object support to hand-object manipulation, while providing dense spatial guidance for geometrically consistent and contact-stable HOI synthesis.

Quantitative Affordance Type Comparison. To isolate the effect of affordance representation, we keep the same PACHO backbone and replace only the affordance guidance with three alternatives: sparse point-level affordance, part-level affordance, and our part-aware point-level affordance. As shown in Table 9, sparse point-level affordance improves contact guidance but remains limited by its sparse spatial representation, leading to larger contact distance and penetration errors. Part-level affordance provides more structured semantic guidance and improves both semantic alignment and physical plausibility, but its coarse granularity cannot precisely localize contact regions within object parts. In contrast, PACHO's part-aware point-level affordance achieves the best results across all metrics. These results suggest that dense point-wise salience, when conditioned on part semantics and dynamically co-predicted during diffusion, provides more effective spatial guidance for contact-consistent HOI synthesis.

Table 9: **Comparison of different affordance representations on the InterAct Xu et al. (2025a) dataset.** We implement sparse point-level affordance following HOI-Diff Peng et al. (2025) and part-level affordance following HOI-PAGE Li & Dai (2026).

Affordance Representation	FID ↓	R-Prec. Top3 ↑	OCD ↓	Pene ↓
Sparse Point-Level Affordance	0.097	0.731	0.108	0.101
Part-Level Affordance	0.091	0.752	0.081	0.077
Part-Aware Point-Level Affordance	**0.085**	**0.768**	**0.065**	**0.062**

Affordance Representation Validation. PACHO models object affordance as a dense, dynamic point-level field $A^o \in [0,1]^{L \times n'_p}$. We compare constrained variants of this representation. Specifically, sparse variants retain only the Top-K point responses through $A^o_{l,i}\mathbb{I}_{i \in \mathrm{TopK}(A^o_{l,:})}$. Part-level variants average responses within predefined regions, yielding $\frac{1}{|P_k|}\sum_{j \in P_k} A^o_{l,j}$. Static variants aggregate responses over the temporal dimension using $\max_l A^o_{l,i}$. These variants respectively restrict spatial coverage, within-part spatial detail, or temporal variation. The comparison uses a common backbone design, training objectives, and optimization schedules, with changes restricted to the representation and coupling mechanisms. We report parameter counts to account for the resulting differences in model capacity.

Table 10: **Capacity-controlled comparison of affordance representations on the OMOMO Li et al. (2023) dataset.**

Representation	Params (M)	R-Prec. Top1 ↑	R-Prec. Top2 ↑	R-Prec. Top3 ↑	FID ↓	OCD ↓	Pene ↓
Direct Motion Interaction	50.69	0.275	0.490	0.650	0.135	0.115	0.028
Sparse Dynamic	49.70	0.298	0.520	0.685	0.110	0.095	0.020
Part-Level Dynamic	49.70	0.302	0.525	0.690	0.105	0.092	0.019
Dense Static	49.70	0.316	0.548	0.712	0.090	0.084	0.015
Dense Dynamic	49.70	**0.325**	**0.560**	**0.725**	**0.082**	**0.075**	**0.012**

As shown in Table 10, Dense Dynamic achieves the best reported R-Precision, FID, OCD, and Pene among the compared variants. Despite having more parameters, Direct Motion Interaction performs worse than the explicit affordance variants, suggesting that increased capacity alone does not explain the gains. At the same parameter count, Dense Dynamic also outperforms Sparse Dynamic, Part-Level Dynamic, and Dense Static. These results support the benefit of preserving both point-level spatial detail and frame-wise temporal variation for human–object motion generation and geometric contact consistency.

G CONTACT-MOTION CONSISTENCY

To complement the geometric contact evaluation, we analyze motion synchronization between human and object surfaces during sustained contact using **Contact-Motion Consistency (CMC)**. We report comparisons with baseline methods on InterAct and representation ablations on OMOMO.

