

OpenHOI: Open-World Hand-Object Interaction Synthesis with Multimodal Large Language Model

Appendix

Appendix:

| | |
|---|----------|
| A Implementation Details | 1 |
| B Instruction Decomposition and Affordance Reasoning via 3D MLLM | 2 |
| B.1 Affordance Reasoning | 2 |
| B.2 Instruction Decomposition | 2 |
| B.3 Diffusion Process | 3 |
| B.4 Loss-guided Physical Refinement | 4 |
| C Additional Experiments | 4 |
| C.1 Results on Extreme-Case: Completely Unseen Datasets | 4 |
| C.2 Evaluation for physical realism | 4 |
| C.3 3D MLLM Fine-tuning | 5 |
| C.4 Sensitivity Analysis | 5 |
| C.5 Ablation Study on Multi <AFF> | 5 |
| C.6 Ablation Study on Motion In-between Refinement | 5 |
| C.7 Visualization on Affordance | 6 |
| C.8 Qualitative results Compare with SOTA | 6 |
| C.9 Statistically Insignificant | 6 |
| D Code and Dataset | 8 |
| D.1 Code | 8 |
| D.2 Dataset | 8 |
| E Discussion | 9 |
| E.1 Generalizability ability in HOI: Affordance as a key | 9 |
| E.2 How to use 3D MLLM in Embodied AI: Choose Powerful Foundation Model and Coarse-to-fine tuning | 10 |
| E.3 Future of Work: Real-World Applications, about AR/VR and Robotics | 10 |

A Implementation Details

3D MLLM. We initialize our model from the ShapeLLM-7B checkpoint, freezing its 3D encoder and augmenting the visual backbone with Uni3D for robust dense 3D prediction, while the projection head is implemented as a shallow MLP, and LoRA is applied to streamline fine-tuning. Training unfolds in two stages: first, we optimize for seven epochs with AdamW (learning rate $2 \cdot 10^{-4}$, zero weight decay) under a cosine-annealing schedule and a 2% linear warm-up; then, we continue for three additional epochs with AdamW (learning rate $5 \cdot 10^{-4}$, zero weight decay) under the same cosine schedule but a 1% warm-up.

Diffusion Model. We employ a $T = 1000$ -step noising process with a cosine noise schedule, and inject positional information at both the frame- and agent-levels using sinusoidal encodings. During sampling, we apply classifier-free guidance by randomly substituting 10% of conditioning inputs with unconditional noise while retaining 90% of the original conditions, and use a guidance scale of 2.5 to steer the denoising trajectory.

B Instruction Decomposition and Affordance Reasoning via 3D MLLM

B.1 Affordance Reasoning

3D Object Point Cloud Encoding. We take as input a point cloud of an object, sampled to N points. The backbone is a ReCon++ [34] network (or a similar architecture) that processes these points and produces per-point feature representations

$$F_{\text{obj}} \in \mathbb{R}^{N \times C},$$

which capture both local geometric details and the overall global context.

Multi-Token Fusion Mechanism. Rather than generating a single $\langle \text{AFF} \rangle$ segmentation token, PixelLM[36] defines, at each visual scale ℓ , a segmentation codebook comprising N learnable $\langle \text{AFF} \rangle$ tokens. After encoding the textual prompt, the model sequentially outputs the N tokens, each associated with a hidden vector h_i^ℓ . A linear projection ϕ then aggregates these vectors into a unified representation

$$h^\ell = \phi(h_1^\ell, \dots, h_N^\ell),$$

which is concatenated with the scale-specific image features and fed into the pixel decoder to produce the final segmentation mask. Experiments on the MUSE validation set indicate that increasing N from 1 to 3 improves cloU, demonstrating that the multi-token fusion mechanism captures more nuanced semantic details and significantly enhances fine-grained segmentation.

B.2 Instruction Decomposition

OpenHOI decomposes a single high-level instruction into an ordered sequence of actionable affordance steps. Each step is marked by a special token and then grounded spatially in the 3D point cloud.

Instruction Text Encoding. We take a natural-language instruction T_{ins} as the model input. The backbone is a LLaMA-style Transformer that produces token-wise hidden states $\{h_t\}_{t=1}^T$ and aggregates them via a pooling operation into a single embedding

$$h_{\text{cls}} \in \mathbb{R}^D,$$

which is then used for downstream tasks.

Segmentation Token Injection. Extend the MLLM’s vocabulary by adding a special marker $\langle \text{AFF} \rangle$, which explicitly denotes the boundary of each sub-task in the generated sequence.

