arXiv:2603.19013v4 [cs.CV] 2 Jul 2026

GenHOI: Generalized Hand-Object Pose Estimation with Occlusion Awareness

Hui Yang[1,2], Wei Sun[1*], Jian Liu[1*], Jian Xiao[1], Tao Xie[1], Hossein Rahmani[3], Ajmal Mian[4], Nicu Sebe[5], and Gim Hee Lee[2]
{huiyang, wei_sun, jianliu}@hnu.edu.cn, gimhee.lee@comp.nus.edu.sg

[1] NERC-RVC, Hunan University, [2] National University of Singapore,
[3] Lancaster University, [4] The University of Western Australia,
[5] University of Trento

Abstract. Generalized 3D hand-object pose estimation from a single RGB image remains challenging due to the large variations in object appearances and interaction patterns, especially under heavy occlusion. We propose GenHOI, a framework for generalized hand-object pose estimation with occlusion awareness. GenHOI integrates hierarchical semantic knowledge with hand priors to enhance model generalization under challenging occlusion conditions. Specifically, we introduce a hierarchical semantic prompt that encodes object states, hand configurations, and interaction patterns via textual descriptions. This enables the model to learn abstract high-level representations of hand-object interactions for generalization to unseen objects and novel interactions while compensating for missing or ambiguous visual cues. To enable robust occlusion reasoning, we adopt a multi-modal masked modeling strategy over RGB images, predicted point clouds, and textual descriptions. Moreover, we leverage hand priors as stable spatial references to extract implicit interaction constraints. This allows reliable pose inference even under significant variations in object shapes and interaction patterns. Extensive experiments on the challenging DexYCB and HO3Dv2 benchmarks demonstrate that our method achieves state-of-the-art performance in hand-object pose estimation.

Keywords: Pose Estimation · Hand-Object Interaction

1 Introduction

Accurate 3D hand-object pose recovery is essential for applications in augmented reality [6, 44], human-robot interaction [3, 35, 37], and robotic manipulation [23, 24, 26, 47]. However, real-world hand-object interactions involve a wide variety of objects and interaction scenarios, which hinders the model ability to generalize to unseen objects and interaction patterns. Additionally, frequent occlusions during interactions obscure key visual cues further exacerbating the challenge of generalization in pose estimation.

* Corresponding authors.

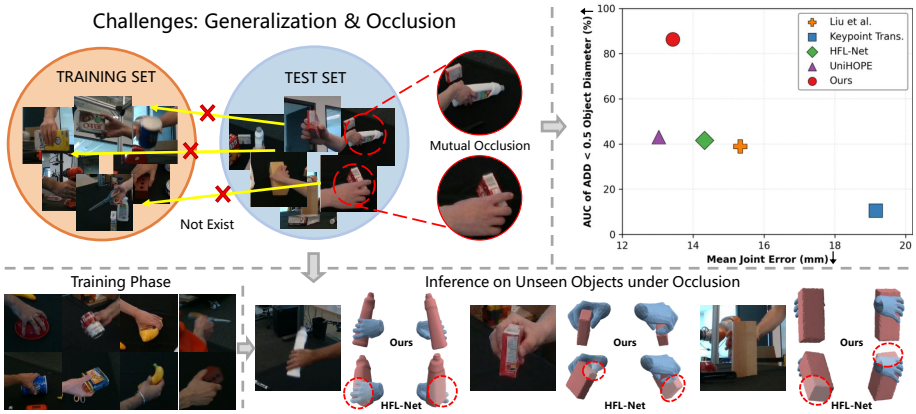

Fig. 1: Illustration of task challenges and key performance comparisons. Hand-object pose estimation (*left*) faces the challenge of generalizing from training set to unseen test set. Additionally, severe hand-object occlusion makes this task more complicated. Our method shows stronger generalization (*bottom:* qualitative comparison with HFL-Net [22] on unseen objects in front and back views, and red dotted circles marking focus areas) and achieves the best performance (*right*) on the challenging generalized DexYCB S3 split [4]. These results confirm effectiveness in generalized settings and under occlusion.

Most existing hand-object pose estimation methods either directly regress poses from visual features [19,34,46] or estimate them through explicit keypoint correspondences [20,22,33]. The former relies only on appearance cues to capture surface patterns rather than intrinsic object properties or interaction intent. This limits generalization to unseen categories and diverse interaction scenarios. The latter depends on accurate keypoint detection, which makes it highly sensitive to occlusions and missing visual evidence. In general, as shown in Fig. 1, the bottleneck limiting the application of hand-object pose estimation methods lies in their generalization and robustness to occlusion.

To address these limitations, we introduce hierarchical semantic information and interaction priors. Hierarchical semantics provide multi-level descriptions of object properties, hand states, and hand-object interaction patterns. This abstract knowledge supports reasoning about unseen objects and novel interactions, and can supplement missing visual cues. Moreover, the hand provides stable structural priors such as joint positions and rotations during interaction. These internal constraints act as reliable spatial references when object shapes vary widely, which further improves generalization. Consequently, fully leveraging semantic cues and interaction priors is essential to overcome the remaining challenges and to improve the generalization of hand-object pose estimation under occlusion.

In this paper, we propose a framework that integrates hierarchical semantic knowledge with hand priors to improve generalization in hand-object pose

estimation, especially under occlusion. Specifically, we introduce a hierarchical semantic prompt that encodes object states, hand configurations, and interaction patterns through textual descriptions. This prompt guides the model toward abstract representations that generalize to unseen objects and novel interactions. Additionally, we introduce a multi-modal masked modeling strategy that leverages RGB images, predicted point clouds, and language descriptions. By randomly masking and reconstructing regions within these modalities, the model learns to perceive occlusions and infer missing visual cues. This occlusion-aware learning process enables it to maintain reliable pose estimation even when visual information is partially missing or corrupted. Furthermore, we incorporate hand priors by using joint positions and rotational parameters as stable references that capture implicit spatial constraints during interaction. These priors improve robustness across large variations in object shapes and interaction patterns, and they support reliable pose estimation under severe occlusion.

Our contributions are summarized as follows:

- We propose a generalized hand-object pose estimation framework tailored to unseen objects and novel interaction patterns under severe occlusion.
- We introduce hierarchical semantic prompts that encode object states, hand configurations, and interaction patterns through textual descriptions. This supports pose estimation under occlusion with stronger generalization.
- We design a multi-modal masked modeling strategy over RGB images, predicted point clouds, and textual descriptions to learn occlusion-aware representations and infer missing visual cues.
- We leverage hand priors as stable spatial references to extract implicit constraints. This helps the model infer object poses across variations in object shapes and interaction patterns, which improves generalization.

2 Related Work

2.1 3D Hand-Object Pose Estimation

Existing hand-object pose estimation approaches can be broadly divided into direct regression and keypoint-based methods. Direct regression methods directly predict the 3D poses of the hand and object from image features. Chen *et al.* [8] proposed a unified framework that jointly learns image features and parametric hand-object representations to infer poses. Chen *et al.* [7] further introduced kinematic constraints to enforce physically plausible hand motion. Qi *et al.* [30] designed a fused image-depth representation to jointly regress hand-object configurations, while Zhang *et al.* [45] incorporated ray-based feature aggregation to better model local hand-object interactions. Keypoint-based approaches estimate the 3D hand-object pose by leveraging spatial relations among detected keypoints. Doosti *et al.* [12] jointly predicted hand and object keypoints and used MANO [31] and PnP to recover 3D poses. Liu *et al.* [27] introduced a semi-supervised framework that utilizes spatiotemporal constraints to provide pseudo-label supervision. Hampali *et al.* [14] integrated cross-attention transformers to

explicitly encode joint hand-object geometry, and Lin *et al.* [22] further enforced dual-stream feature disentanglement for more stable mutual pose refinement.

Despite the progress of both paradigms, most existing methods rely primarily on visual cues or fixed keypoint representations and lack semantic and structural priors. As a result, they struggle in heavy-occlusion scenarios and exhibit limited generalization to unseen objects and novel interaction patterns.

2.2 Textual Semantics for Hand-Object Interaction

Recent studies [28, 43] indicate that purely visual cues often struggle under severe occlusion and large intra-class variation, where textual semantic can provide effective complementary guidance. This has motivated the integration of textual semantics into interaction modeling. Zhang *et al.* [48] used large language models to produce contextual interaction descriptions and combine them with image generators to synthesize hand-object background images. Following this direction, Cha *et al.* [2] introduced a text-guided 3D interaction generation pipeline to produce plausible hand-object configurations. Huang *et al.* [18] further build a generative language-to-3D framework that enables interpretable and expressive interaction synthesis. In parallel, textual supervision has been adopted in object pose estimation [25, 39, 41]. Corsetti *et al.* [9] leveraged vision-language priors for open-vocabulary segmentation and 6D pose estimation, while Lin *et al.* [21] leveraged pretrained vision-language models to incorporate semantic knowledge from both image and text modalities for enhanced category-level understanding.