Metric Definition. Let $\mathcal{W}$ denote the set of sustained contact windows used for evaluation. Each window $w = (i, j, m, n)$ specifies a human vertex i and an object surface vertex j that maintain contact over the frame interval $[m, n]$, where $n > m$. The vertex identities i and j remain fixed within the window. Let $\hat{\mathbf{h}}_{l,i}$ and $\hat{\mathbf{o}}_{l,j}$ denote their positions in the decoded interaction sequence, expressed in world coordinates. Their inter-frame displacements are

$$\begin{aligned} \Delta\hat{\mathbf{h}}_{l,i} &= \hat{\mathbf{h}}_{l,i} - \hat{\mathbf{h}}_{l-1,i}, \\ \Delta\hat{\mathbf{o}}_{l,j} &= \hat{\mathbf{o}}_{l,j} - \hat{\mathbf{o}}_{l-1,j}. \end{aligned} \tag{18}$$

For each contact window, we compute

$$\mathrm{CMC}_w = \frac{1}{n-m} \sum_{l=m+1}^{n} \left\| \Delta\hat{\mathbf{h}}_{l,i} - \Delta\hat{\mathbf{o}}_{l,j} \right\|_2 . \tag{19}$$

The reported CMC is averaged over all evaluated sustained contact windows:

$$\mathrm{CMC} = \frac{1}{|\mathcal{W}|} \sum_{w \in \mathcal{W}} \mathrm{CMC}_w, \qquad |\mathcal{W}| > 0. \tag{20}$$

CMC is reported in meters. Lower values indicate better motion synchronization and less relative displacement between the contacting vertices within the evaluated windows. CMC is a kinematic measure of contact consistency rather than a direct evaluation of contact forces or dynamic feasibility. To construct $\mathcal{W}$, a human–object vertex pair (i, j) is considered in contact at frame t when $\|\hat{\mathbf{h}}_{t,i} - \hat{\mathbf{o}}_{t,j}\|_2 \le \tau_{\mathrm{contact}}$, where $\tau_{\mathrm{contact}} = 0.04\,\mathrm{m}$. A sustained-contact window is retained only when the same vertex pair satisfies this condition for at least $K_{\min} = 15$ consecutive frames, corresponding to approximately $0.5\,\mathrm{s}$ at 30 fps.

Table 11: **Contact-Motion Consistency on the InterAct dataset.** Lower values indicate better motion synchronization. The best result among generated methods is shown in **bold**.

Metric	CHOIS Li et al. (2024)	HOI-Diff Peng et al. (2025)	InterDiff Xu et al. (2023)	LIGHT Wang et al. (2026b)	**PACHO**	Ground Truth
CMC (m) ↓	0.045	0.052	0.038	0.042	**0.017**	0.008

Comparison with Baselines on InterAct. Table 11 compares PACHO with CHOIS, HOI-Diff, InterDiff, and LIGHT. Ground Truth is included as a reference and is excluded from best-method highlighting. PACHO achieves the best CMC among all generated methods and yields results closest to the Ground Truth reference, indicating better synchronized human–object motion and reduced relative displacement during sustained contact. Nevertheless, a gap to real interaction sequences remains.

Effect of Affordance Representation on OMOMO. We further isolate the CMC results of the representation variants examined in Table 10. Table 12 reports their contact-motion consistency separately from the other generation metrics.

Dense Dynamic obtains the lowest CMC of $0.017\,\mathrm{m}$, compared with $0.032\,\mathrm{m}$ for Direct Motion Interaction. Sparse Dynamic and Part-Level Dynamic reduce CMC to $0.025\,\mathrm{m}$ and $0.024\,\mathrm{m}$, respectively, but remain less consistent than the dense dynamic representation. Dense Dynamic also improves upon Dense Static ($0.020\,\mathrm{m}$) at the same parameter count. Together, these results support the use of spatially detailed, time-varying affordance guidance for improving motion synchronization during sustained human–object contact.

