A^{3}fford-HOI: Anatomy-Aligned Affordance Disentanglement for Fine-grained and Generalizable Hand Object Interaction

Xurui Hu[1], Xingqun Qi[2], Bingkun Yang[1], Chen Su[1], Yiwei Ru[1,3], Junhui Yin[4], Muyi Sun[1], Man Zhang[1]

[1]BUPT,
[2]HKUST,
[3]CASIA,
[4]USTB

{xingqunqi, junhuiyin23}@gmail.com, {2024141206, zhihenxue, suchen1201, muyi.sun, zhangman}@bupt.edu.cn, yiwei.ru@cripac.ia.ac.cn

Abstract

Hand Object Interaction (HOI) generation provides a feasible solution for virtual reality simulation and embodied AI deployment. Recent studies explore the instruction-driven HOI synthesis, yet they overlook the fine-grained interactive contact and struggle with robust generalization in data-scarce scenarios. In addition, existing datasets are frequently constrained with limited object categories and monotonous motion patterns. To address these limitations, we first integrate a comprehensive HOI dataset, namely **HOI-X**, which contains diverse objects, rich motion patterns, and part-level contact annotations. Along with HOI-X, we propose **A^{3}fford-HOI**, a fine-grained and generalizable **H**and **O**bject **I**nteraction framework with **A**natomy-**A**ligned **Afford**ance Disentanglement.Specifically, we define a **Disentangle-2-Interact** paradigm in **A^{3}fford-HOI**. Considering the various interactions of distinct hand anatomical structures, in the **Disentangle** stage, we devise an **Anatomy-Aligned Affordance Predictor**, which leverages the external plausible physical priors in 3D multimodal large language models for affordance disentanglement prediction. Then, the disentangled affordance is employed in the **Interact** stage, dubbed **Affordance Driven Interactor**. The interactor integrates textual instructions, object geometric features, and the external knowledge enhanced (*i.e.*MLLMs enhanced) affordance via a Diffusion Transformer (DiT) for generalizable HOI generation. Extensive experiments demonstrate that **A^{3}fford-HOI** displays superior generation quality and contact plausibility for fine-grained and generalizable hand object interaction.

1 Introduction

Hand Object Interaction (HOI) generation aims to synthesize realistic and interactive hand-object motion sequences. In practice, HOI has broad applications in various fields, including Virtual Reality (VR) [Zhang and et al., 2024; Fang *et al.*, 2023], Embodied AI [Wan *et al.*, 2023], and robotics [Wang and et al., 2024]. Conventionally, HOI generation demands not merely visual realism, but also a grounding in physically plausible contact and interaction dynamics.

Recently, instruction-driven methods [Cha and et al., 2024; Huang *et al.*, 2025; Zhang *et al.*, 2025] have provided feasible paradigms for HOI generation by mapping natural language to 3D motion spaces. However, they remain circumscribed by monotonous interaction patterns and limited object categories. Researchers [Christen *et al.*, 2024; Huang *et al.*, 2025] primarily focus on kinematic motion synthesis, ignoring the explicit contact constraints, inevitably resulting in unnatural artifacts and physical implausibility. Only a few approaches [Cha and et al., 2024; Zhang *et al.*, 2025] incorporate coarse contact information to guide generation. Specifically, they leverage interaction cues as guidance by calculating the contact probabilities between hand and object surfaces. However, they overlook the distinct physical roles of hand anatomical structures. Crucially, different hand parts — such as fingertips for precision and palms for stability — must be aligned with specific disentangled affordance regions on the object to guarantee functionally valid manipulation. As illustrated in Figure 2, different colors on the hand denote the contact regions of distinct anatomical structures, visually revealing the specific functional contribution to the interaction. Notably, due to the diversity of manipulation modes, each anatomical part may exhibit high contact probabilities across multiple distinct regions of the object surface. While pursuing fine-grained accuracy, generalization is also a more critical challenge. When confronted with rare object shapes and complex interaction patterns, existing methods [Cha and et al., 2024; Christen *et al.*, 2024; Huang *et al.*, 2025] exhibit significant performance degradation, due to the lack of external physical knowledge and geometric reasoning capabilities. Moreover, current datasets are limited in object diversity and manipulation patterns, which is a main reason for limiting the model generalization.