Conditioned Autoregressive Generation. Given the fused 3D point features F_{obj} and the instruction embedding h_{cls} , the Transformer predicts an interleaved stream of action words and $\langle \text{AFF} \rangle$ tokens, for example:

$$\text{Pick} \rightarrow \langle \text{AFF} \rangle \quad \text{Twist} \rightarrow \langle \text{AFF} \rangle \quad \text{Lift} \rightarrow \langle \text{AFF} \rangle$$

Let S be the total number of $\langle \text{AFF} \rangle$ tokens generated.

Boundary Localization & Hidden-State Extraction. Record the positions t_1, \dots, t_S where $\langle \text{AFF} \rangle$ appears. For each $i = 1, \dots, S$, extract the corresponding last-layer hidden vector

$$z_i = h_{t_i} \in \mathbb{R}^D,$$

which encodes the full context immediately preceding the end of sub-task i .

Sequential Mask Decoding. For each step i , use the query E_i to perform cross-attention over the point features F_{obj} and decode a per-point mask:

$$M_i(p) = \sigma(\text{Decoder}(\tilde{F}_p, z_i)) \quad \text{for } p = 1, \dots, N,$$

where \tilde{F} are the fused features and σ is the sigmoid activation. Collect the ordered set $\{M_1, M_2, \dots, M_S\}$ to obtain the final sequence of affordance masks, each aligned with its corresponding sub-task.

B.3 Diffusion Process

Our framework employs diffusion models to learn the conditional distribution $p(X|\mathbf{C})$, of hand-object interaction (HOI) sequences, where the conditioning signal \mathbf{C} combines:

- Object affordance prior $\tilde{\mathbf{A}}_{\text{obj}}$
- Sub-task embedding $f^{\text{clip}}(\tilde{\mathbf{T}}_{\text{sub_tasks}})$
- Object point cloud features \mathbf{F}_{obj}

Forward Process. The diffusion process gradually corrupts the input data through the forward process with a fixed noise schedule $\alpha_t \in [0, T]$

$$p(X_t|X_0) \sim \mathcal{N}(\sqrt{\alpha_t}X_0, (1 - \alpha_t)I). \quad (15)$$

where X_0 is the original HOI sequence, X_t represents its noisy version at timestep t and $\bar{\alpha}_t = \prod_{i=1}^t \alpha_i$. This forward process progressively transforms the data distribution into a tractable Gaussian distribution $\mathcal{N}(0, I)$.

Loss Function. Like VAEs, the diffusion model can be optimized by maximizing the ELBO:

$$\log p_\theta(X_0|\mathbf{C}) = \log \int p_\theta(X_{0:T}|\mathbf{C}) dX_{1:T} \quad (16)$$

$$= \log \int \frac{p_\theta(X_{0:T}|\mathbf{C})p(X_{1:T}|X_0, \mathbf{C})}{p(X_{1:T}|X_0, \mathbf{C})} dX_{1:T} \quad (17)$$

$$= \log \mathbb{E}_{p(X_{1:T}|X_0, \mathbf{C})} \left[\frac{p_\theta(X_{0:T}|\mathbf{C})}{p(X_{1:T}|X_0, \mathbf{C})} \right] \quad (18)$$

$$\geq \mathbb{E}_{p(X_{1:T}|X_0, \mathbf{C})} \left[\log \frac{p_\theta(X_{0:T}|\mathbf{C})}{p(X_{1:T}|X_0, \mathbf{C})} \right] \quad (19)$$

By Simplification, Eq. 19 can be reduced to the following:

$$\arg \max_{\theta} \mathbb{E}_{q(X_{1:T}|X_0, \mathbf{C})} \left[\log \frac{p_\theta(X_{0:T}|\mathbf{C})}{p(X_{1:T}|X_0, \mathbf{C})} \right] \Leftrightarrow \arg \min_{\theta} \frac{1}{2\sigma_t^2} \frac{\bar{\alpha}_{t-1}(1 - \alpha_t)^2}{(1 - \bar{\alpha}_t)^2} \|\hat{X}_\theta(X_t, t, \mathbf{C}) - X_0\|_2^2 \quad (20)$$

where $\sigma_t = \frac{1 - \bar{\alpha}_{t-1}}{1 - \bar{\alpha}_t} \beta_t$. After removing constant terms, we obtain the denoising loss in diffusion models:

$$L_{\text{hoi_diff}}(\hat{X}_\theta, \mathbf{C}) = \mathbb{E}_{X_0 \sim p(X_0), X_t \sim p(X_t|X_0), t \sim [1, T]} \|X_0 - X_\theta(X_t, t, \mathbf{C})\|_2^2 \quad (21)$$

We also introduce geometric loss, including distance map loss $L_{\text{hoi_distance}}$ and relative orientation loss $L_{\text{hoi_orient}}$ for physical plausibility. To enable classifier-free guidance, we randomly mask 10% of the condition to train an unconditional model $X_\theta(X_t, t, \emptyset)$. Since the unconditional model captures the natural HOI sequence, it is then utilized as a prior to generate seamless transitions between different HOI sequences.