However, these methods mainly rely on textual cues that describe the global scene or coarse object categories, which provide only high-level context. In contrast, our approach tackles the more challenging hand-object pose estimation task, which demands fine-grained spatial reasoning under severe occlusion. We introduce a hierarchical textual representation that encodes the object, the hand, and their interaction to provide structured multi-level cues for detailed understanding. These hierarchical semantics are integrated with visual and geometric modalities through a multimodal masked modeling strategy, which reconstructs masked signals under random occlusion and improves occlusion awareness and cross-modal reasoning.

3 Method

Fig. 2 shows our proposed GenHOI, a novel framework that integrates hierarchical semantic prompts and hand priors to enhance both generalization and robustness under occlusion.

3.1 Hierarchical Textual Semantic Embedding

We construct a hierarchical prompt template that encodes multi-level semantics of the hand, object, and their interaction. Leveraging this template with the input image, we generate a unified textual prompt that captures both fine-grained visual details and high-level semantic relationships.

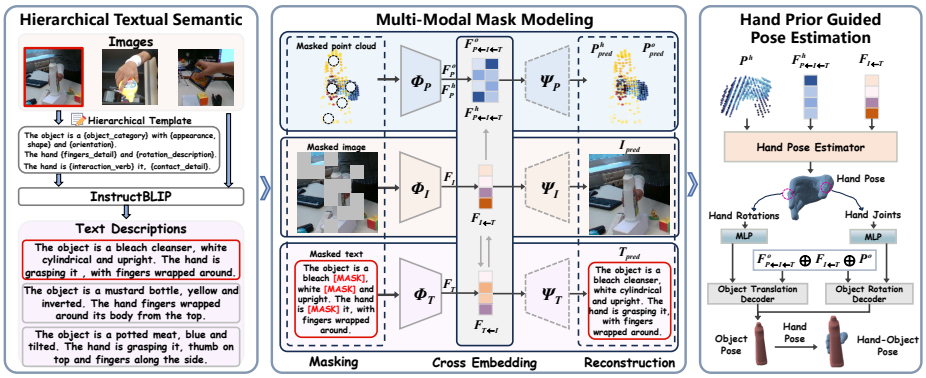

Fig. 2: GenHOI framework, which consists of three key components: (1) **Hierarchical Textual Semantics.** Given RGB images and hierarchical templates with the object, hand, and interaction levels, we employ InstructBLIP to generate hierarchical textual descriptions. (2) **Multi-Modal Mask Modeling.** Given the RGB image, textual description, and corresponding hand-object point cloud, we first apply a modality-specific masking strategy to the inputs. The masked inputs are processed by the cross-modal embedding module, which fuses visual, geometric, and textual features to produce representations used for reconstruction and pose estimation. The dashed boxes indicate that masking and reconstruction are used only during training. (3) **Hand Prior Guided Pose Estimation.** The cross-modal features learned in the previous stage are aggregated to a robust representation. Using these fused features, we first estimate hand pose parameters, which then serve as reliable priors for object pose reasoning.

Hierarchical Prompt Template. We define three semantic components to construct the hierarchical prompt template: object-level semantics s_o, hand-level semantics s_h, and interaction-level semantics s_i.

The object-level semantics s_o describe object properties such as category, appearance, and orientation, formulated as "the object is a {object_category} with a {shape, appearance}, and {orientation_description}." The hand-level semantics s_h characterize hand articulation and grasp configuration, expressed as "the hand {fingers_detail} and {rotation_description}." The interaction-level semantics s_i capture the relational context between the hand and the object, following the structure "the hand is {interaction_verb} it, {contact_detail}." Finally, these three components are integrated into a unified hierarchical prompt $\mathcal{T}$. This process enables the model to capture a comprehensive multi-level semantic context by linking cues from the object, hand, and their interaction. The final hierarchical template is shown in Fig. 2.

Hierarchical Textual Semantic Generation. Given an input image $I \in \mathbb{R}^{3\times H\times W}$ and a hierarchical template $\mathcal{T}$, we leverage a vision-language model such as InstructBLIP [10] to generate a textual description T. Formally, the generation process is defined as:

$$T = \mathcal{G}(I, \mathcal{T}), \tag{1}$$

where $\mathcal{G}(\cdot)$ denotes the generative vision-language model. This model fills the template slots with image-specific content to produce a textual description that aligns the visual observations with the hierarchical semantics encoded in $\mathcal{T}$. As shown in Fig. 2, the model generates the hierarchical prompt $\mathcal{T}$ based on the visual content. To improve robustness against potential noise or bias in automatically generated descriptions, textual perturbations are introduced during training to reduce dependence on perfectly consistent inputs. More details are provided in the supplementary material.

This unified hierarchical textual description provides rich semantic information that encapsulates both local and global contextual cues of the hand-object interaction. It is then fused with visual and geometric representations to enable a more robust and generalizable pose estimation under occlusion.

3.2 Multi-Modal Masked Modeling

Visual cues often become incomplete under severe occlusion due to hand object interaction, while point cloud geometry and textual semantics can provide complementary information. To enhance occlusion awareness and improve cross-modal consistency, we propose a multi-modal masked modeling strategy that jointly masks and reconstructs visual, geometric, and textual modalities. During training, each image is paired with a point cloud sampled from ground-truth hand and object mesh. During inference, we dynamically sample a point cloud from a voxel grid guided by image features. The sampling follows [30, 40] and provides a geometric representation for each image without requiring ground-truth meshes. Masking and reconstruction are applied only during training to help the model infer missing information and align multi-modal representations. Moreover, the model directly utilizes the learned cross-modal fusion features from image, point cloud, and text, and these integrated representations are then used for subsequent hand-object pose estimation.

Modality-specific Masking Strategy. We apply three modality-specific masks: 1) Image masking corrupts selected patches in the hand-object region and the background. 2) Point cloud masking drops random local patches for both hand and object geometry. 3) Text masking replaces a subset of tokens in the hierarchical description with [MASK].

Image Masking: Given an input image I, the masked image $\tilde{I}$ is defined as:

$$\begin{aligned} M_I &= M_{ho}(\lfloor \alpha N_I \rfloor) \cup M_{bg}(N_I - \lfloor \alpha N_I \rfloor), \\ \tilde{I} &= (1 - M_I) \odot I + M_I \odot \mathcal{N}(0, \sigma^2), \end{aligned} \tag{2}$$

where $\odot$ denotes element-wise multiplication, and $M_I \in \{0,1\}^{H \times W}$ is the mask for the image. Here, M_{ho} and M_{bg} are the sets of masked patches sampled from the hand-object region and background region, respectively. N_I is the total number of masked patches and $\alpha \in [0,1]$ controls the proportion of patches masked

within the hand-object region. $\mathcal{N}(0, \sigma^2)$ denotes Gaussian noise. Our target-centric masking balances patches from the hand-object region and background, which preserves interaction-focused cues and global contextual information.

Point Cloud Masking: Given a hand point cloud $P^h \in \mathbb{R}^{3\times N_h}$, we first partition it into K local patches $\bigcup_{k=1}^{K} \mathbf{P}_k^h = P^h$. The masked hand point cloud $\tilde{\mathbf{P}}_k^h$ is then defined as:

$$\tilde{\mathbf{P}}_k^h = \begin{cases} \mathbf{0}, & k \in \mathcal{K}^h, \\ \mathbf{P}_k^h, & k \notin \mathcal{K}^h, \end{cases} \tag{3}$$

where $\mathcal{K}^h$ is the set of patch indices randomly selected according to the masking ratio $\beta \in [0, 1]$. The object point cloud P^o is masked in the same manner as hand, yielding the final masked point clouds $\tilde{P}^h$ and $\tilde{P}^o$.

Textual Masking: Given a hierarchical textual description $\mathbf{T} = \{t_l\}_{l=1}^{L}$ of length L, we randomly select a subset of token positions $\mathcal{L}$ to mask according to a masking ratio $\gamma \in [0, 1]$. The masked sequence $\tilde{\mathbf{T}} = \{\tilde{t}_l\}_{l=1}^{L}$ is then defined as:

$$\tilde{t}_l = \begin{cases} \text{[MASK]}, & l \in \mathcal{L}, \\ t_l, & l \notin \mathcal{L}, \end{cases} \tag{4}$$

where [MASK] denotes the special mask token [11].

Multi-Modal Masked Reconstruction. Given the masked inputs $\tilde{I}$, $\tilde{P}$, and $\tilde{T}$ obtained from the previous stage, the model first extracts modality-specific representations to encode visual, geometric, and textual information. Specifically, the masked image $\tilde{I}$ is encoded by the visual encoder ϕ_I to produce $F_I \in \mathbb{R}^{H_I \times W_I \times D_I}$. The masked point cloud $\tilde{P}$ is encoded by the geometric encoder ϕ_P to produce $F_P \in \mathbb{R}^{N_P \times D_P}$. The masked textual description $\tilde{T}$ is encoded by the textual encoder ϕ_T to produce $F_T \in \mathbb{R}^{L_T \times D_T}$. These representations serve as the foundational features for subsequent cross-modal embedding and reconstruction.