Thus, the HOI generation task still encounters two main

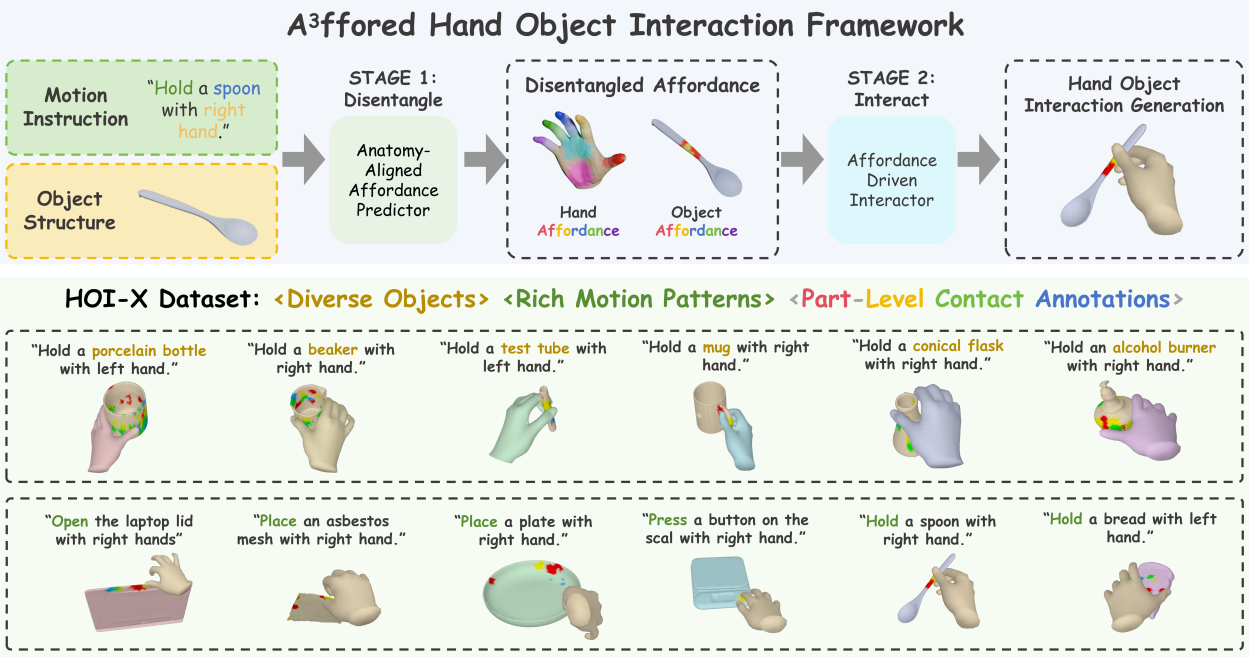

Figure 1: The proposed **A^{3}fford-HOI** framework for fine-grained and generalizable hand object interaction. **A^{3}fford-HOI** mainly consist of two stages, Disentangle and Interact. Disentangle realizes anatomy-aligned affordance prediction with text instruction and object structure as inputs. Interact integrates disentangled affordance as principal condition to guide the final interaction generation. **HOI-X** is the specifically collected hand object interaction dataset with diverse objects, rich motion patterns, and part-level contact annotations.

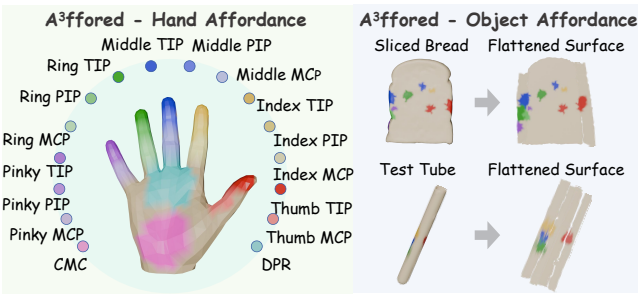

Figure 2: Visualization of Anatomy-Aligned Affordance. Based on anatomical structures, we partition the hand into 16 distinct regions, distinguished by unique colors. The same color area between hand and object represents the high probability of contact.

challenges: 1) Building an extensive HOI dataset, including diverse objects and rich motion patterns. 2) Modeling fine-grained and generalizable hand-object interaction motion following physically plausible contact and dynamics.

To address the data scarcity, we present **HOI-X**, an integrated comprehensive HOI dataset. By consolidating several public HOI datasets [Kwon *et al.*, 2021; Taheri *et al.*, 2020; Zhan and et al., 2024], our HOI-X contains various object categories and rich motion patterns. The rigorous pre-processing is then performed to unify data formats across different data sources. Significantly, we conduct part-level contact annotations (*i.e.*, fine-grained quantification of object affordance) by calculating contact probabilities between specific hand anatomical structures and object surfaces, as shown in Figure 1. Along with HOI-X, we propose **A^{3}fford-HOI**, a novel two-stage framework that enables fine-grained and generalizable HOI motion generation. We formulate the generation process as a **Disentangle-2-Interact** paradigm. To decouple the complexities of interaction, in the **Disentangle** stage, we devise an **Anatomy-Aligned Affordance Predictor**. Specifically, we prompt 3D MLLMs with textual instructions and object geometries to extract latent physical commonsense—external knowledge that is often scarce in limited HOI datasets. Here, the predictor effectively transforms high-level semantic knowledge into fine-grained, part-specific functional cues. In this manner, the textual instructions and object geometries are adaptively transformed into spatially explicit and disentangled affordance maps. Subsequently, these disentangled affordance priors are seamlessly integrated into the **Interact** stage to guide the interaction synthesis. In particular, we present the **Affordance-Driven Interactor**, where textual instructions and object geometric features are fused to provide the global condition for coarse motion generation. Once we obtain initial results, we employ the MLLM-enhanced affordance map as an interactive guidance to refine local hand-object interactions, ensuring high-fidelity contact plausibility. The modulated features are finally fed into the diffusion-based transformer backbone to produce harmonic and physically plausible HOI sequences.

Overall, our contributions are summarized as follows:

- We present HOI-X, a comprehensive hand object interaction dataset with diverse objects, rich motion patterns, and part-level contact annotations for fine-grained interaction.
- We propose A^{3}fford-HOI, a novel two-stage framework

Datasets	#Frames	#Sequences	#Objects	#Motion Patterns	Label Method	Caption	Affordance
GRAB [Taheri *et al.*, 2020]$_{ECCV}$	1.62M	1.33K	51	276	Mocap	✗	✗
HO3D [Hampali and et al., 2020]$_{CVPR}$	0.08M	0.14K	10	92	Auto	✗	✗
H2O [Kwon *et al.*, 2021]$_{ICCV}$	0.57M	1K	20	68	Auto	✗	✗
DexYCB [Chao *et al.*, 2021]$_{CVPR}$	0.57M	0.67K	8	243	Mocap	✗	✗
ARCTIC [Fan *et al.*, 2023]$_{CVPR}$	2.1M	4.6K	11	161	Mocap	✗	✗
OakInk2 [Zhan and et al., 2024]$_{CVPR}$	4.01M	2.2K	75	254	Mocap	✓	✗
HOI-X [Ours 2026]	**6.2M**	**5.1K**	**134**	**598**	**Semi-Auto**	✓	**Disentangled**

Table 1: Statistical comparison of our **HOI-X** dataset against various counterparts. HOI-X contains diverse objects, rich motion patterns and part-level contact annotations.

that enables fine-grained and generalizable hand object interaction generation.

- We design an Anatomy-Aligned Affordance Predictor that injects external physical priors provided by 3D MLLMs for disentangled affordance prediction. Furthermore, we present an Affordance-Driven Interactor, integrating MLLM-enhanced affordance priors to ensure fine-grained and generalizable interaction synthesis.
- Extensive experiments demonstrate that A^{3}fford-HOI achieves SOTA performances, producing highly realistic and functionally accurate hand-object interaction on fine-grained and generalizable scenarios.