Sampling Process. During sampling, we employ classifier-free guidance to enhance alignment with the conditioning input \mathbf{C} . This approach demonstrates superior performance compared to using only the conditional model $X_\theta(X_t, t, \mathbf{C})$:

$$X_\theta^s(X_t, t, \mathbf{C}) = X_\theta(X_t, t, \emptyset) + s \cdot (X_\theta(X_t, t, \mathbf{C}) - X_\theta(X_t, t, \emptyset)), \quad (22)$$

where $s \geq 1$ controls the guidance strength. We generate samples through an iterative denoising process using the reverse diffusion posterior:

$$p(X_{t-1}|X_0, X_t) = \frac{q(X_t|X_{t-1}, X_0)q(X_{t-1}|X_0)}{q(X_t|X_0)} \quad (23)$$

$$= \frac{\mathcal{N}(X_t; \sqrt{\alpha_t}X_{t-1}, (1 - \alpha_t)\mathbf{I}) \mathcal{N}(X_{t-1}; \sqrt{\bar{\alpha}_{t-1}}X_0, (1 - \bar{\alpha}_{t-1})\mathbf{I})}{\mathcal{N}(X_t; \sqrt{\bar{\alpha}_t}X_0, (1 - \bar{\alpha}_t)\mathbf{I})} \quad (24)$$

$$= \mathcal{N}\left(X_{t-1}; \frac{\sqrt{\alpha_t}(1 - \bar{\alpha}_{t-1})X_t + \sqrt{\bar{\alpha}_{t-1}}(1 - \alpha_t)X_0}{1 - \bar{\alpha}_t}, \frac{(1 - \alpha_t)(1 - \bar{\alpha}_{t-1})}{1 - \bar{\alpha}_t}\mathbf{I}\right) \quad (25)$$

$$\approx \mathcal{N}\left(X_{t-1}; \frac{\sqrt{\alpha_t}(1 - \bar{\alpha}_{t-1})X_t + \sqrt{\bar{\alpha}_{t-1}}(1 - \alpha_t)X_\theta^s(X_t, t, \mathbf{C})}{1 - \bar{\alpha}_t}, \frac{(1 - \alpha_t)(1 - \bar{\alpha}_{t-1})}{1 - \bar{\alpha}_t}\mathbf{I}\right) \quad (26)$$

B.4 Loss-guided Physical Refinement

Loss guidance is a technique that minimizes the off-the-shelf loss function $L(X_0, y)$ during the sampling time:

$$\begin{aligned} & \min_{X_0} L(X_0, y) \\ & \text{s.t. } X_0 \in \mathcal{M} \end{aligned} \quad (27)$$

where y is the conditioning input, \mathcal{M} denotes the conditional data manifold that follows the conditional distribution $p(X_0|\mathbf{C})$ learned by the diffusion model. In this work, we propose a novel loss-guided sampling strategy that explicitly enforces physical constraints during the denoising process to achieve more realistic hand-object interactions.

C Additional Experiments

C.1 Results on Extreme-Case: Completely Unseen Datasets

Result on H2O. We further subject our model to extreme-case testing to stress its generalization under the most challenging conditions. All objects and instructions in the **H2O** dataset are **entirely novel** to models trained on GRAB and ARCTIC, making evaluation on H2O a particularly stringent test of generalization. Despite this extreme distribution shift, our experiments demonstrate that the proposed model nonetheless delivers **robust** and **state-of-the-art** performance (shown in Table A1).

Table A1: **Unseen Results on H2O.** Training on GRAB / ARCTIC and evaluation on H2O.

| | Method | MPJPE↓ | FOL↓ | FID↓ | Diversity→ | MModality↑ |
|---------------|---------------|-------------------|------------------|-------------------|------------------|-------------------|
| | GT | – | – | – | 3.43 | – |
| GRAB | MDM[39] | 95.12±3.21 | 0.64±0.04 | 70.54±2.75 | 2.36±0.11 | 12.50±0.50 |
| | TM2T[10] | 90.45±4.50 | 0.68±0.03 | 65.47±3.20 | 2.51±0.10 | 14.45±0.60 |
| | MotionGPT[16] | 85.38±4.00 | 0.61±0.02 | 60.12±2.80 | 2.73±0.13 | 15.93±0.70 |
| | Text2HOI[2] | 80.25±3.80 | 0.63±0.025 | 55.23±2.50 | 1.90±0.14 | 17.00±0.80 |
| | Ours | 75.78±4.68 | 0.52±0.49 | 51.33±3.41 | 3.07±0.27 | 18.15±1.48 |
| ARCTIC | MDM[39] | 105.32±5.00 | 0.80±0.04 | 75.89±3.50 | 1.90±0.09 | 11.27±0.35 |
| | TM2T[10] | 98.47±4.80 | 0.82±0.03 | 72.55±3.30 | 2.10±0.10 | 13.05±0.45 |
| | MotionGPT[16] | 93.15±4.50 | 0.74±0.02 | 65.37±3.00 | 2.30±0.11 | 12.96±0.55 |
| | Text2HOI[2] | 88.02±4.30 | 0.71±0.025 | 60.28±2.80 | 1.50±0.12 | 14.78±0.65 |
| | Ours | 81.36±5.77 | 0.63±0.12 | 55.78±3.62 | 2.69±0.43 | 15.44±1.48 |