Visual and textual features are first fused through cross-attention since text-guided visual features help produce more accurate point cloud sampling. For visual feature F_I and textual feature F_T, the attention is formulated as:

$$\begin{aligned} \mathbf{Q}_I = F_I \mathbf{W}_Q^I, \quad \mathbf{K}_T = & F_T \mathbf{W}_K^T, \quad \mathbf{V}_T = F_T \mathbf{W}_V^T, \\ F_{I \leftarrow T} = \text{Softmax}\Big(\frac{\mathbf{Q}_I \mathbf{K}_T^\top}{\sqrt{d}}\Big)\mathbf{V}_T, \end{aligned} \tag{5}$$

with $F_{T \leftarrow I}$ obtained through the same formulation as above. Reconstruction heads ψ_I and ψ_T predict the recovered image I_{pred} and textual description T_{pred}:

$$I_{pred} = \psi_I(F_{I \leftarrow T}), \quad T_{pred} = \psi_T(F_{T \leftarrow I}). \tag{6}$$

The semantically enhanced visual feature $F_{I \leftarrow T}$ is fused with the hand point cloud feature F_P^h via cross-attention to produce the fused enhanced hand point

cloud feature $F^h_{P\leftarrow T\leftarrow I}$. The enhanced features of the hand point cloud $F^h_{P\leftarrow I\leftarrow T}$ is directly fed into a reconstruction head ψ_P to predict the reconstructed hand point cloud:

$$P^h_{\text{pred}} = \psi_P(F^h_{P\leftarrow I\leftarrow S}). \tag{7}$$

The object point cloud P^o_{pred} is reconstructed in the same manner.

This hierarchical fusion and reconstruction mechanism lets each modality draw on complementary signals from the others. It supports occlusion-aware completion and preserves semantic coherence and structural consistency across visual, geometric, and textual modalities.

3.3 Hand Prior Guided Pose Estimation

To enhance the generalization capability of the model across diverse objects and interaction scenarios, we introduce a hand prior guided hand-object pose estimation module. This module leverages the inherent structural regularities of the hand as reliable prior information to guide object pose reasoning. Compared to diverse and deformable objects, the hand exhibits a relatively fixed motion topology and limited shape variation. The estimated hand pose and geometric information thus provide stable and effective spatial constraints for inferring object pose.

Hand Pose Estimation. We incorporate an implicit geometric by predicting the signed distance field (SDF). Given hand points P_h, we apply Fourier positional encoding $\gamma(\cdot)$ to enrich spatial frequency representation, and fuse it with $F_{I\leftarrow T}$ to regress the SDF_h:

$$SDF_h = \text{MLP}_{\text{SDF}}\left(\gamma(P_h) \oplus F_{I\leftarrow T} \oplus P_h\right). \tag{8}$$

The process of estimating object SDF_o follows the same steps as in SDF_h. We estimate the hand pose by integrating multiple cues, including the explicit geometric feature, the semantically enhanced visual feature, and the implicit geometric from the predicted SDF_h. The SDF-guided aggregation encourages features to focus on the hand-object surface.

$$F^h_{\text{agg}} = \left(F^h_{P\leftarrow I\leftarrow S} \oplus F_{I\leftarrow T}\right) \cdot \frac{1}{\beta_h} \cdot \sigma_h\left(\frac{SDF_{\text{h}}}{\beta_h}\right), \tag{9}$$

$$(\{\theta_i\}_{i=1}^{16}, \alpha) = \text{MLP}_h\Big(\text{Transformer}_h(F^h_{\text{agg}})\Big), \tag{10}$$

where $\sigma_h(\cdot)$ is the sigmoid function, β_h is a learnable scale, $\theta \in \mathbb{R}^{16\times 3}$ represents joint rotations, and $\alpha \in \mathbb{R}^{10}$ denotes MANO hand shape coefficients. The aggregation object feature F^o_{agg} is obtained in the same manner as for the hand.

Object Pose Estimation. The object rotation and translation are jointly regressed by incorporating hand cues. Specifically, the joint rotations θ are projected and globally aggregated to produce a compact hand motion descriptor, and the reconstructed 3D hand joints $J \in \mathbb{R}^{21\times 3}$ provide spatial cues of the grasp. These hand features are concatenated with the object feature F^o_{agg} to estimate the full object pose:

$$\begin{aligned} R_{\text{obj}} &= \text{MLP}_R\Big(F^o_{\text{agg}} \oplus \text{Pool}\big(\phi_{\text{rot}}(\theta_{i\,i=1}^{16})\big)\Big), \\ T_{\text{obj}} &= \text{MLP}_T\Big(F^o_{\text{agg}} \oplus \text{Pool}\big(\phi_{\text{pos}}(J)\big)\Big). \end{aligned} \tag{11}$$

The final object pose is represented as $[R_{\text{obj}}, T_{\text{obj}}]$, where $R_{\text{obj}} \in \mathbb{R}^{3\times 3}$ is a 3D rotation matrix and $T_{\text{obj}} \in \mathbb{R}^3$ is a 3D translation vector.

3.4 Loss Functions

Our overall training objective:

$$\begin{aligned} L_{\text{total}} = \lambda_1 L_{\text{rec}} + \lambda_2 L_{\text{text}} + \lambda_3 L_{\text{pc}} \\ + \lambda_4 L_{\text{mano}} + \lambda_5 L_{\text{obj}} + \lambda_6 L_{\text{SDF}} + \lambda_7 L_{\text{others}}. \end{aligned} \tag{12}$$

combines multiple loss terms to jointly optimize multi-modal reconstruction quality and pose estimation accuracy. Specifically, we implement the image reconstruction loss L_{rec}, the point cloud reconstruction loss L_{pc}, the SDF supervision loss L_{sdf}, the hand pose loss L_{mano}, and the object pose loss L_{obj} as L1 losses. These loss terms are supervised by image cues, geometric consistency, implicit geometry, hand pose, and object pose, respectively. In contrast, the text reconstruction loss L_{text} is formulated as a cross-entropy loss to reconstruct masked text tokens. The auxiliary loss terms L_{others} follow Qi *et al* [30] to provide task-specific constraints. Each term is weighted by its corresponding coefficient λ to balance its contribution during optimization.

4 Experiments

Our method is implemented in PyTorch and trained on a single NVIDIA RTX 4090 GPU. We adopt the Adam optimizer with a batch size of 24. The initial learning rate is set to 1e-4 and decayed by a factor of 0.7 every 5 epochs. The model is trained for 60 epochs on both the DexYCB [4] and HO3Dv2 [13] datasets, which is sufficient to achieve satisfactory performance. Further details are in supplementary material.

Datasets. We evaluate our method on two widely used hand-object interaction datasets. **DexYCB** [4] contains 582K images from over 1000 sequences, where 10 subjects manipulate 20 YCB objects [36]. We follow the S0 split for standard evaluation. The split uses sequences from 8 subjects for training and 2 subjects for testing, which keeps the test subjects and interaction patterns unseen during

Table 1: Evaluation of hand and object pose estimation on the DexYCB S3 split [4]. Results are for unseen objects and novel hand-object interactions. 'avg' denotes average results over all evaluated objects. ↓ means lower is better, ↑ indicates higher is better. Best results are bolded, and the second best are underlined. HPE: Hand Pose Estimation. HOPE: Hand-Object Pose Estimation.

	Method	Object Pose (mAP)				Hand Pose (mm)			
		AUC of ADD-0.5D ↑				MJE ↓	PA-MJE ↓	V-PE ↓	PA-V-PE ↓
		gelatin_box	bleach_cleanser	wood_block	avg				
HPE	HandOccNet [29]	–	–	–	–	14.58	6.73	14.10	6.49
HPE	MobRecon [5]	–	–	–	–	15.40	7.25	14.24	6.39
HPE	H2ONet [38]	–	–	–	–	15.20	6.35	15.03	6.74
HPE	SimpleHand [49]	–	–	–	–	14.88	6.74	14.21	6.45
HOPE	Liu et al. [27]	26.31	25.07	68.56	38.89	15.43	6.61	14.91	6.40
HOPE	Keypoint Trans. [14]	0.00	1.31	32.61	10.47	18.79	7.77	18.35	7.94
HOPE	HFL-Net [22]	25.88	32.08	70.16	41.59	14.77	6.64	14.29	6.41
HOPE	UniHOPE [34]	26.23	32.32	74.29	43.06	**13.31**	**6.08**	**12.89**	**5.87**
HOPE	**GenHOI (ours)**	**86.70**	**89.27**	**92.06**	**89.34**	13.67	6.39	13.32	5.93

Table 2: Evaluation of object pose estimation using ADD-S on DexYCB S3 split [4].

Method	ADD-S (mm) ↓			
	gelatin_box	bleach_cleanser	wood_block	average
HFL-Net [22]	94.74	176.84	72.93	114.83
UniHOPE [34]	109.23	173.38	49.76	110.79
HOISDF [30]	52.31	89.92	109.89	84.04
GenHOI (ours)	**13.14**	**26.91**	**25.23**	**21.76**

training. We also adopt the S3 split to evaluate generalization to unseen objects. The split uses 15 objects for training and holds out the remaining 3 objects for testing. **HO3Dv2** [13] contains 77K images from 68 sequences. The dataset covers 10 subjects interacting with 10 YCB objects. We use the official train/test split, where the test set contains 1 unseen object category to evaluate cross-category generalization, and submit test results to the official evaluation server.