2 Related Work

Hand Object Interaction Generation: Research in 3D Hand-Object Interaction (HOI) can be broadly categorized into two primary streams: reconstruction and generation. While reconstruction methods [Tzionas and Gall, 2015; Cao *et al.*, 2021; Zhang *et al.*, 2019; Hu *et al.*, 2022; Liu *et al.*, 2025] typically focus on recovering 3D hand-object dynamics from visual observations such as RGB videos or depth maps, the core objective of HOI generation is to synthesize realistic interaction sequences from multi-modal inputs. Early generation approaches [Ghosh *et al.*, 2023; Brahmbhatt *et al.*, 2019; Corona and et al., 2020] predominantly concentrated on synthesizing static grasping postures or simple manipulation primitives. Propelled by the advancements in Diffusion Models and Large Language Models (LLMs), recent frameworks like DiffH2O [Christen *et al.*, 2024] and HOIGPT [Huang *et al.*, 2025] have harnessed textual guidance to facilitate diverse interaction synthesis. Building on this, Text2HOI [Cha and et al., 2024] and OpenHOI [Zhang *et al.*, 2025] explicitly incorporate interaction cues to further refine generation quality. However, existing methodologies largely overlook the distinct functional roles of different hand anatomical structures during interaction. In addition, existing datasets [Fan *et al.*, 2023; Kwon *et al.*, 2021; Taheri *et al.*, 2020; Zhan and et al., 2024] are often limited by repetitive interaction patterns and a scarce variety of object shapes. To address these limitations, we introduce HOI-X, a comprehensive dataset enriched with diverse objects and motions, and propose the A^{3}fford-HOI framework, aiming to achieve fine-grained and generalizable HOI generation.

Affordance Learning: The genesis of affordance learning lies in the image domain, with pioneering works [Do *et al.*, 2018; Zhao *et al.*, 2020] primarily targeting the detection of functional objects. Subsequent advancements [Lu *et al.*, 2022; Mi *et al.*, 2020] integrate linguistic context to refine this process. Nevertheless, these approaches remain constrained to coarse, object-level predictions. To overcome this limitation, contemporary research [Fang *et al.*, 2018; Luo *et al.*, 2022; Nagarajan *et al.*, 2019] shifts towards specific affordance parts, establishing a new norm for precision in the field. With the rise of Embodied AI, affordance learning has rapidly expanded into the 3D domain. 3D AffordanceNet [Deng *et al.*, 2021] constructs the first dataset for affordance detection in 3D point clouds. Building on this, works like LASO [Li *et al.*, 2024] and AffordanceLLM [Chu *et al.*, 2025] establish a bridge between the powerful natural language capabilities of LLMs and 3D object affordances, significantly enhancing static 3D understanding. Further advancements, such as SeqAfford [Wang *et al.*, 2024] and OpenAfford [Wu *et al.*, 2025], refine and expand this capability, realizing compositional affordance reasoning. However, these approaches remain confined to fundamental perception tasks, leaving the application of affordance insights to hand-object interaction generation unexplored. In contrast, our work introduces a "Disentangle-2-Interact" paradigm. Through this framework, we seamlessly integrate disentangled affordance into the HOI task, achieving superior fine-grained and generalizable motion synthesis.

3 HOI-X Dataset

To alleviate the limitations of object diversity and motion patterns in current datasets, we introduce **HOI-X**, a comprehensive hand object interaction dataset with Anatomy-Aligned Affordance (A^{3}fford) annotations, which consists of 5.1K high-quality motion sequences, with a total of 6.2M frames and 134 object templates.

We devise a rigorous data processing pipeline to integrate and canonicalize interaction data from several sources: GRAB [Taheri *et al.*, 2020], H2O [Kwon *et al.*, 2021], and OakInk2 [Zhan and et al., 2024]. The pipeline consists of two key phases. **1) Unified Spatial Projection.** To eliminate spatial discrepancies caused by different capture systems, all hand-object interaction sequences are projected into a unified world coordinate system and normalized to a standard scale.

2) Object Representation Standardization. To facilitate efficient geometric learning, we perform uniform sampling on the diverse object meshes, converting them into standardized point cloud models with a fixed resolution of 1,024 points. The specific data statistics and comparisons are shown in Table 1.

4 Proposed Method

4.1 Problem Formulation

Given the multi-modal inputs of text descriptions $text$ and object point clouds P_o, the goal of our framework is to generate realistic and interactive hand-object motion sequences M. In practice, we leverage the parameterized hand model MANO [Romero and et al., 2022] to represent hand poses. We tackle this challenging problem with the designed framework **A^{3}fford-HOI**. The overall objective can be expressed as:

$$M = \mathrm{A^3ffordHOI}(text, P_o) \tag{1}$$

4.2 Disentangle-2-Interact Paradigm

Drawing from classical grasp taxonomy theory [Cutkosky and others, 1989], the anatomical structures of the hand assume distinct functional roles during interaction, governed by dual constraints of task requirements and object geometry. Inspired by these theoretical insights, we propose **Anatomy-Aligned Affordance (A^{3}fford)** to explicitly model disentangled part-specific contact potentials for fine-grained interaction synthesis. To implement this guidance, we introduce a **Disentangle-2-Interact** paradigm. As shown in Figure 3, we utilize the anatomy disentangled affordance, augmenting with external knowledge from MLLMs to guide the fine-grained and generalizable interactive motion generation.

Disentangle Stage: A^{3}fford Predictor To harness the advanced reasoning capabilities of large models for part-level contact prior prediction, we build an A^{3}fford Predictor upon the 3D MLLM to enhance fine-grained and external physical knowledge. The backbone integrates a 3D Vision Encoder (ReCon++ [Qi and et al., 2023]) for geometric object embedding, a pre-trained text encoder (CLIP [Sanghi and et al., 2022]) for text embedding, and a generic Large Language Model (LLaMA [Touvron and et al., 2023]) for affordance prediction. The multi-modal input features include representations of object structure (point features F_{point}) , text prompt (prompt features F_{prompt}), and motion instruction (instruction features F_{ins}), which are then transformed into hybrid latent tokens H via LLM. Thus, the LLM inference process can be formulated as:

$$H = LLM(F_{point}, F_{prompt}, F_{ins}), \tag{2}$$

where H is the hybrid latent tokens and can also represent the textual tokens H_{text}, constrained by the annotated answer text in the following LLM tuning.