Table 12: **Effect of affordance representation on Contact-Motion Consistency on the OMOMO Li et al. (2023) dataset.**

Representation	CMC (m) ↓
Direct Motion Interaction	0.032
Sparse Dynamic	0.025
Part-Level Dynamic	0.024
Dense Static	0.020
Dense Dynamic	**0.017**

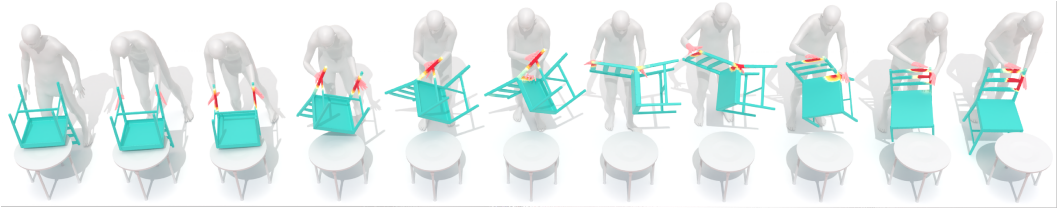


Figure 8: **Qualitative result on an unseen wooden chair from the OMOMO Li et al. (2023) dataset.** The red regions indicate high predicted contact-affordance salience.

H Additional Visual Comparison

Generalization to Unseen Object Geometry. Figure 8 shows a qualitative result on an unseen wooden chair from the OMOMO dataset. Although the object geometry is not observed during training, PACHO generates a coherent multi-stage interaction sequence in which the human first establishes contact with the chair, then lifts and rotates it while maintaining stable hand-object contact. This indicates that the proposed part-aware point-level affordance modeling can generalize to unseen object instances and provide effective spatial guidance for contact-consistent manipulation. The generated sequence also shows smooth object motion and plausible body adaptation, suggesting that the decoupled motion priors and cross-affordance coupling help preserve both object dynamics and human kinematic realism under novel object geometry.

Cross-Category Unseen Object Generalization. To evaluate generalization to unseen object categories, we follow the in-category and cross-category splits used by LIGHT Wang et al. (2026b). In the in-category split, the training and test sets share the same object categories, whereas the cross-category split contains object categories that are not observed during training. This setting therefore evaluates whether the learned interaction representation can transfer beyond the object categories seen during training.

As shown in Table 13, both methods experience a substantial performance drop under the cross-category setting, highlighting the difficulty of generalizing to unseen object categories. Nevertheless, PACHO consistently outperforms LIGHT across all reported metrics. In particular, PACHO improves R-Precision while substantially reducing FID, OCD, and Pene under the cross-category split. These results suggest that PACHO is less sensitive to category-level distribution shifts and can better transfer learned contact-affordance patterns to previously unseen object categories.

Additional visual comparisons. Figure 10 provides additional qualitative comparisons with CHOIS Li et al. (2024), HOI-Diff Peng et al. (2025), and LIGHT Wang et al. (2026b) across diverse interaction scenarios. In simple holding and raising actions, baseline methods often produce severe interpenetration or unstable object contact, causing the object to intersect the body or drift away from plausible support regions. In the long-horizon monitor manipulation sequence, LIGHT further exhibits unnatural joint rigidity and contact breakage during lifting, rotation, and placement. In contrast, PACHO maintains more stable human-object contact, reduces visible penetration, and generates smoother interaction transitions, demonstrating the benefit of dynamic contact-affordance guidance for physically plausible HOI synthesis.

Figure 9: **Generalization to unseen objects on the InterAct Xu et al. (2025a) dataset.** The in-category split contains unseen object instances from seen categories, whereas the cross-category split contains object categories not observed during training.