C.2 Evaluation for physical realism

We supplemented our experiments by evaluating Physical Realism and IV metrics against the closest baseline, Text2HOI (HOIGPT’s code is not publicly available) in A2, and conducted ablation studies on our Physical Refinement module in A3.

Table A2: **Comparison with Text2HOI**

| Method | Physical realism \uparrow | IV \downarrow |
|---------------|---------------------------------|----------------------------------|
| Seen | | |
| Text2HOI | 0.87 \pm 0.03 | 11.74 \pm 1.22 |
| Ours | 0.93\pm0.02 | 9.25\pm0.73 |
| Unseen | | |
| Text2HOI | 0.79 \pm 0.05 | 14.63 \pm 1.07 |
| Ours | 0.89\pm0.01 | 10.35\pm0.82 |

Table A3: **Ablation Study on Physical Refinement**

| Method | Physical realism \uparrow | IV \downarrow |
|-------------------------|---------------------------------|----------------------------------|
| Seen | | |
| w/o Physical Refinement | 0.89 \pm 0.07 | 10.75 \pm 0.80 |
| Ours | 0.93\pm0.02 | 9.25\pm0.73 |
| Unseen | | |
| w/o Physical Refinement | 0.84 \pm 0.03 | 12.27 \pm 0.48 |
| Ours | 0.89\pm0.01 | 10.35\pm0.82 |

C.3 3D MLLM Fine-tuning

We fine-tune the MLLM on the Affordance dataset [52] and the HOI dataset [38, 6]. We first perform coarse-grained fine-tuning on the Affordance dataset to instill strong affordance priors, and then carry out fine-grained tuning on the HOI dataset to produce our final model [43]. The results shown in Table. A4.

Table A4: **MLLM Coarse-to-Fine Affordance Tuning**

| Method | AUC \uparrow |
|------------------------------------|----------------|
| w/o Fine-tuning | 68.77 |
| Coarse-grained tuning | 84.65 |
| Coarse-to-Fine Tuning (full model) | 87.02 |

C.4 Sensitivity Analysis

We conducted a sensitivity analysis on the guidance rate, and the results are as follows (shown in Table A5 and Table A6). Our experimental results demonstrate that the proposed model maintains robust performance even under these challenging conditions.

C.5 Ablation Study on Multi <AFF>

The additional ablation study results are as follows A7.

C.6 Ablation Study on Motion In-between Refinement

Motion In-between Metric. To the best of our knowledge, no previous work has defined an evaluation metric for motion in-between hand-object interaction. We thus introduce a simple yet effective measure, the ‘‘Smooth Rate,’’ to quantify the temporal continuity of interpolated motion segments as follows,

$$\text{SmoothRate} = \frac{d\text{FID}}{dt}, \quad (28)$$

where dt is the derivative of time. The results are shown in Table A8 and Table A9.

Table A5: Guidance Rate on GRAB

| | Guidance Rate | MPJPE↓ | FOL↓ | FID ↓ | Diversity → | MModality ↑ |
|--------|---------------|-------------------|------------------|-------------------|------------------|-------------------|
| | GT | - | - | - | 4.66 | - |
| Seen | 0.5 | 58.08±0.87 | 0.38±0.01 | 34.40±0.57 | 3.35±0.06 | 18.21±0.31 |
| | 2.0 | 51.86±0.62 | 0.29±0.03 | 27.45±1.13 | 3.63±0.02 | 23.35±0.46 |
| | 2.5 | 47.64±1.03 | 0.26±0.02 | 26.43±0.77 | 3.69±0.27 | 24.59±2.01 |
| | 3.0 | 50.81±1.07 | 0.32±0.03 | 26.75±0.30 | 3.57±0.10 | 23.86±0.77 |
| | 5.0 | 58.92±1.28 | 0.33±0.02 | 34.29±0.47 | 3.40±0.08 | 23.55±0.51 |
| Unseen | 0.5 | 61.39±2.45 | 0.40±0.02 | 35.44±1.45 | 3.32±0.11 | 13.34±0.47 |
| | 2.0 | 54.95±1.30 | 0.30±0.06 | 29.21±2.15 | 3.55±0.07 | 18.70±0.79 |
| | 2.5 | 51.34±0.85 | 0.27±0.01 | 28.29±0.62 | 3.61±0.09 | 19.91±0.63 |
| | 3.0 | 54.62±1.61 | 0.33±0.01 | 28.61±0.81 | 3.50±0.33 | 19.16±1.59 |
| | 5.0 | 62.98±1.93 | 0.35±0.04 | 37.25±0.82 | 3.34±0.14 | 18.81±1.41 |