Meanwhile, to bridge the gap between textual semantics and geometric interactions in LLM, we expand the LLM vocabulary with two specialized tokens: `<Left/Right Hand Affordance>` and `<Object Affordance>`, as shown in Figure 3 (left), serving as learnable instructors for affordance disentanglement. Thus, the hand mask tokens H_{hand} and object affordance tokens H_{object} are separated from the hybrid latent tokens H_{text}/H. Finally, H_{text}, H_{hand}, H_{object} are then fed into the specific Text Decoder, Hand Mask Decoder, and Object Affordance Decoder for subsequent features extraction, respectively.

The textual tokens are decoded into the answer text by Text Decoder (TD), which serves as a semantic constraint to ensure the synthesized interaction affordance aligns with the linguistic instructions.

$$T_{answer} = TD(H_{text}), \tag{3}$$

where T_{answer} is the answer text. To acquire the disentangled contact priors, the Hand Mask Decoder (HMD) projects hand mask tokens into active and anatomical hand contact affordance. Simultaneously, the Object Affordance Decoder (OAD) projects object affordance tokens into the geometric object affordance.

$$A_{hand} = HMD(H_{hand}), \tag{4}$$

$$A_{object} = OAD(H_{object}), \tag{5}$$

where A_{hand} is the predicted hand affordance, A_{object} is the predicted object affordance. We employ the LoRA techniques [Hu and et al., 2022] to efficiently fine-tune the overall large model constrained by the annotated answer text. The specific decoders are trained with the supervision of annotated affordances.

Therefore, the A^{3}fford Predictor effectively transforms high-level interaction intent into spatially explicit, fine-grained and disentangled affordance maps, ensuring strict alignment between anatomical hand structures and their corresponding functional regions on the object. Meanwhile, the introduction of large model provides external knowledge cues for the next Interact stage, improving both the fine-grained fidelity of the interaction and the generalization ability across diverse object categories.

Interact Stage: Affordance Driven Interactor In the Interact stage, we employ the anatomy-aligned affordance from A^{3}fford Predictor to facilitate high-fidelity and physically plausible hand-object interaction synthesis. The generation process is fundamentally guided by two-level conditions: holistic semantic constraints and the affordance features acquired from Affordance Encoder. The holistic constraints are composed of motion instruction feature, textural prompt feature and object structure feature. These global cues are integrated into our backbone via multi-head attention, which can generate rough motion trajectory and hand pose. Though the above holistic conditions determine the overall motion trend, they can not guarantee precise finger-level contacts without the contact priors. Hence, we introduce the Affordance Control mechanism that employs the aforementioned fine-grained affordance priors to explicitly guide the interaction details.

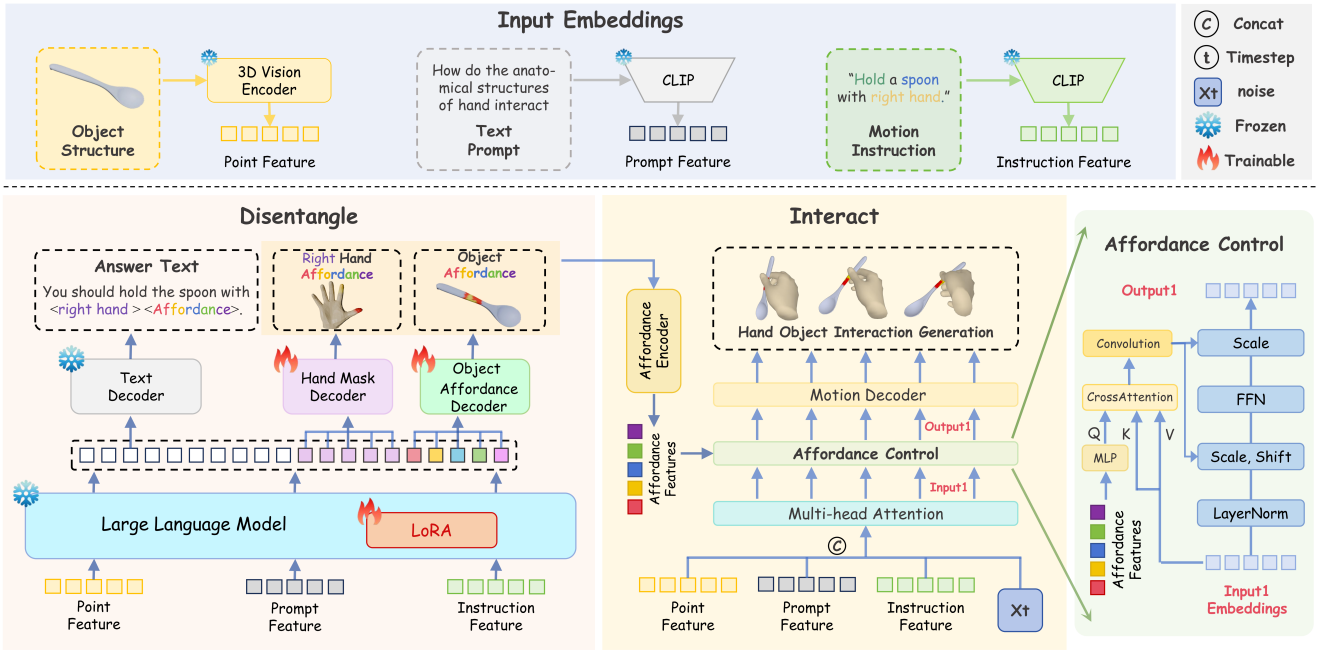

Figure 3: A^{3}fford-HOI framework: We conduct a Disentangle-2-Interact paradigm. In the Disentangle stage, we utilize the external physical priors in 3D MLLMs for anatomy-aligned affordance prediction. In the Interact stage, we integrate the motion instructions, object structures and external knowledge enhanced affordance to generate HOI motions.

Affordance Control . As illustrated in Figure 3 (bottom right corner), we utilize the affordance features as the query Q to match the motion embeddings as key $K$and value V, through the cross-attention mechanism. Then we obtain the affordance-guided motion features via the AdaLN [Huang and et al., 2017] layers. The modulated motion features are derived as:

$$F_{motion} = \text{AdaLN}(F_{motion}, \text{Conv}(F_{aff})), \tag{6}$$

where F_{motion} donates the interaction motion features and F_{aff} indicates the affordance features. Ultimately, the Motion Decoder translates these features into the final fine-grained and physically plausible HOI sequences.