Evaluation Metrics. Hand Pose and Mesh Reconstruction: To assess hand pose accuracy, we report Mean Joint Error (M-JE), Procrustes-Aligned MJE (PA-MJE) [50], and Scale-Translation aligned MJE (ST-MJE) [51]. Since our method directly regresses MANO parameters, we further evaluate hand mesh reconstruction quality using Mean Mesh Error (MME), the area under the curve of the percentage of correct vertices (V-AUC), and F-scores at 5mm and 15mm thresholds (F@5, F@15), together with their Procrustes-aligned versions following H2ONet [38]. **Object Pose:** For object pose estimation, we report Object Center Error (OCE), Mean Corner Error (MCE), Object Mesh Error (OME), and average closest point distance (ADD-S). We further evaluate the average 3D distance (ADD) of the object. Specifically, we report ADD-0.5D, which measures the percentage of predictions whose ADD is within 50% of the object diameter.

Table 3: Quantitative comparison of hand mesh with across occlusion levels on DexYCB S3 split dataset [4].

Methods	Occlusion (25%-50%)				Occlusion (50%-75%)				Occlusion (75%-100%)			
	J-PE↓	PA-J-PE↓	V-PE↓	PA-V↓	J-PE↓	PA-J-PE↓	V-PE↓	PA-V↓	J-PE↓	PA-J-PE↓	V-PE↓	PA-V↓
HFL-Net [22]	16.33	7.00	15.81	6.77	18.66	7.33	18.11	7.11	28.95	8.80	27.94	8.53
UniHOPE [34]	**14.59**	**6.39**	**14.13**	**6.17**	16.27	**6.51**	**15.78**	6.29	26.42	7.64	**25.51**	**7.40**
Ours	14.75	6.49	14.36	6.23	**16.18**	6.61	15.89	**6.25**	**26.28**	**7.55**	25.64	7.57

Table 4: Quantitative results on the unseen object ("019 pitcher base") from the HO3Dv2 dataset [13].

Method	HFL-Net [22]	HOISDF [30]	Ours
Metrics in [mm]	ADD-0.5D	ADD-0.5D	ADD-0.5D
019 pitcher base	24.1	88.4	**92.7**

4.1 Comparison with SOTA Methods

Comparisons on DexYCB. We evaluate our method on the DexYCB S3 split dataset [4], which focuses on unseen objects. Following UniHOPE [34], we evaluate only hand-object interaction frames. This split presents significant challenges due to severe hand-object occlusions and the inclusion of unseen objects, providing a rigorous evaluation of the generalization ability and robustness. Table 1 presents quantitative results for hand and object pose estimation. Our method achieves competitive hand pose accuracy, closely matching state-of-the-art (SOTA) approaches. More importantly, it substantially outperforms existing methods in object pose estimation by achieving an average AUC of 89.34% across all evaluated categories. The relatively smaller improvement in hand pose accuracy compared to object pose stems from the inherent differences in variability. The hand exhibits a consistent structure and appearance across samples, which makes it less influenced by the generalization design of our framework. To enable a clearer comparison under a shared mm-based object metric, we further report the ADD-S results in Table 2. The results show that our method achieves the lowest ADD-S error among all compared methods, reducing the error from 84.04 mm of HOISDF to 21.76 mm. In contrast, objects exhibit much larger shape and texture variation. Our hierarchical textual semantics and hand priors therefore provide stronger semantic and geometric constraints. Following UniHOPE, we evaluate hand pose accuracy across occlusion ratios under the same protocol in Table. 3. Our method achieves comparable performance under heavy occlusion despite UniHOPE includes a dedicated hand-occlusion removal module. Fig. 3 shows the corresponding qualitative comparison on unseen objects under the S3 split. The results highlight improved hand-object pose accuracy over HFL-Net and UniHOPE.

Comparisons on HO3Dv2. We evaluate our method on the HO3Dv2 dataset [13], which mainly consists of objects seen during training, along with a single unseen object ("019 pitcher base"). To assess the generalization capability, we

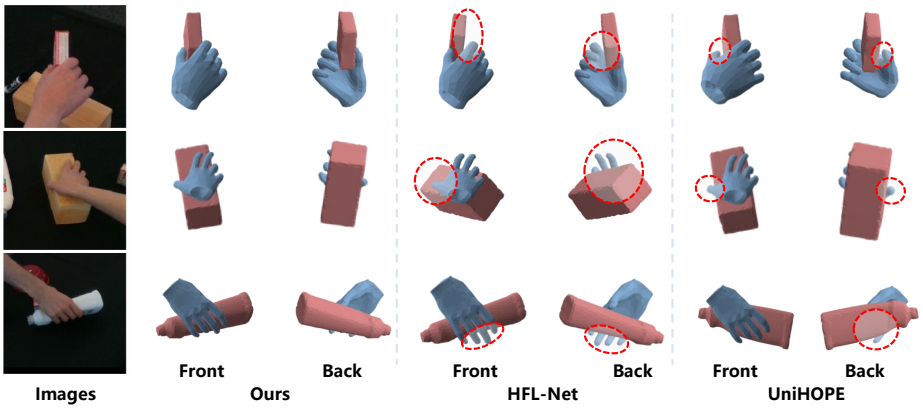

Fig. 3: Qualitative comparison of hand-object pose estimation on unseen objects under the DexYCB S3 split [4]. Front and back indicate the front and rear views, respectively. Red dotted circles highlight regions where other methods produce less accurate pose estimates than our method. This demonstrates the superior generalization and robustness of our method on unseen objects.

Table 5: Comparison with SOTA hand-object pose estimation methods on the HO3Dv2 dataset [13].

Metric in [mm]	MJE ↓	ST-MJE ↓	PA-MJE ↓	OME ↓	ADD-S ↓
Hasson *et al.* [17]	-	31.8	11.0	-	-
Hasson *et al.* [15]	-	36.9	11.4	67.0	22.0
Hasson *et al.* [16]	-	26.8	12.0	80.0	40.0
Liu *et al.* [27]	-	31.7	10.1	-	-
Hampali *et al.* [14]	25.5	25.7	10.8	68.0	21.4
HFL-Net [22]	28.9	28.4	**8.9**	64.3	32.4
HOISDF [30]	23.6	22.8	9.6	48.5	17.8
LCP [20]	-	23.7	**8.9**	-	-
UniHOPE [34]	-	25.5	9.6	-	-
GenHOI (ours)	**21.3**	**21.1**	9.9	**35.9**	**13.4**

evaluate our model on the unseen object, as presented in Table 4. Despite not being seen during training, our method achieves a competitive ADD-0.5D score of 92.7% that surpasses existing approaches. This strong generalization to unseen objects benefits from the integration of hierarchical textual semantics and hand priors. Table 5 reports quantitative comparisons on the seen objects. Our method consistently surpasses previous approaches across all metrics in achieving the lowest hand and object pose errors. These results demonstrate that the proposed model accurately captures hand–object interactions and yields robust estimation even under challenging occlusions.

Qualitative comparisons on both unseen and seen objects are shown in Fig. 4. Specifically, the top-left example corresponds to the unseen object, while the re-

Table 6: Ablation study of our proposed components on the DexYCB dataset S3 split full [4]. 'w/o' denotes 'without'.

	Ablation	MJE ↓	PA-MJE ↓	V-PE ↓	PA-V-PE ↓	ADD-0.5D ↑
	Hierarchical Textual Semantic					
1	w/o Text Generation Template	16.78	6.73	15.12	6.23	86.51
2	w/o Multi-level Semantic Knowledge	17.04	6.98	15.75	6.41	86.32
	Multi-Modal Masked Modeling					
3	w/o Image Masking	15.33	6.52	14.78	6.03	87.65
4	w/o Text Masking	14.98	6.21	14.50	5.88	88.12
5	w/o Point Cloud Masking	15.21	6.47	14.91	6.07	87.90
	Hand Priors					
6	w/o Hand Rotation Prior	14.65	6.12	14.32	5.78	85.45
7	w/o Hand Joint Prior	14.88	6.29	14.55	5.91	84.01
	GenHOI (Full)	**13.42**	**5.81**	**12.96**	**5.63**	**90.44**

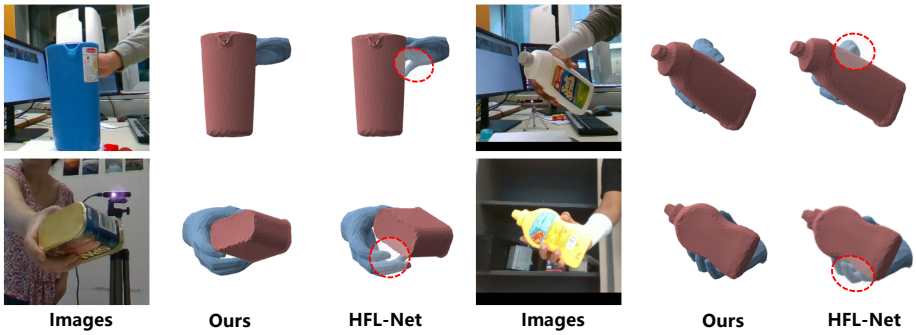

Fig. 4: Qualitative comparison on the HO3Dv2 dataset [13]. The top-left example show results on the unseen object ("019 pitcher base"), while the remaining examples correspond to objects seen during training.

maining examples show results on seen objects. The qualitative results further confirm that our method attains SOTA performance on HO3Dv2 while maintaining robust generalization beyond the training distribution.