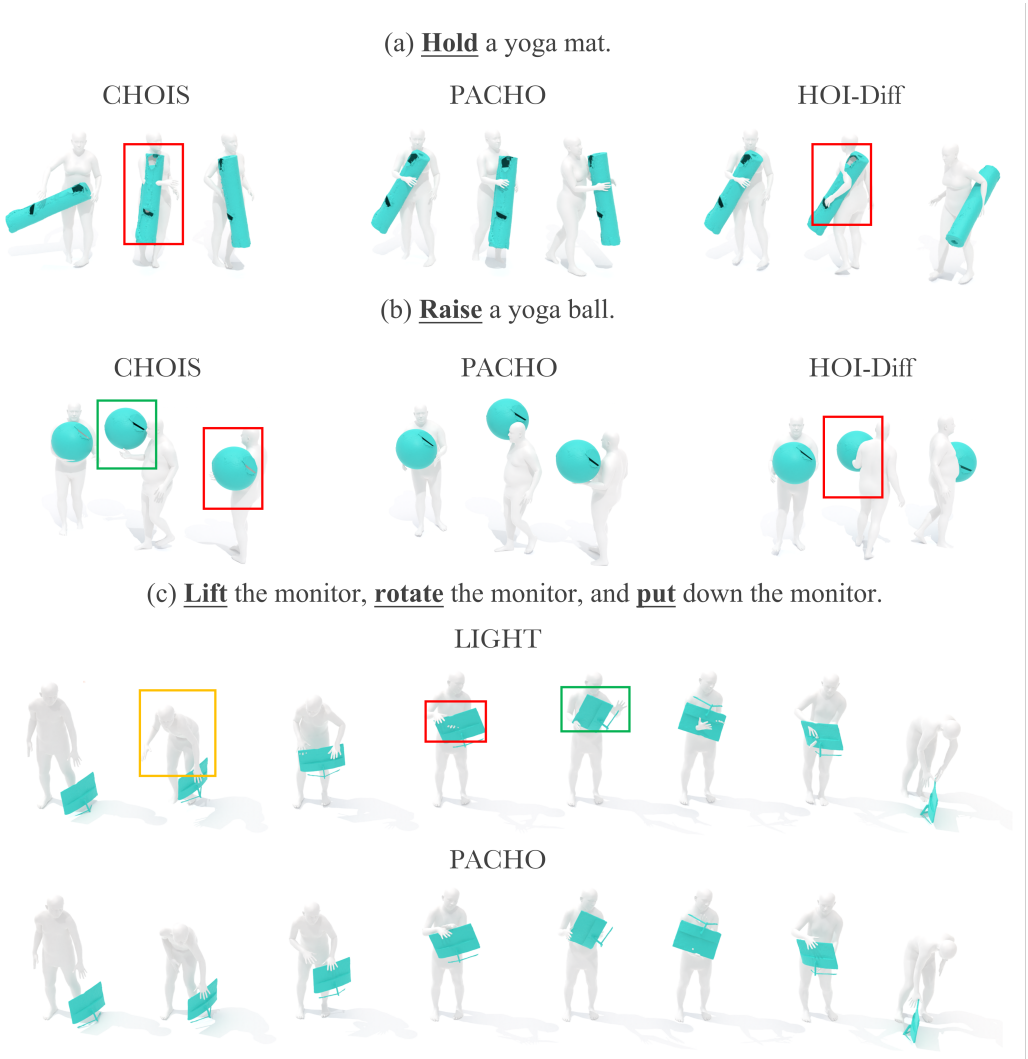


Figure 10: **Additional visual comparisons.** We compare PACHO with the state-of-the-art baselines CHOIS Li et al. (2024), HOI-Diff Peng et al. (2025), and LIGHT Wang et al. (2026b). The yellow boxes indicate unnatural joint rigidity, the red boxes highlight severe interpenetration, and the green boxes denote contact breakage during the interaction.

I Limitations and Future Work

PACHO Failure Analysis. Although PACHO improves contact consistency in most cases, it may still fail under rapid contact-affordance transitions, as shown in Figure 11. The Cross-Affordance Interaction module updates the human motion stream using object affordance cues and the object motion stream using human contact cues. When the predicted contact-affordance salience shifts rapidly or becomes inaccurate, the two motion streams may receive inconsistent or temporally unstable interaction guidance, resulting in abrupt pose transitions, contact drift, or delayed human-object coordination. In addition, our salience-weighted contact alignment loss relies on pseudo-label salience to identify contact-relevant human and object regions for geometric supervision. Inaccurate or overly broad pseudo-labels may therefore introduce inappropriate alignment constraints, encouraging the

Table 13: **Generalization to unseen objects on the InterAct Xu et al. (2025a) dataset.** The in-category split contains unseen object instances from seen categories, whereas the cross-category split contains object categories not observed during training.