4.3 Objective Function

In the **Disentangle stage**, to effectively fit the complex distribution of anatomy-aligned affordances, we employ the Channel-Wise Tversky Loss ($\mathcal{L}_{Tversky}$) [Salehi and et al., 2017]. Formally, we calculate the weighted overlap between the predicted probability map P and the ground truth G as follows:

$$\mathcal{L}_{Tversky} = \frac{PG}{PG + P(1-G) + (1-P)G}. \tag{7}$$

Similar to [Jiang and et al., 2021], we optimize the **Interact stage** using the composition of simple loss L_{simple}, contact loss $L_{contact}$, and penetration loss L_{penet} to generate the high-fidelity interaction sequences that are both geometrically precise and physically plausible. The overall objective function is expressed as:

$$L_{total} = L_{simple} + \lambda_c L_{contact} + \lambda_p L_{penet}, \tag{8}$$

where λ_c, λ_p are the trade-off weight coefficient.

5 Experiments

5.1 Experimental Settings

Implementation Details: In our experiments, we set the generated motion length to 200 with FPS = 30. We employ a frozen CLIP-ViT-B-32 model as the text encoder. To train the disentangle stage, the batchsize is set to 128, and the epoch number is 100. The interact stage is trained for 600 epochs with a batchsize as 64. The learning rate of the disentangle stage is 1e-3, and weight decay is 2e-4. The interact stage is trained with 5e-5 learning rate within 2e-6 weight decay. All experiments are deployed on an NVIDIA A100 GPU for 24 hours. The trade-off weight coefficients λ_c, λ_p are set to 20 and 10, respectively.

Evaluation Metrics: To comprehensively evaluate the quality, diversity, and realism of our generated interaction motions, we utilize a set of quantitative metrics inspired by [Ghosh *et al.*, 2023]. We roughly divide the 6 evaluation indicators into three categories:

- **Motion Accuracy:** Average Variance Error (**AVE**) measures the variance error between the joint positions. Fréchet Inception Distance (**FID**) quantifies the distributional discrepancy between the synthesized results and the GT.
- **Diversity:** **Diversity** and multi-modality (**MModality**) measure the variability across different prompts.
- **Interaction Quality:** Contact percentage(**C%**) calculates the percentage of frames detected with contact. **Penetra-**

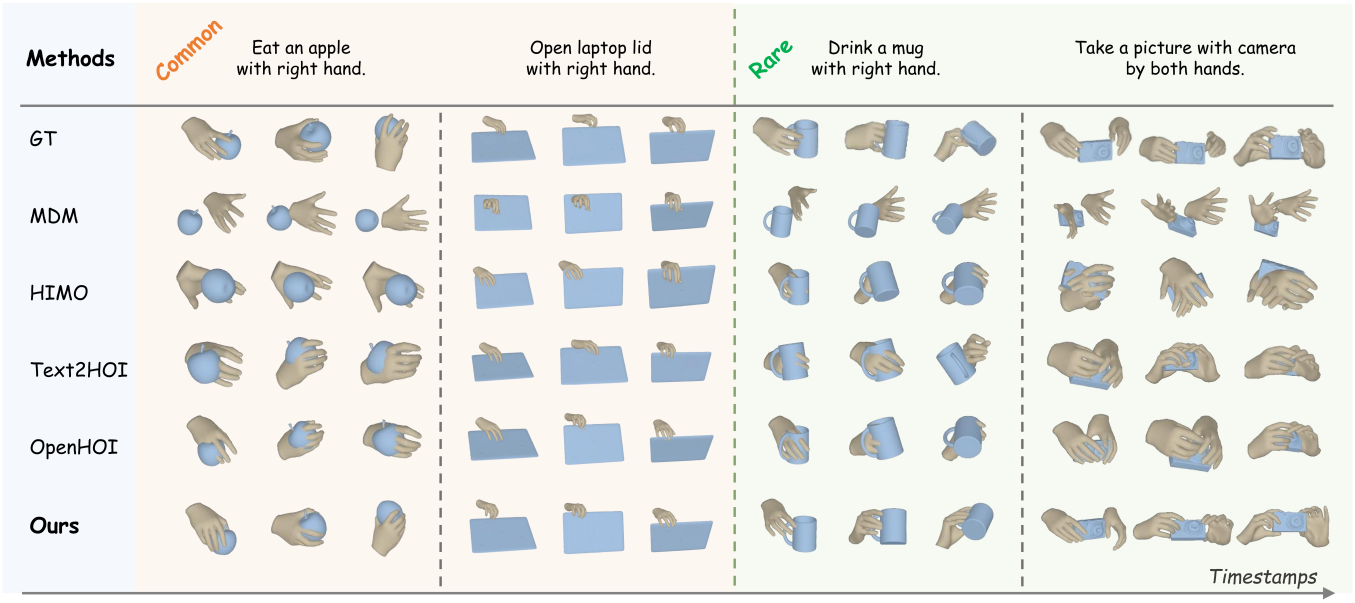

Figure 4: **Visual comparisons on HOI-X.** We compare the predicted hand object interaction visualization of our framework with other SOTA methods. **Common** and **Rare** represent objects with different degrees of scarcity.

Methods	FID ↓	AVE ↓	Diversity →	MModality ↑	C% ↑	Penetrarion ↓
GT	-	-	$1.573^{\pm.008}$	$0.142^{\pm.002}$	$0.549^{\pm.001}$	$0.175^{\pm.003}$
MDM [Tevet and et al., 2022]$_{ICLR}$	$4.901^{\pm.106}$	$0.207^{\pm.025}$	$0.948^{\pm.082}$	$0.032^{\pm.008}$	$0.295^{\pm.013}$	$0.408^{\pm.016}$
HIMO [Lv and et al., 2024]$_{ECCV}$	$3.325^{\pm.023}$	$0.114^{\pm.004}$	$0.952^{\pm.037}$	$0.048^{\pm.005}$	$0.313^{\pm.008}$	$0.388^{\pm.010}$
Text2HOI [Cha and et al., 2024]$_{CVPR}$	$2.391^{\pm.031}$	$0.098^{\pm.001}$	$0.976^{\pm.041}$	$0.045^{\pm.008}$	$0.421^{\pm.007}$	$0.412^{\pm.019}$
OpenHOI [Zhang *et al.*, 2025]$_{NeurIPS}$	$2.448^{\pm.014}$	$0.097^{\pm.002}$	$1.005^{\pm.049}$	$0.044^{\pm.005}$	$0.397^{\pm.004}$	$0.322^{\pm.009}$
A^{3}ffordance-HOI (Ours)	$\mathbf{1.995^{\pm.014}}$	$\mathbf{0.093^{\pm.001}}$	$\mathbf{1.139^{\pm.042}}$	$\mathbf{0.069^{\pm.006}}$	$\mathbf{0.481^{\pm.006}}$	$\mathbf{0.261^{\pm.009}}$

Table 2: Comparison of our A^{3}fford-HOI with the state-of-the-art methods on our HOI-X dataset in the **common** scenarios. ↑ donates the higher the better, ↓ means the lower the better, and →indicates that closer to GT is better. ± indicates the 95 % confidence interval.

tion quantifies physical implausibility by measuring the extent to which the hand mesh intersects with the object mesh.