4.2 Ablation Studies

We perform the ablations on the DexYCB S3 split full [4], using all frames to evaluate the contribution of each component.

Hierarchical Textual Semantic. Rows (1-2) of Table 6 show apparent performance drops when removing either the text template or hierarchical semantic descriptions. Our motivation for introducing hierarchical semantics is to provide structured level-dependent priors about object and interaction attributes. These cues offer high-level concepts as well as fine-grained part-level hints, which guide

the model to reason beyond purely visual correlation. The model is forced to rely mainly on appearance cues without them, and thus weakening its generalizability. The resulting performance gap confirms that hierarchical semantics serve as a semantic generalization prior, which is crucial when handling novel hand-object interaction scenarios with severe occlusion.

Multi-Modal Masked Modeling. Rows (3-5) of Table 6 indicate that image masking contributes the most to performance, and text and point cloud masking further enhance occlusion awareness. The multi-modal masking strategy enables the model to infer occluded information. The model develops occlusion-aware reasoning and cross-modal feature agreement by learning to reconstruct from incomplete observations, which are crucial for robust performance in complex interaction scenes. The consistent performance degradation when the masks are removed validates that explicitly learning to reason under occlusion is central to improving generalization in real-world settings.

Hand Priors. Rows (6-7) of Table 6 show that the removal of hand priors only slightly affects hand pose estimation but leads to a clear drop in object pose accuracy, which indicates that stable hand priors provide essential spatial constraints for interaction reasoning. This is because hand configuration and articulation inherently encode grasp intent and contact topology, which implicitly restrict the feasible orientation and position of the object in 3D space. Even when the hand pose can still be estimated from appearance cues, the loss of these structured priors weakens the geometric coupling between hand and object. As a result, object pose inference becomes more sensitive to local visual variations. The model relies more on appearance similarity without such constraints. This makes object pose inference harder when shapes and interaction patterns vary, which reduces generalization to unseen objects and novel interaction scenarios.

Additional ablations and analyses on multimodal mask ratios, multimodal fusion strategies, robustness to inaccurate or noisy text descriptions, as well as inference efficiency are provided in the supplementary material.

5 Conclusion

In this work, we presented GenHOI, a novel hand-object pose estimation framework that improves generalization to unseen objects and interaction scenarios. We introduced hierarchical textual semantics to encode object states, hand configurations, and interaction patterns that provide high-level guidance beyond visual appearance. We proposed a multi-modal masked modeling strategy that enables the model to infer occluded regions and maintain cross-modal consistency under partial observations. Furthermore, hand priors serve as stable spatial constraints that anchor interaction geometry to reinforce object pose reasoning even when object shapes vary. Extensive experiments demonstrate that our GenHOI achieves strong generalization and robust performance under severe occlusion. For future work, we may extend the model to egocentric datasets and further validate its generalization.

Acknowledgements

This work is supported by the National Natural Science Foundation of China under Grants 62473141 and U22A2059, National Science and Technology Major Project of China under Grants 2026ZD1610900, China Scholarship Council under Grant 202506130069, the Open Foundation of the State Key Laboratory of Advanced Design and Manufacturing for Vehicle Body, the National Research Foundation (NRF) Singapore, under its NRF-Investigatorship Programme (Award ID. NRF-NRFI09-0008), and the Tier 2 grant MOET2EP20124-0015 from the Singapore Ministry of Education.

References

1. Akiva, P., Purri, M., Dana, K., Tellman, B., Anderson, T.: H2o-net: Self-supervised flood segmentation via adversarial domain adaptation and label refinement. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 111–122 (2021) 20
2. Cha, J., Kim, J., Yoon, J.S., Baek, S.: Text2hoi: Text-guided 3d motion generation for hand-object interaction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1577–1585 (2024) 4
3. Chao, Y.W., Paxton, C., Xiang, Y., Yang, W., Sundaralingam, B., Chen, T., Murali, A., Cakmak, M., Fox, D.: Handoversim: A simulation framework and benchmark for human-to-robot object handovers. In: Proceedings of the IEEE International Conference on Robotics and Automation. pp. 6941–6947 (2022) 1
4. Chao, Y.W., Yang, W., Xiang, Y., Molchanov, P., Handa, A., Tremblay, J., Narang, Y.S., Van Wyk, K., Iqbal, U., Birchfield, S., et al.: Dexycb: A benchmark for capturing hand grasping of objects. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9044–9053 (2021) 2, 9, 10, 11, 12, 13, 19, 20, 24
5. Chen, X., Liu, Y., Dong, Y., Zhang, X., Ma, C., Xiong, Y., Zhang, Y., Guo, X.: Mobrecon: Mobile-friendly hand mesh reconstruction from monocular image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20544–20554 (2022) 10, 20
6. Chen, Y., Wang, Q., Chen, H., Song, X., Tang, H., Tian, M.: An overview of augmented reality technology. In: Journal of Physics: Conference Series. vol. 1237, p. 022082 (2019) 1
7. Chen, Z., Chen, S., Schmid, C., Laptev, I.: Gsdf: Geometry-driven signed distance functions for 3d hand-object reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12890–12900 (2023) 3, 20
8. Chen, Z., Hasson, Y., Schmid, C., Laptev, I.: Alignsdf: Pose-aligned signed distance fields for hand-object reconstruction. In: European Conference on computer vision. pp. 231–248. Springer (2022) 3, 20
9. Corsetti, J., Boscaini, D., Oh, C., Cavallaro, A., Poiesi, F.: Open-vocabulary object 6d pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18071–18080 (2024) 4
10. Dai, W., Li, J., Li, D., Tiong, A., Zhao, J., Wang, W., Li, B., Fung, P.N., Hoi, S.: Instructblip: Towards general-purpose vision-language models with instruction tuning. Advances in Neural Information Processing Systems **36**, 49250–49267 (2023) 5, 25

11. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the conference of the North American chapter of the association for computational linguistics: human language technologies. pp. 4171–4186 (2019) 7
12. Doosti, B., Naha, S., Mirbagheri, M., Crandall, D.J.: Hope-net: A graph-based model for hand-object pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6608–6617 (2020) 3
13. Hampali, S., Rad, M., Oberweger, M., Lepetit, V.: Honnotate: A method for 3d annotation of hand and object poses. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3196–3206 (2020) 9, 10, 11, 12, 13
14. Hampali, S., Sarkar, S.D., Rad, M., Lepetit, V.: Keypoint transformer: Solving joint identification in challenging hands and object interactions for accurate 3d pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11090–11100 (2022) 3, 10, 12, 20
15. Hasson, Y., Tekin, B., Bogo, F., Laptev, I., Pollefeys, M., Schmid, C.: Leveraging photometric consistency over time for sparsely supervised hand-object reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 571–580 (2020) 12
16. Hasson, Y., Varol, G., Schmid, C., Laptev, I.: Towards unconstrained joint hand-object reconstruction from rgb videos. In: International Conference on 3D Vision. pp. 659–668 (2021) 12, 20
17. Hasson, Y., Varol, G., Tzionas, D., Kalevatykh, I., Black, M.J., Laptev, I., Schmid, C.: Learning joint reconstruction of hands and manipulated objects. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11807–11816 (2019) 12, 20
18. Huang, M., Chu, F.J., Tekin, B., Liang, K.J., Ma, H., Wang, W., Chen, X., Gleize, P., Xue, H., Lyu, S., et al.: Hoigpt: Learning long-sequence hand-object interaction with language models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 7136–7146 (2025) 4
19. Jiang, S., Ye, Q., Xie, R., Huo, Y., Li, X., Zhou, Y., Chen, J.: In-hand 3d object reconstruction from a monocular rgb video. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 2525–2533 (2024) 2
20. Kuang, Z., Ding, C., Yao, H.: Learning context with priors for 3d interacting hand-object pose estimation. In: Proceedings of the ACM International Conference on Multimedia. pp. 768–777 (2024) 2, 12, 20
21. Lin, X., Zhu, M., Dang, R., Zhou, G., Shu, S., Lin, F., Liu, C., Chen, Q.: Clipose: Category-level object pose estimation with pre-trained vision-language knowledge. IEEE Transactions on Circuits and Systems for Video Technology **34**(10), 9125–9138 (2024) 4
22. Lin, Z., Ding, C., Yao, H., Kuang, Z., Huang, S.: Harmonious feature learning for interactive hand-object pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12989–12998 (2023) 2, 4, 10, 11, 12, 20
23. Liu, J., Sun, W., Liu, C., Zhang, X., Fu, Q.: Robotic continuous grasping system by shape transformer-guided multi-object category-level 6d pose estimation. IEEE Transactions on Industrial Informatics **19**(11), 11171–11181 (2023) 1
24. Liu, J., Sun, W., Yang, H., Deng, P., Liu, C., Sebe, N., Rahmani, H., Mian, A.: Diff9d: Diffusion-based domain-generalized category-level 9-dof object pose estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence **47**(7), 5520–5537 (2025) 1