Table A6: Guidance Rate on ARCTIC

| | Guidance Rate | MPJPE↓ | FOL↓ | FID ↓ | Diversity → | MModality ↑ |
|--------|---------------|-------------------|------------------|-------------------|------------------|-------------------|
| | GT | - | - | - | 3.39 | - |
| Seen | 0.5 | 52.23±1.06 | 0.40±0.01 | 31.05±1.56 | 1.97±0.19 | 12.77±0.45 |
| | 2.0 | 46.04±1.19 | 0.28±0.03 | 20.98±2.30 | 2.62±0.04 | 14.96±0.61 |
| | 2.5 | 45.15±0.94 | 0.25±0.04 | 19.74±0.16 | 2.65±0.03 | 15.25±1.44 |
| | 3.0 | 46.55±1.74 | 0.27±0.02 | 21.03±0.67 | 2.68±0.15 | 15.03±1.75 |
| | 5.0 | 53.25±2.07 | 0.38±0.02 | 32.85±2.04 | 3.40±0.06 | 12.86±2.07 |
| Unseen | 0.5 | 51.67±1.14 | 0.35±0.02 | 31.54±0.68 | 2.07±0.05 | 10.18±0.26 |
| | 2.0 | 47.70±0.88 | 0.30±0.02 | 20.81±2.12 | 2.46±0.07 | 12.36±0.89 |
| | 2.5 | 47.25±0.39 | 0.28±0.03 | 20.05±0.80 | 2.49±0.08 | 12.66±0.71 |
| | 3.0 | 47.76±0.66 | 0.29±0.01 | 21.07±0.41 | 2.51±0.28 | 12.50±1.77 |
| | 5.0 | 54.19±0.89 | 0.34±0.02 | 27.32±0.56 | 2.21±0.06 | 10.53±0.69 |

To determine the most suitable window size [1] for the motion in-between algorithm, we conducted the following comparative experimentsA10.

C.7 Visualization on Affordance

In this subsection, we present visualizations of open-world affordances on seen and unseen objects in Fig. A1.

C.8 Qualitative results Compare with SOTA

This section presents additional visual comparisons between our approach and existing state-of-the-art (SOTA) methods.

Seen Objects. Qualitative results on seen objects in Fig. A2.

Unseen Objects. Qualitative results on unseen objects in Fig. A3.

C.9 Statistically Insignificant

For the ablation study, we performed paired two-sample t-tests, repeating each test five times and reporting the mean p-value. As summarized in Table A11, the proposed method is significant at the 95% confidence level for the majority of metrics(P-value ≤ 0.05).