5.2 Quantitative Evaluation

Comparing to SOTA Methods: We compare A^{3}fford-HOI with four state-of-the-art approaches, MDM [Tevet and et al., 2022], HIMO [Lv and et al., 2024], Text2HOI [Cha and et al., 2024], and OpenHOI [Zhang *et al.*, 2025]. These methods are retrained and tested on the HOI-X dataset with the open-source code released by the authors. In particular, to evaluate the generalization capability of our model, we divide the test data into two categories: **common** objects and **rare** objects. The details of the dataset splitting are provided in the Supplementary Materials. The results in Table 2 indicate that our model outperforms the competitive methods on all the metrics in common objects, with the most notable lead observed in Interaction Quality. By leveraging anatomy-aligned affordance, our model employs the Affordance Control to seamlessly align interactive motions with fine-grained contact information, successfully enhancing the synthesis quality of HOI motions. Besides, in terms of scarce-category objects as depicted in Table 3, other methods suffer a notable decline in FID, C%, and Penetration scores. In contrast, our approach maintains strong performance in both motion accuracy and fine-grained interaction quality. This is attributed to our disentangle-to-interact paradigm, which enhances generalization by predicting object affordances that are fused with external physical priors.

Ablation Study: To evaluate the Disentangle-2-Iteract paradigm, we conduct an ablation study with different key components. Considering the fair comparison, our ablation study is conducted on the whole testing set, directly. For verifying the effects of the Disentangle stage, we conduct ablation studies on A^{3}fford Predictor. As shown in Table 4, removing the A^{3}fford Predictor leads to a severe drop in both contact ratio (C%) and penetration, which reveals that without external physical priors enhanced affordance (*i.e.*LLM tuning), the model can not generate reasonable hand-object contact. When we employ simple affordance without anatomy-aligned distinction, interaction quality decrease significantly, indicating the anatomy-aligned affordance plays a crucial role in fine-grained control. To demonstrate the effectiveness of the Interact stage, we further conduct an ablation by removing the

Methods	FID ↓	AVE ↓	Diversity →	MModality ↑	C% ↑	Penetrarion ↓
Ground Truth	-	-	$1.566^{\pm.008}$	$0.141^{\pm.002}$	$0.552^{\pm.001}$	$0.177^{\pm.003}$
MDM [Tevet and et al., 2022]$_{ICLR}$	$5.391^{\pm.064}$	$0.089^{\pm.013}$	$0.944^{\pm.096}$	$0.031^{\pm.014}$	$0.191^{\pm.021}$	$0.382^{\pm.028}$
HIMO [Lv and et al., 2024]$_{ECCV}$	$3.477^{\pm.030}$	$0.117^{\pm.005}$	$0.935^{\pm.044}$	$0.034^{\pm.006}$	$0.284^{\pm.008}$	$0.382^{\pm.011}$
Text2HOI [Cha and et al., 2024]$_{CVPR}$	$2.709^{\pm.033}$	$0.106^{\pm.001}$	$0.960^{\pm.039}$	$0.029^{\pm.006}$	$0.347^{\pm.008}$	$0.455^{\pm.024}$
OpenHOI [Zhang *et al.*, 2025]$_{NeurIPS}$	$2.682^{\pm.017}$	$0.104^{\pm.004}$	$0.905^{\pm.055}$	$0.024^{\pm.006}$	$0.304^{\pm.001}$	$0.418^{\pm.007}$
A^{3}ffordance-HOI (Ours)	$\mathbf{2.051^{\pm.0020}}$	$\mathbf{0.099^{\pm.005}}$	$\mathbf{1.139^{\pm.055}}$	$\mathbf{0.066^{\pm.002}}$	$\mathbf{0.467^{\pm.006}}$	$\mathbf{0.285^{\pm.018}}$

Table 3: Comparison on HOI-X dataset in the **rare** scenarios. Our method achieves a substantial improvement in **generalization**.

Methods	FID ↓	AVE ↓	Diversity →	MModality ↑	C% ↑	Penetration ↓
Ground Truth	-	-	$1.573^{\pm.008}$	$0.142^{\pm.002}$	$0.549^{\pm.001}$	$0.175^{\pm.003}$
w/o affordance	$2.432^{\pm.013}$	$0.098^{\pm.002}$	$0.981^{\pm.052}$	$0.044^{\pm.001}$	$0.351^{\pm.008}$	$0.414^{\pm.017}$
w/o anatomy-aligned	$2.389^{\pm.021}$	$0.096^{\pm.002}$	$0.984^{\pm.058}$	$0.045^{\pm.005}$	$0.424^{\pm.004}$	$0.389^{\pm.021}$
w/o control	$2.397^{\pm.011}$	$0.097^{\pm.003}$	$1.084^{\pm.047}$	$0.058^{\pm.004}$	$0.453^{\pm.007}$	$0.305^{\pm.011}$
A^{3}ffordance-HOI (Ours)	$\mathbf{1.997^{\pm.016}}$	$\mathbf{0.094^{\pm.001}}$	$\mathbf{1.139^{\pm.036}}$	$\mathbf{0.069^{\pm.006}}$	$\mathbf{0.4814^{\pm.008}}$	$\mathbf{0.275^{\pm.014}}$

Table 4: Ablation study on HOI-X. "w/o affordance" indicates we remove the A^{3}fford Predictor and do not use affordance in the Interact Stage. "w/o anatomy-aligned" means we do not disentangle the hand into distinct components for affordance prediction. "w/o control" indicates that we concatenate affordance, input embeddings, and noise without the Affordance Control module.