25. Liu, J., Sun, W., Yang, H., Zeng, Z., Liu, C., Zheng, J., Liu, X., Rahmani, H., Sebe, N., Mian, A.: Deep learning-based object pose estimation: A comprehensive survey. International Journal of Computer Vision **134**(2), 81 (2026) 4
26. Liu, J., Sun, W., Zeng, K., Zheng, J., Yang, H., Rahmani, H., Mian, A., Wang, L.: Scalable unseen objects 6-dof absolute pose estimation with robotic integration. IEEE Transactions on Robotics **42**, 1884–1901 (2026) 1
27. Liu, S., Jiang, H., Xu, J., Liu, S., Wang, X.: Semi-supervised 3d hand-object poses estimation with interactions in time. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14687–14697 (2021) 3, 10, 12
28. Liu, Y., Long, X., Yang, Z., Liu, Y., Habermann, M., Theobalt, C., Ma, Y., Wang, W.: 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. pp. 7037–7047 (2025) 4
29. Park, J., Oh, Y., Moon, G., Choi, H., Lee, K.M.: Handoccnet: Occlusion-robust 3d hand mesh estimation network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1496–1505 (2022) 10, 20
30. Qi, H., Zhao, C., Salzmann, M., Mathis, A.: Hoisdf: Constraining 3d hand-object pose estimation with global signed distance fields. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10392–10402 (2024) 3, 6, 9, 10, 11, 12, 19, 20, 24
31. Romero, J., Tzionas, D., Black, M.J.: Embodied hands: Modeling and capturing hands and bodies together. ACM Transactions on Graphics **36**(6) (2017) 3
32. Tse, T.H.E., Kim, K.I., Leonardis, A., Chang, H.J.: Collaborative learning for hand and object reconstruction with attention-guided graph convolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1654–1664 (2022) 20
33. Wang, R., Mao, W., Li, H.: Interacting hand-object pose estimation via dense mutual attention. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 5735–5745 (2023) 2
34. Wang, Y., Xu, H., Heng, P.A., Fu, C.W.: Unihope: A unified approach for hand-only and hand-object pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12231–12241 (2025) 2, 10, 11, 12, 20
35. Wang, Z., Chen, J., Chen, Z., Xie, P., Chen, R., Yi, L.: Genh2r: Learning generalizable human-to-robot handover via scalable simulation demonstration and imitation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16362–16372 (2024) 1
36. Xiang, Y., Schmidt, T., Narayanan, V., Fox, D.: Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. arXiv preprint arXiv:1711.00199 (2017) 9
37. Xu, H., Li, H., Wang, Y., Liu, S., Fu, C.W.: Handbooster: Boosting 3d hand-mesh reconstruction by conditional synthesis and sampling of hand-object interactions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10159–10169 (2024) 1
38. Xu, H., Wang, T., Tang, X., Fu, C.W.: H2onet: Hand-occlusion-and-orientation-aware network for real-time 3d hand mesh reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17048–17058 (2023) 10
39. Yang, H., Sun, W., Liu, J., Zheng, J., Dai, Z., Mian, A.: Lancope: Language-guided category-level object pose estimation from a single rgb image. IEEE Robotics and Automation Letters (2025) 4

40. Yang, H., Sun, W., Liu, J., Zheng, J., Xiao, J., Mian, A.: Occlusion-aware 3d hand-object pose estimation with masked autoencoders. arXiv preprint arXiv:2506.10816 (2025) 6
41. Yang, H., Sun, W., Liu, J., Zheng, J., Zeng, Z., Mian, A.: Rgb-based category-level object pose estimation via depth recovery and adaptive refinement. IEEE Robotics and Automation Letters (2025) 4
42. Yang, L., Li, K., Zhan, X., Lv, J., Xu, W., Li, J., Lu, C.: Artiboost: Boosting articulated 3d hand-object pose estimation via online exploration and synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2750–2760 (2022) 20
43. Ye, Y., Gupta, A., Kitani, K., Tulsiani, S.: G-hop: Generative hand-object prior for interaction reconstruction and grasp synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1911–1920 (2024) 4
44. Ye, Y., Li, X., Gupta, A., De Mello, S., Birchfield, S., Song, J., Tulsiani, S., Liu, S.: Affordance diffusion: Synthesizing hand-object interactions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22479–22489 (2023) 1
45. Zhang, C., Di, Y., Zhang, R., Zhai, G., Manhardt, F., Tombari, F., Ji, X.: Ddf-ho: Hand-held object reconstruction via conditional directed distance field. Advances in Neural Information Processing Systems **36**, 56871–56884 (2023) 3
46. Zhang, C., Jiao, G., Di, Y., Wang, G., Huang, Z., Zhang, R., Manhardt, F., Fu, B., Tombari, F., Ji, X.: Moho: Learning single-view hand-held object reconstruction with multi-view occlusion-aware supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9992–10002 (2024) 2
47. Zhang, H., Christen, S., Fan, Z., Hilliges, O., Song, J.: Graspxl: Generating grasping motions for diverse objects at scale. In: European Conference on Computer Vision. pp. 386–403 (2024) 1
48. Zhang, M., Fu, Y., Ding, Z., Liu, S., Tu, Z., Wang, X.: Hoidiffusion: Generating realistic 3d hand-object interaction data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8521–8531 (2024) 4
49. Zhou, Z., Zhou, S., Lv, Z., Zou, M., Tang, Y., Liang, J.: A simple baseline for efficient hand mesh reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1367–1376 (2024) 10
50. Zimmermann, C., Brox, T.: Learning to estimate 3d hand pose from single rgb images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4903–4911 (2017) 10
51. Zimmermann, C., Ceylan, D., Yang, J., Russell, B., Argus, M., Brox, T.: Freihand: A dataset for markerless capture of hand pose and shape from single rgb images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 813–822 (2019) 10

Appendix

In this supplementary material, we first introduce the implementation details (Section A). Section B presents additional qualitative results and ablation experiments to further demonstrate the robustness of GenHOI, and Section C discusses the limitations and future directions.

A. Implementation Details

The input images are cropped around the object and resized to 224×224. In our proposed masking strategy, the number of image masking blocks is denoted as N_I and set to 12. For masks falling within the object bounding box, the parameter α is configured to 50%. For text masking, the masking ratio is denoted as γ and set to 50%. For point cloud masking, the number of groups is denoted as K and set to 64, while the proportion of point cloud masking groups corresponding to the parameter β is 30%. The number of hand points is denoted as N_H and set to 600, and the number of object points is denoted as N_O and set to 200. The feature dimensions for images, point clouds and text are respectively denoted as D_I, D_P and D_T. These three feature dimensions are set to 768, 768 and 512. Through extensive experimental validation, it has been found that setting λ_1, λ_2, λ_3, λ_4, λ_5 and λ_6 to 1 achieves the optimal performance. The parameter λ_7 is configured according to the parameter settings specified in HOISDF [30].

B. More Experimental Results

B.1 Quantitative Results on the DexYCB S0 Split

We report quantitative results on the DexYCB S0 split dataset [4] in Tables A and B. We evaluate our method on the standard S0 split of the DexYCB dataset [4], where the test set contains objects seen during training. Table A summarizes the quantitative comparison with SOTA methods. Our approach achieves the best performance across all reported metrics. The corresponding qualitative results for the S0 split are provided in the supplementary material.

In addition, as our method involves MANO-based hand reconstruction, we further compare it with SOTA hand mesh reconstruction methods in Table B. Our method achieves the best performance under both procrustes-aligned and non-aligned metrics, demonstrating its joint-level accuracy and detailed mesh reconstruction capability under complex occlusions and hand-object interactions.

B.2 Qualitative Results on the DexYCB S0 Split

We demonstrate the superiority of our method on the DexYCB S0 split dataset [4] in Figs. A and Figs. B.

Table A: Comparison with SOTA hand-object pose estimation methods on the DexYCB S0 split dataset [4].

Metric in [mm]	MJE ↓	PA-MJE ↓	OCE ↓	MCE ↓	ADD-S ↓
Hasson *et al.* [17]	17.6	-	-	-	-
Hasson *et al.* [16]	18.8	-	-	52.5	-
Tze *et al.* [32]	15.3	-	-	-	-
ArtiBoost [42]	12.8	-	-	-	-
AlignSDF [8]	19.0	-	27.0	-	-
gSDF [7]	14.4	-	19.1	-	-
Keypoint Trans. [14]	12.7	6.86	27.3	32.6	15.9
HFL-Net [22]	11.9	5.81	39.8	45.7	31.9
LCP [20]	11.1	5.30	-	-	-
HOISDF [30]	10.1	5.31	18.4	27.4	13.3
UniHOPE [34]	12.6	5.54	-	-	-
GenHOI (ours)	**9.83**	**5.12**	**17.2**	**25.1**	**12.3**

Table B: Quantitative comparison with hand mesh metrics on the DexYCB S0 split dataset [4].