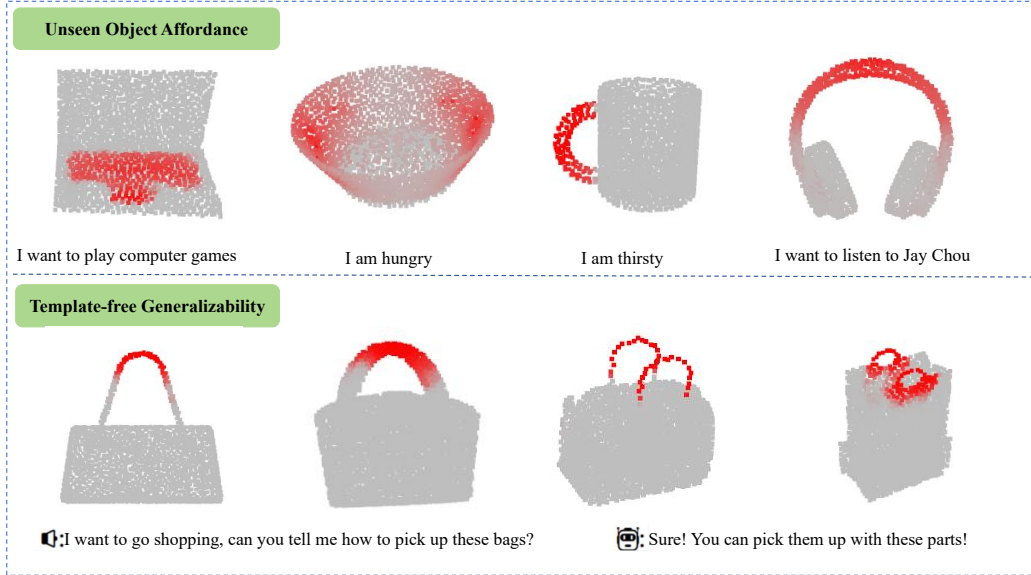


Figure A1: Visualization on Affordance

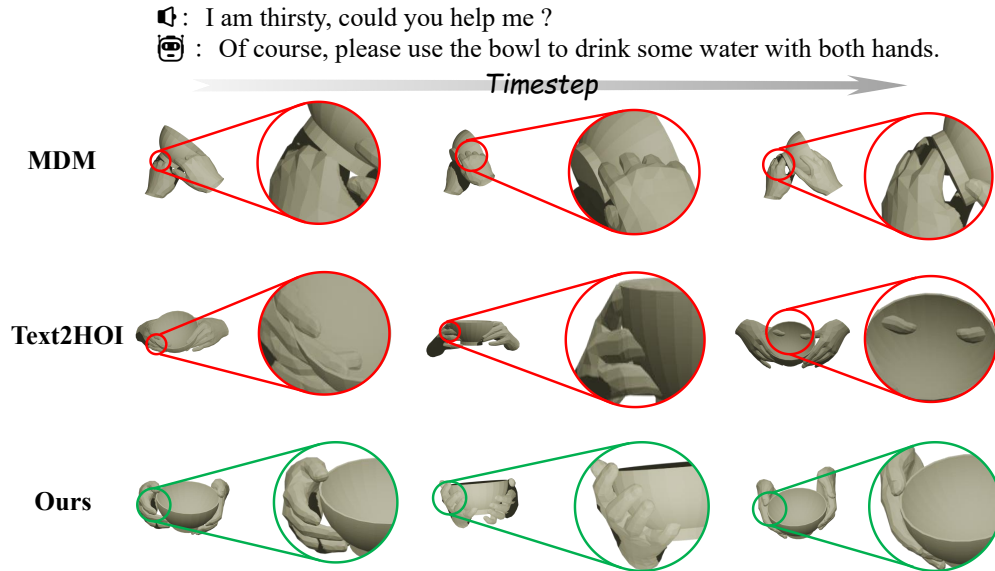


Figure A2: Qualitative results on seen object

Table A7: Ablation on Affordance Configuration (Single vs. Multi)

| Method | MPJPE ↓ | FOL ↓ | FID ↓ | Diversity → | MModality ↑ |
|---------------|-------------------|------------------|-------------------|------------------|-------------------|
| Seen | | | | | |
| Single <AFF> | 52.05±0.82 | 0.33±0.01 | 29.31±0.76 | 3.50±0.12 | 21.05±1.47 |
| Multi <AFF> | 47.64±1.03 | 0.26±0.02 | 26.43±0.77 | 3.69±0.27 | 24.59±2.01 |
| Unseen | | | | | |
| Single <AFF> | 56.48±1.06 | 0.39±0.02 | 34.15±1.08 | 3.40±0.22 | 17.03±1.25 |
| Multi <AFF> | 51.34±0.85 | 0.27±0.01 | 28.29±0.62 | 3.61±0.09 | 19.91±0.63 |

Table A8: Ablation Study of Motion In-between Results on GRAB

| Setting | Method | SmoothRate ↓ |
|---------|-----------------------|--------------------|
| Seen | w/o Motion In-between | 38.18 ± 6.75 |
| | Ours | 2.98 ± 0.43 |
| Unseen | w/o Motion In-between | 35.25 ± 4.91 |
| | Ours | 3.70 ± 0.61 |

D Code and Dataset

D.1 Code

We will release our code as soon as possible. GitHub is OpenHOI

D.2 Dataset

H2O. This dataset contains 571,645 synchronized multi-view RGB-D frames captured with five Kinect sensors in three indoor scenes. Each frame includes 3D poses for both hands, 6-DoF object poses, and verb–noun action labels (36 classes). Split into training (344,645), validation (73,380), and test (153,620) frames, H2O supports egocentric interaction recognition and manipulation benchmarks.