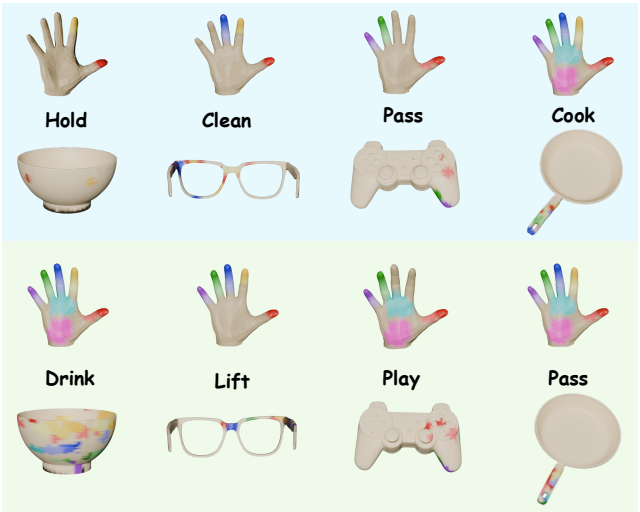

Figure 5: Visualization of various anatomy-aligned affordances. Given the same object (in each column), varying motion patterns result in different predictions of contact information (affordances) on both hand and object, which indicate the fine-grained interaction.

Affordance Control module. In this impact, the interaction motions display misalignment with the affordance guidance, leading to the inaccurate contact pattern.

5.3 Qualitative Evaluation

To demonstrate the effectiveness of our method, we present the visualized comparison between A^{3}fford-HOI and other works with the same motion patterns and object templates, as shown in Figure 4. A^{3}fford-HOI synthesizes superior, natural, and physically plausible HOI sequences. Other approaches are limited in their ability to capture sufficiently detailed contact information, which leads to unnatural interaction motions and severe penetration. Especially, in scenarios with scarce objects, our approach demonstrates stronger generalization capabilities. The key capability lies in generating fine-grained, physically plausible interactions without being confined to specific object shapes.

Additionally, we display various disentangled affordances across **diverse objects** and **motion patterns** corresponding to different actions, as shown in Figure 5. The LLM-enhanced external knowledge embedded within the affordance provides a strong prior and guidance for our method to generate fine-grained and generalizable interactive motions. Moreover, we conduct a user study to assess the quality of the generated results from the perspective of the subjective metrics. The details are shown in the Supplementary Materials.

6 Conclusion

In this paper, we propose A^{3}fford-HOI, a fine-grained and generalizable Hand Object Interaction generation framework with Anatomy-Aligned Affordance Disentanglement. Meanwhile, we introduce HOI-X, a comprehensive HOI dataset which enriches object diversity, motion patterns, and fine-grained contact information. Specifically, we propose a Disentangle-2-Interact Paradigm, including an A^{3}fford Predictor to acquire disentangled affordance and the Affordance Control in the interact stage to guide HOI synthesis. Finally, our method achieves SOTA performances with physically plausible, fine-grained and generalizable HOI sequences. In the future, we will focus on interaction tasks in more complex scenarios, such as simultaneously generation for hand and multi-object interaction.

Acknowledgements

This work is supported by NSFC No. 62306309, 62276031 and NKRDPC No. 2023YFF0904700.

Contribution Statement

Xurui Hu conducted the algorithm design, experiments, and paper writing. Xingqun Qi served as the project leader. Muyi Sun served as the corresponding author and provided research guidance. All authors contributed to the discussion and approved the final manuscript.

References

[Brahmbhatt *et al.*, 2019] Samarth Brahmbhatt, Ankur Handa, and Hays et al. Contactgrasp: Functional multi-finger grasp synthesis from contact. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2386–2393. IEEE, 2019.

[Cao *et al.*, 2021] Zhe Cao, Ilija Radosavovic, and Kanazawa et al. Reconstructing hand-object interactions in the wild. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12417–12426, 2021.

[Cha and et al., 2024] Junuk Cha and Kim et al. Text2hoi: Text-guided 3d motion generation for hand-object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1577–1585, 2024.

[Chao *et al.*, 2021] Yu-Wei Chao, Wei Yang, and Xian et al. Dexycb: A benchmark for capturing hand grasping of objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9044–9053, 2021.

[Christen *et al.*, 2024] Sammy Christen, Shreyas Hampali, and Sener et al. Diffh2o: Diffusion-based synthesis of hand-object interactions from textual descriptions. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–11, 2024.

[Chu *et al.*, 2025] Hengshuo Chu, Xiang Deng, and Lv et al. 3d-affordancellm: Harnessing large language models for open-vocabulary affordance detection in 3d worlds. *arXiv preprint arXiv:2502.20041*, 2025.

[Corona and et al., 2020] Enric Corona and Pumarola et al. Ganhand: Predicting human grasp affordances in multi-object scenes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5031–5041, 2020.

[Cutkosky and others, 1989] Mark R Cutkosky et al. On grasp choice, grasp models, and the design of hands for manufacturing tasks. *IEEE Transactions on robotics and automation*, 5(3):269–279, 1989.

[Deng *et al.*, 2021] Shengheng Deng, Xun Xu, and Wu et al. 3d affordancenet: A benchmark for visual object affordance understanding. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1778–1787, 2021.

[Do *et al.*, 2018] Thanh-Toan Do, Anh Nguyen, and Ian Reid. Affordancenet: An end-to-end deep learning approach for object affordance detection. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 5882–5889. IEEE, 2018.

[Fan *et al.*, 2023] Zicong Fan, Omid Taheri, and Tzionas et al. Arctic: A dataset for dexterous bimanual hand-object manipulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12943–12954, 2023.

[Fang *et al.*, 2018] Kuan Fang, Te-Lin Wu, and Yang et al. Demo2vec: Reasoning object affordances from online videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2139–2147, 2018.

[Fang *et al.*, 2023] Hao-Shu Fang, Chenxi Wang, and Fang et al. Anygrasp: Robust and efficient grasp perception in spatial and temporal domains. *IEEE Transactions on Robotics*, 39(5):3929–3945, 2023.