Methods	PA-J-AUC↑	PA-V-PE↓	PA-V-AUC↑	PA-F@5↑	PA-F@15↑	J-AUC↑	V-PE↓	V-AUC↑	F@5↑	F@15↑
HandOccNet [29]	88.4	5.5	89.0	78.0	99.0	74.8	13.1	76.6	51.5	92.4
MobRecon [5]	87.3	5.6	88.9	78.5	98.8	73.7	13.1	76.1	50.8	92.1
H2ONet [1]	88.9	5.5	89.1	80.1	99.0	74.6	13.0	76.2	51.3	92.1
HOISDF [30]	89.8	5.1	89.8	80.7	99.2	**80.7**	9.8	80.6	59.6	95.0
UniHOPE [34]	-	5.4	-	-	-	-	12.2	-	-	-
Ours	**90.0**	**4.9**	**90.1**	**81.3**	**99.4**	**80.7**	**9.5**	**81.1**	**60.4**	**95.4**

B.3 Additional Ablation Results

To further analyze our model design and in accordance with the main paper, we conducted all of the following ablation studies on DexYCB S3 split [4].

Ablation on Proposed Components under Different Occlusion Levels. As shown in Table C, we further analyze the robustness of our proposed components under different occlusion levels on the DexYCB S3 split. The results show that performance degradation becomes more pronounced as the occlusion ratio increases, indicating the importance of these components for occlusion-aware hand-object reasoning. Hierarchical textual semantics provide high-level object and interaction guidance to compensate for missing or ambiguous visual evidence. Multi-modal masked modeling improves robustness by encouraging reasoning from complementary image, text, and point-cloud cues. Hand priors regularize plausible hand articulation and rotation when visible hand regions are incomplete. The consistent drops caused by removing these components further validate their effectiveness under occlusion.

Ablation on Image Object-Centric Masking Ratio and the Total Number of Masking Patches. As shown in Table D, we analyze how the image masking ratio and the number of masking patches affect model robustness. With

Table C: Occlusion-aware ablation study of the proposed components on the DexYCB S3 split, following the ablation indices in Table 5.

Ablation	Occlusion 25%–50%					Occlusion 50%–75%					Occlusion 75%–100%				
	J-PE	PA-J-PE	V-PE	PA-V	ADD	J-PE	PA-J-PE	V-PE	PA-V	ADD	J-PE	PA-J-PE	V-PE	PA-V	ADD
1	16.87	6.82	15.62	6.51	86.20	18.87	7.06	17.88	6.93	84.35	30.45	8.54	29.38	8.42	78.70
2	17.16	7.03	16.18	6.68	85.95	19.24	7.31	18.42	7.05	84.05	31.10	8.87	30.08	8.73	78.10
3	15.68	6.72	15.22	6.38	87.45	17.24	6.92	16.88	6.50	85.90	28.10	8.05	27.52	8.02	80.65
4	15.39	6.60	14.95	6.32	88.10	16.83	6.80	16.48	6.43	86.55	27.42	7.91	26.84	7.89	81.45
5	15.58	6.68	15.14	6.37	87.72	17.08	6.87	16.73	6.49	86.15	27.90	8.00	27.26	7.98	80.95
6	15.28	6.56	14.82	6.30	84.90	16.62	6.74	16.31	6.38	83.25	27.24	7.83	26.68	7.86	77.40
7	15.36	6.62	14.97	6.34	83.70	16.75	6.82	16.45	6.45	81.95	27.52	7.95	26.96	7.94	75.85

Table D: Ablation on the image masking ratio α and the number of masking patches N_I. Here, α controls the proportion of masked regions within the object bounding box, and N_I denotes the total number of masked image patches. "All" indicates that all patches inside the bounding box are masked.

Image Mask Ratio	**Image Mask Pacth**	MJE ↓	PA-MJE ↓	V-PE ↓	PA-V-PE ↓	ADD-0.5D ↑
0%	12	15.33	6.52	14.78	6.03	87.65
25%	12	14.15	6.21	13.86	5.94	88.17
50%	**12 (Ours)**	**13.42**	**5.81**	**12.96**	**5.63**	**90.44**
75%	12	15.48	6.43	14.96	6.31	87.95
100%	12	16.75	6.89	16.29	6.56	85.68
50%	4	14.98	6.39	14.63	6.15	88.32
50%	8	13.98	6.10	13.64	5.93	89.12
50%	**12 (Ours)**	**13.42**	**5.81**	**12.96**	**5.63**	**90.44**
50%	16	13.75	6.08	13.41	5.95	89.01
100%	All	24.86	11.69	25.73	10.95	73.21

Table E: Ablation study on the text masking ratio γ. We vary the proportion of masked tokens while keeping all other settings fixed.

Text Masking Ratio	MJE ↓	PA-MJE ↓	V-PE ↓	PA-V-PE ↓	ADD-0.5D ↑
0%	14.98	6.21	14.50	5.88	88.12
25%	14.26	5.96	14.13	5.82	88.75
50% (Ours)	**13.42**	**5.81**	**12.96**	**5.63**	**90.44**
75%	15.79	6.45	15.13	6.13	84.05
100%	17.85	7.25	16.98	6.98	79.21

a fixed patch count, increasing the masking ratio initially improves performance by encouraging reliance on stable global structure rather than occlusion-prone local textures. When the ratio too large, excessive visual information loss prevents reliable structural reasoning and leads to degradation, indicating that the ratio must balance occlusion simulation with essential cue preservation. With a fixed masking ratio, a moderate number of patches performs best. Too few patches offer limited occlusion diversity, while too many fragment the masked

Table F: Ablation study on the point cloud masking ratio. We fix the number of point cloud groups to $K = 64$ and vary the proportion of groups being masked.

Point Cloud Masking Ratio	MJE ↓	PA-MJE ↓	V-PE ↓	PA-V-PE ↓	ADD-0.5D ↑
0%	15.21	6.47	14.91	6.07	87.90
30% (Ours)	**13.42**	**5.81**	**12.96**	**5.63**	**90.44**
50%	17.35	7.71	17.04	7.28	82.23
70%	24.89	9.56	25.45	8.42	71.19
100%	36.05	13.79	36.67	14.84	54.93

Table G: Ablation study on the cross-attention order among text, image, and point cloud modalities.

Fusion Order	MJE ↓	PA-MJE ↓	V-PE ↓	PA-V-PE ↓	ADD-0.5D ↑
Text→Point→Image	14.27	6.23	13.85	6.12	87.18
Text→Image & Text→Point	13.76	6.02	13.24	5.89	88.76
Text→Image→Point (Ours)	**13.42**	**5.81**	**12.96**	**5.63**	**90.44**

regions and disrupt object structure. Overall, image masking strategy provides the greatest robustness gain only when the masking ratio and patch count are jointly balanced. As shown in the last row of Table D, when the entire object bounding box is masked, most object-related and some hand-related visual cues are occluded, leading to a severe drop in pose estimation accuracy.

Ablation on Text Masking Ratio. To evaluate the impact of masking ratio on text semantic infusion, we conduct an ablation study by varying the text masking ratio while keeping all other settings fixed. As shown in Table E, compared with low masking ratios, a moderate masking ratio yields better performance because the model leverages textual contextual semantics to reconstruct missing information, enhancing its robustness and contextual understanding. Conversely, excessively high masking ratios remove too much semantic information, limiting the model to infer missing information from context and weakening the textual semantic supplement for occluded visual cues. These results indicate that an appropriate masking ratio enhances the model ability to reason over contextual semantics.

Ablation on Point Cloud Masking Ratio. To evaluate how the point cloud masking ratio affects geometric reasoning, we vary the proportion of masked point cloud groups while fixing the total number of groups at $K = 64$. As shown in Table F, a moderate masking ratio of 30% achieves the best overall performance. This level of masking provides enough geometric degradation for the model to learn occlusion-aware reasoning, encouraging it to infer missing geometry using context. In contrast, a very low masking ratio offers limited improvement because the model seldom encounters incomplete geometry and thus cannot effectively learn to handle occluded or sparse point clouds. However, when the masking ratio becomes excessively high, performance drops sharply

Table H: Ablation on the hierarchical template. We assess each component by retaining only the object-, hand-, or interaction-related part of the hierarchical template.

Template Configuration	MJE ↓	PA-MJE ↓	V-PE ↓	PA-V-PE ↓	ADD-0.5D ↑
Object Component Only	14.52	6.37	14.31	6.28	88.56
Hand Component Only	13.98	6.13	13.65	6.06	87.15
Interaction Component Only	13.82	6.05	13.36	5.81	89.27
w/o YCB category tokens	13.45	5.78	13.03	5.61	90.22
Full Template (ours)	**13.42**	**5.81**	**12.96**	**5.63**	**90.44**

Table I: Ablation on noise augmentation. We study the effect of adding noise to text and point cloud to improve model robustness.