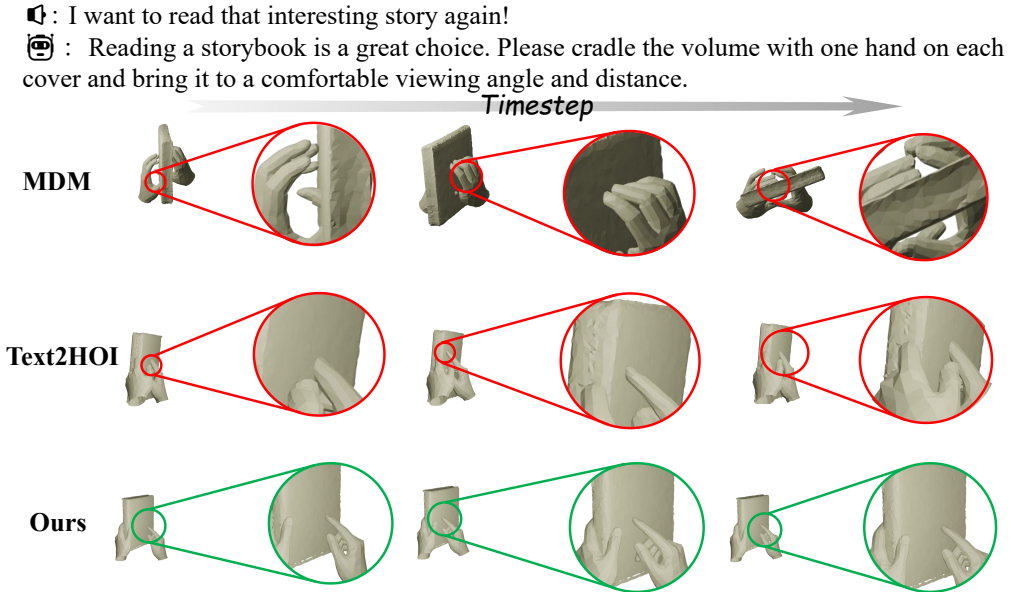


Figure A3: Qualitative results on unseen object

Table A9: Ablation Study of Motion In-between Results on ARCTIC

| Setting | Method | SmoothRate ↓ |
|---------|-----------------------|--------------------|
| Seen | w/o Motion In-between | 41.37 ± 5.98 |
| | Ours | 6.77 ± 2.17 |
| Unseen | w/o Motion In-between | 44.05 ± 5.63 |
| | Ours | 6.04 ± 3.25 |

Table A10: Window Size Compare on GRAB

| Size | SmoothRate (Seen) ↓ | SmoothRate (Unseen) ↓ |
|------|---------------------|-----------------------|
| 1 | 4.77 ± 0.68 | 5.26 ± 2.47 |
| 3 | 3.59 ± 0.81 | 4.58 ± 1.05 |
| 5 | 2.98 ± 0.43 | 3.70 ± 0.61 |
| 10 | 3.24 ± 0.65 | 4.08 ± 0.82 |
| 20 | 3.30 ± 0.57 | 4.19 ± 0.74 |

Dataset annotation. For both the GRAB [38] and ARCTIC [6] datasets, we preprocess the datasets and annotate the semantics. After that, the object point clouds are first upsampled to gain accurate affordance maps using our inference model while ensuring fine geometric details in the meantime. Our model then processes the upsampled data to infer accurate affordance maps, which are subsequently downsampled to match the original resolution for efficient computation.

We preprocess both datasets and get their semantic annotations. First, we upsample the object point clouds to enhance geometric details and generate accurate affordance maps using our inference model. The upsampled data is then processed to infer affordance maps, which are downsampled back to the original resolution for computational efficiency.

To enhance the semantic alignment between language and interaction, we employ a large language model (LLM) to refine the original hand-object interaction (HOI) descriptions. The LLM generates high-level, natural language annotations that better capture the intent and dynamics of HOI.

E Discussion

E.1 Generalizability ability in HOI: Affordance as a key

Affordance is a powerful and explicit prior for interaction that can guide complex and fine-grained HOI synthesis. Our model achieves strong generalization by using affordance as a middleware layer.

- **Open-World Affordance Grounding:** We first employ a coarse-to-fine tuning strategy to equip the model with strong affordance reasoning capabilities, enabling it to generate open-world affordance grounding. This enables the synthesis of realistic HOI sequences, demonstrating strong template-free generalization capabilities.
- **Affordance serves as a crucial condition:** The open-world affordance grounding serves as a crucial condition for the affordance-driven HOI Diffusion. We incorporate affordance not only during training but also in the loss-guidance applied during inference.
- **Affordance-based Refinement:** we design an improved refinement strategy based on affordance: for the interaction between the hand and the object, we optimize based on affordance grounding rather than the conventional closest-surface-point approach.
- **A Template-free Example:** Our 3D MLLM has learned from a wide variety of cups, it can still generate accurate affordance grounding for a completely unseen mug. Accurate affordance grounding will guide the synthesis of realistic HOI sequences.