[Ghosh *et al.*, 2023] Anindita Ghosh, Rishabh Dabral, and Golyanik et al. Imos: Intent-driven full-body motion synthesis for human-object interactions. In *Computer Graphics Forum*, volume 42, pages 1–12. Wiley Online Library, 2023.

[Hampali and et al., 2020] Shreyas Hampali and Rad et al. Honnotate: A method for 3d annotation of hand and object poses. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3196–3206, 2020.

[Hu and et al., 2022] Edward J Hu and Shen et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

[Hu *et al.*, 2022] Haoyu Hu, Xinyu Yi, Hao Zhang, Jun-Hai Yong, and Feng Xu. Physical interaction: Reconstructing hand-object interactions with physics. In *SIGGRAPH Asia 2022 conference papers*, pages 1–9, 2022.

[Huang and et al., 2017] Xun Huang and Belongie et al. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pages 1501–1510, 2017.

[Huang *et al.*, 2025] Mingzhen Huang, Fu-Jen Chu, and Tekin et al. Hoigpt: Learning long-sequence hand-object interaction with language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7136–7146, 2025.

[Jiang and et al., 2021] Hanwen Jiang and Liu et al. Hand-object contact consistency reasoning for human grasps generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11107–11116, 2021.

[Kwon *et al.*, 2021] Taein Kwon, Bugra Tekin, and Stühmer et al. H2o: Two hands manipulating objects for first person interaction recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10138–10148, 2021.

[Li *et al.*, 2024] Yicong Li, Na Zhao, and Xiao et al. Laso: Language-guided affordance segmentation on 3d object. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14251–14260, 2024.

[Liu *et al.*, 2025] Yumeng Liu, Xiaoxiao Long, and Yang et al. Easyhoi: Unleashing the power of large models for reconstructing hand-object interactions in the wild. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7037–7047, 2025.

[Lu *et al.*, 2022] Liangsheng Lu, Wei Zhai, and Luo et al. Phrase-based affordance detection via cyclic bilateral interaction. *IEEE Transactions on Artificial Intelligence*, 4(5):1186–1198, 2022.

[Luo *et al.*, 2022] Hongchen Luo, Wei Zhai, and Zhang et al. Learning affordance grounding from exocentric images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2252–2261, 2022.

[Lv and et al., 2024] Xintao Lv and Xu et al. Himo: A new benchmark for full-body human interacting with multiple objects. In *European Conference on Computer Vision*, pages 300–318. Springer, 2024.

[Mi *et al.*, 2020] Jinpeng Mi, Hongzhuo Liang, Nikolaos Katsakis, Song Tang, Qingdu Li, Changshui Zhang, and Jianwei Zhang. Intention-related natural language grounding via object affordance detection and intention semantic extraction. *Frontiers in Neurorobotics*, 14:26, 2020.

[Nagarajan *et al.*, 2019] Tushar Nagarajan, Christoph Feichtenhofer, and Kristen Grauman. Grounded human-object interaction hotspots from video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8688–8697, 2019.

[Qi and et al., 2023] Zekun Qi and Dong et al. Contrast with reconstruct: Contrastive 3d representation learning guided by generative pretraining. In *International Conference on Machine Learning*, pages 28223–28243. PMLR, 2023.

[Romero and et al., 2022] Javier Romero and Tzionas et al. Embodied hands: Modeling and capturing hands and bodies together. *arXiv preprint arXiv:2201.02610*, 2022.

[Salehi and et al., 2017] Seyed Sadegh Mohseni Salehi and Erdogmus et al. Tversky loss function for image segmentation using 3d fully convolutional deep networks. In *International workshop on machine learning in medical imaging*, pages 379–387. Springer, 2017.

[Sanghi and et al., 2022] Aditya Sanghi and Chu et al. Clip-forge: Towards zero-shot text-to-shape generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18603–18613, 2022.

[Taheri *et al.*, 2020] Omid Taheri, Nima Ghorbani, and Black et al. Grab: A dataset of whole-body human grasping of objects. In *European conference on computer vision*, pages 581–600. Springer, 2020.

[Tevet and et al., 2022] Guy Tevet and Raab et al. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*, 2022.

[Touvron and et al., 2023] Hugo Touvron and Lavril et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[Tzionas and Gall, 2015] Dimitrios Tzionas and Juergen Gall. 3d object reconstruction from hand-object interactions. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 729–737, 2015.

[Wan *et al.*, 2023] Weikang Wan, Haoran Geng, and Liu et al. Unidexgrasp++: Improving dexterous grasping policy learning via geometry-aware curriculum and iterative generalist-specialist learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3891–3902, 2023.

[Wang and et al., 2024] Chen Wang and Shi et al. Dexcap: Scalable and portable mocap data collection system for dexterous manipulation. *arXiv preprint arXiv:2403.07788*, 2024.

[Wang *et al.*, 2024] Hanqing Wang, Chunlin Yu, and Luo et al. Seqafford: Sequential 3d affordance reasoning via multimodal large language model. 2024.

[Wu *et al.*, 2025] Lin Wu, Wei Wei, and Yu et al. Open-vocabulary 3d affordance understanding via functional text enhancement and multilevel representation alignment. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 7988–7997, 2025.

[Zhan and et al., 2024] Xinyu Zhan and Yang et al. Oakink2: A dataset of bimanual hands-object manipulation in complex task completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 445–456, 2024.

[Zhang and et al., 2024] Hui Zhang and Christen et al. Graspxl: Generating grasping motions for diverse objects at scale. In *European Conference on Computer Vision*, pages 386–403. Springer, 2024.

[Zhang *et al.*, 2019] Hao Zhang, Zi-Hao Bo, Jun-Hai Yong, and Feng Xu. Interactionfusion: real-time reconstruction of hand poses and deformable objects in hand-object interactions. *ACM Transactions on Graphics (ToG)*, 38(4):1–11, 2019.

[Zhang *et al.*, 2025] Zhenhao Zhang, Ye Shi, and Yang et al. Openhoi: Open-world hand-object interaction synthesis with multimodal large language model. *arXiv preprint arXiv:2505.18947*, 2025.

[Zhao *et al.*, 2020] Xue Zhao, Yang Cao, and Yu Kang. Object affordance detection with relationship-aware network. *Neural Computing and Applications*, 32(18):14321–14333, 2020.