Noise Augmentation	MJE ↓	PA-MJE ↓	V-PE ↓	PA-V-PE ↓	ADD-0.5D ↑
w/o Text Noise	14.55	6.18	13.48	6.24	89.12
w/o Point Cloud Noise	14.41	6.21	13.82	6.28	88.53
Full Noise (ours)	**13.42**	**5.81**	**12.96**	**5.63**	**90.44**

because too many point cloud groups are removed, leaving insufficient local samples for reliable SDF learning and too little geometric structure for the model to reconstruct the missing regions from contextual cues. Overall, these results demonstrate that a balanced masking ratio is essential: it must be large enough to promote occlusion-aware learning, yet not so large that geometric supervision and contextual reconstruction collapse.

Ablation on Cross-Attention Order. Table G compares different cross-attention orders among text, image, and point cloud modalities. When the model directly fuses text with point cloud features before visual modality (*Text* → *Point* → *Image*), the performance drops significantly. This is because textual tokens encode high-level semantic concepts, while point cloud features represent raw local geometry; without intermediate visual grounding, the model struggles to establish reliable correspondences between abstract semantic cues and local geometry. The parallel fusion strategy (*Text* → *Image & Text* → *Point*) alleviates this issue but still suffers from inconsistent feature correspondence between the two branches. In contrast, our sequential order (*Text* → *Image* → *Point*) first injects textual semantics into the visual features to establish semantic-visual alignment and then transfers the enriched visual cues to the 3D representation, resulting in more coherent multi-modal interaction and improved estimation accuracy across all metrics.

Ablation on Template Ablation. Table H evaluates the contribution of each component in the hierarchical template. Using only the object-related template provides global cues about category and shape but omits hand articulation and contact information, limiting its ability to resolve interaction geometry. Conversely, the hand-only template lacks object attributes, making it inadequate for constraining object pose under occlusion. The interaction-only template cap-

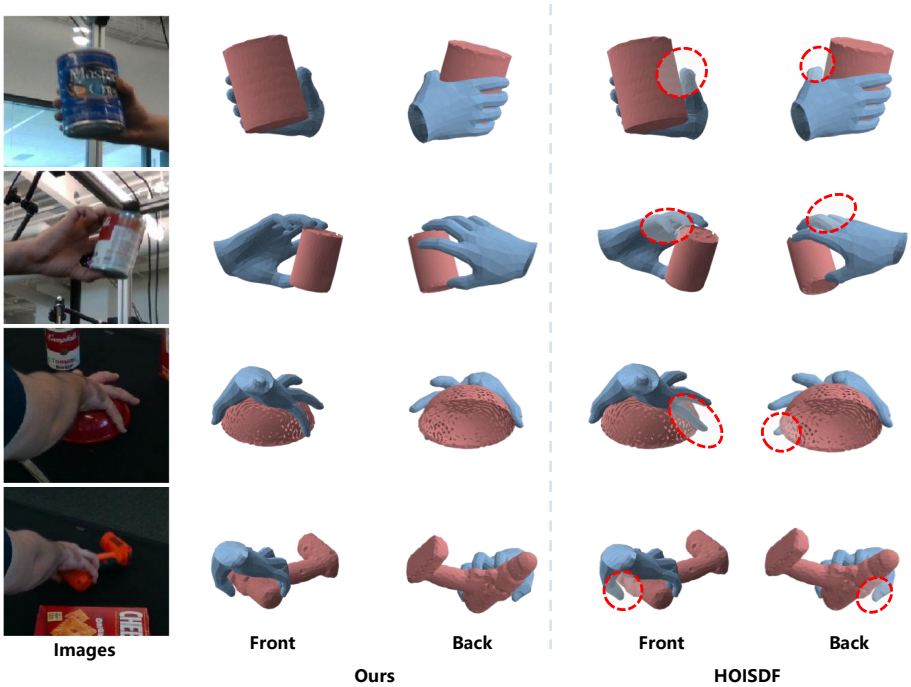

Fig. A: Qualitative comparison between our method and HOISDF [30] on the DexYCB S0 split [4].

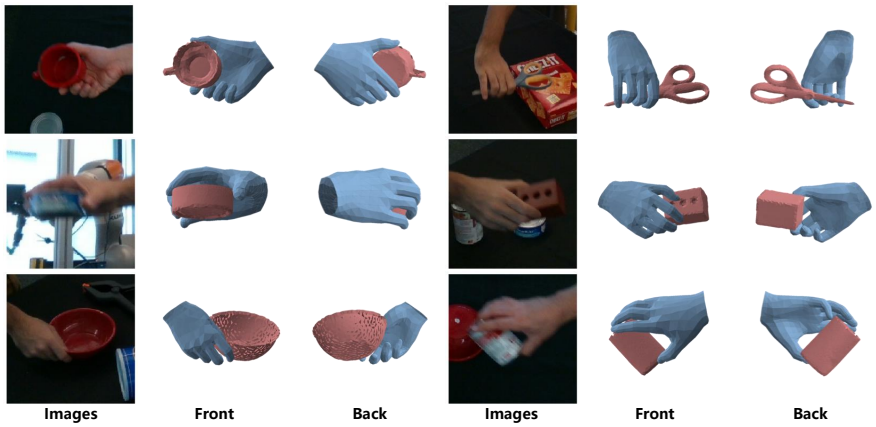

Fig. B: Qualitative results of our method on the DexYCB S0 split [4].

tures relational intent but excludes hand-obejct specific details, leading to poor performance when visual cues are ambiguous. In contrast, the full hierarchical template, which combines object, hand, and interaction semantics, consistently improves accuracy across all metrics. This demonstrates that a richer semantic

context helps the model better understand the interaction and infer more precise hand-object poses.

In addition, we include an ablation without YCB category tokens, where object categories are inferred by the LLM rather than provided as ground-truth labels, to explicitly examine whether category information is leaked to the model. As shown in In the fourth row of Tab. H, removing YCB category tokens results in a marginal change in accuracy. This indicates that explicit category labels are not essential for pose estimation. Instead, the model mainly relies on high-level appearance cues and hand-object interaction semantics encoded in the hierarchical template.

Ablation on Noise Augmentation. Table I evaluates the effect of adding noise to text descriptions and point clouds. Our method uses hierarchical prompts generated by InstructBLIP [10], whose quality, consistency, and potential biases may affect pose estimation. To reduce sensitivity to partially incorrect, biased, or domain-shifted textual inputs, we introduce token-level perturbations during training, allowing the model to learn robustness to imperfect text. Similarly, point cloud noise is added during training to mitigate the distribution gap between ideal training samples (from ground-truth sampling) and the voxel-based sampling used at inference. The results show that removing either type of noise degrades performance, while applying both jointly (*Full Noise*) achieves the best accuracy, demonstrating that these augmentations improve robustness and generalization under imperfect inputs.

Runtime Analysis. To assess the computational efficiency of our proposed framework, we evaluate the inference speed, parameter count, and memory consumption. The model achieves a inference speed of 18 FPS (text generated offline), demonstrating its practicality for real-world applications. The total parameter count remains moderate at 292 MB, and the memory footprint is limited to approximately 5109 MB. These results indicate that our framework achieves a good balance between efficiency and accuracy, making it suitable for deployment in scenarios requiring real-time or near real-time hand–object pose estimation.

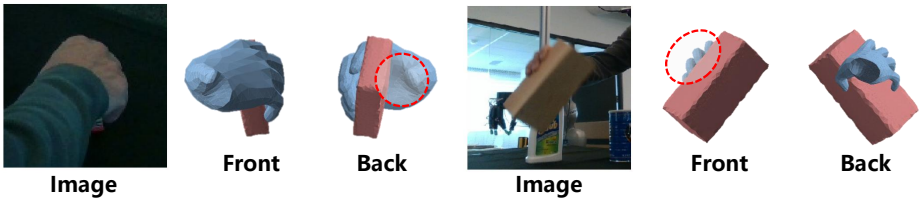

Fig. C: Failure cases. In scenarios with severe occlusions, the predicted hand-object poses exhibit inaccuracies.

C. Limitations and Future Works

Limitations: We observe that failures mainly occur in extreme occlusion scenarios. Failure cases in extreme-occlusion interaction scenarios are shown in Fig.

C. In cases where both the hand and object are severely occluded, the model struggles to capture sufficient visual cues for accurate spatial reasoning, often resulting in erroneous pose estimations. Particularly, when the object is entirely obscured by the hand, the lack of observable features significantly hampers the model ability to infer object orientation and position. These limitations highlight the need for future work on enhancing the model ability to reason under severe occlusions.

Future Work: To address the limitation under severe occlusion, future work will focus on incorporating temporal cues from previous interaction frames together with physical and motion constraints to better infer occluded hand-object interaction relationship. In addition, we aim to extend the evaluation to both third-person and egocentric views by constructing a multi-view dataset that integrates third-person data and egocentric recordings. Training on multi-view data will enhance the model viewpoint adaptability, and we will further explore multi-view feature fusion to improve hand-object interaction pose estimation accuracy.