Table A11: **Two-test Statistically Insignificant**

| Metric | w/o Affordance | w/o CFG | w/o $l_{\text{penetration}}$ | w/o l_{aff} |
|--------------------|----------------------|----------------------|------------------------------|----------------------|
| MPJPE (Seen) | 3.1×10^{-6} | 2.0×10^{-4} | 7.5×10^{-5} | 2.6×10^{-2} |
| MPJPE (Unseen) | 4.4×10^{-5} | 1.2×10^{-4} | 5.5×10^{-4} | 2.6×10^{-4} |
| FOL (Seen) | 1.7×10^{-3} | 6.8×10^{-5} | 1.7×10^{-5} | 3.1×10^{-5} |
| FOL (Unseen) | 2.2×10^{-2} | 2.0×10^{-3} | 4.9×10^{-2} | 3.0×10^{-1} |
| FID (Seen) | 2.4×10^{-5} | 1.5×10^{-3} | 1.0×10^{-2} | 2.0×10^{-3} |
| FID (Unseen) | 2.2×10^{-5} | 8.7×10^{-5} | 2.0×10^{-4} | 2.5×10^{-4} |
| Diversity (Seen) | 1.5×10^{-1} | 1.0×10^{-1} | 4.4×10^{-2} | 1.4×10^{-1} |
| Diversity (Unseen) | 1.8×10^{-3} | 3.0×10^{-4} | 4.6×10^{-2} | 4.8×10^{-2} |
| MModality (Seen) | 2.6×10^{-4} | 2.5×10^{-2} | 5.0×10^{-3} | 1.1×10^{-2} |
| MModality (Unseen) | 1.0×10^{-3} | 4.7×10^{-2} | 2.2×10^{-2} | 2.2×10^{-2} |

E.2 How to use 3D MLLM in Embodied AI: Choose Powerful Foundation Model and Coarse-to-fine tuning

The 3D multimodal large model has been widely applied in HOI and embodied intelligence, and it is very important to choose a basic model suitable for downstream tasks. In OpenHOI, we chose ShapeLLM as our base model, ShapeLLM is a powerful 3D foundation model that performs exceptionally well on multiple downstream tasks (e.g., Embodied Visual Grounding, Visual Question Answering, and Scene Understanding), making it highly suitable for HOI tasks. After research, we found that ShapeLLM has the following advantages and disadvantages

Advantages: ShapeLLM is trained on a large amount of 3D embodied interaction data and achieves state-of-the-art performance across various downstream tasks. It possesses strong priors in 3D interaction and demonstrates impressive zero-shot 3D representation capabilities.

Disadvantages: ShapeLLM has not been trained on part-level object annotations, which limits its reasoning capabilities for fine-grained object understanding.

In order to make the selected 3D base model as suitable as possible for our task, we need to use data to fine tune the model. We adopt a coarse-to-fine tuning strategy: we first pre-train the model on an object-centric affordance dataset to enable it to acquire strong affordance priors. Then, we fine-tune the model on HOI datasets to better align the semantics with the target domain. This allows the model to learn highly effective affordance representations.

E.3 Future of Work: Real-World Applications, about AR/VR and Robotics

OpenHOI can be extended to a wide range of future work in other fields. We have listed several noteworthy areas and provided preliminary solutions for the challenges in future applications

Robotics Manipulation: OpenHOI can be integrated into real-world robotic manipulation systems, including industrial robot arms and service robots, to enable more flexible and human-like interactions [48, 54, 55].

- **Open-World Affordance Grounding as Powerful Guidance for Robots:** Leveraging OpenHOI’s 3D MLLM, our method performs open-world affordance grounding to facilitate the identification of feasible grasping, pushing, and tool-use regions on novel objects, thereby significantly improving success rates for pick-and-place, assembly, and tool-handling tasks.
- **Realistic HOI sequences synthesis for Robot Manipulation:** The HOI sequences generated by OpenHOI can be adapted into robotic manipulation sequences. First, we can use an extraction algorithm to obtain the object’s 6-DoF pose. Then, inverse kinematics are applied to the wrist parameters to compute the robot arm’s pose. Finally, a retargeting algorithm transfers the human hand motions onto various robotic hand configurations for manipulation. This approach ensures smooth, precise, and robust manipulation behaviors in real-world deployments.

Virtual Reality Vision: By synthesizing realistic 3D hand–object interaction sequences, OpenHOI enables users to manipulate virtual objects naturally, for example, by picking up, twisting, or pouring items, thereby enhancing immersion in training simulators, gaming, and virtual prototyping.

Challenges: Robotic manipulation tasks typically demand rapid inference. We plan to employ DPM-Solver to accelerate the diffusion inference process, which is an important direction for our future work[31].