Split	Method	R-Top1 ↑	R-Top2 ↑	R-Top3 ↑	FID ↓	OCD ↓	Pene ↓	FSR ↓
In-category	LIGHT	0.279	0.386	0.466	0.133	0.265	0.305	0.059
	PACHO	**0.342**	**0.495**	**0.551**	**0.088**	**0.092**	**0.068**	**0.052**
Cross-category	LIGHT	0.022	0.043	0.071	4.924	0.742	0.203	0.032
	PACHO	**0.105**	**0.132**	**0.156**	**1.525**	**0.295**	**0.112**	**0.026**



Figure 11: **Failure cases of PACHO.** The red box highlights severe human-object interpenetration, where the human body penetrates the object surface. The yellow box indicates contact inconsistency and unstable support, where the human body fails to maintain physically plausible contact with the object, leading to floating or sliding artifacts.

human body or object to approach implausible contact regions. These errors can lead to severe interpenetration, as highlighted by the red box, or unstable support and sliding contact, as highlighted by the yellow box. These failure cases suggest that future work should incorporate temporally smoother contact-affordance prediction, more reliable contact supervision, and physically grounded contact constraints to better handle fast multi-stage interactions.

Future Work. The observed failure cases point to broader directions for advancing contact-affordance-guided HOI synthesis. A natural next step is to elevate affordance modeling from local frame-wise prediction to hierarchical interaction reasoning, where high-level action stages and low-level contact transitions are jointly modeled to support smoother multi-stage motion generation. Beyond temporal coherence, physically realistic HOI also requires reasoning about support, force, friction, object mass, and dynamic stability, rather than relying solely on geometric contact consistency. Integrating physics-aware priors, differentiable simulation, or embodied feedback could therefore further reduce penetration and unstable support artifacts. Looking further ahead, open-world generalization remains a key challenge: extending the framework to articulated and deformable objects, multi-object scenes, and long-horizon compositional interactions would make contact-affordance reasoning more robust and applicable to realistic 3D environments.

J Potential Broader Impacts

The proposed PACHO introduces a text-driven 3D human-object interaction synthesis framework that explicitly models dynamic human contact and object affordance. By generating physically plausible and contact-consistent human-object motions from language and object geometry, PACHO may have the following broader impacts:

- **Safer Human-Robot Interaction and Embodied AI.** PACHO can support the simulation of realistic human-object interactions for embodied agents and robotic systems. By modeling time-varying contact regions and object affordances, the framework may help robots better anticipate how humans interact with everyday objects, reducing unsafe or physically implausible behaviors in shared environments.
- **Improved Virtual Reality, Gaming, and Digital Content Creation.** Text-driven HOI synthesis can make it easier to generate realistic character-object interactions in virtual environments. PACHO may benefit VR, gaming, animation, and digital storytelling by

producing motions with more stable contact, fewer floating artifacts, and better alignment between human actions and object functionality.

- **Synthetic Data for Interaction Understanding.** PACHO can potentially be used to generate diverse human-object interaction sequences for training or evaluating perception, motion understanding, and interaction recognition models. Such synthetic data may be especially useful when real 3D HOI data are expensive, difficult to capture, or limited in interaction diversity.
- **Responsible Use and Limitations.** Although PACHO improves physical plausibility and contact consistency, generated motions are not guaranteed to be physically executable in real-world robotic or safety-critical settings. Deployment in robotics, healthcare, or assistive applications should therefore require additional physical validation, human oversight, and domain-specific safety checks. Moreover, like other generative motion models, PACHO may inherit biases from training data and should be used carefully when modeling diverse body types, abilities, objects, and interaction contexts.