SG-MoE: Structured Geometric Mixture-of-Experts for Monocular Human–Object Interaction Reconstruction

Anonymous authors
Paper under double-blind review

Abstract

Monocular human–object interaction reconstruction requires recovering both individual surfaces and their coherent spatial configuration from a single image, which is challenging due to severe depth ambiguity and mutual occlusion. Existing methods typically rely on a fixed combination of interaction and geometric reasoning mechanisms throughout progressive reconstruction, despite the evolving demands of different decoding stages. We introduce SG-MoE, a structured geometric mixture-of-experts framework that adaptively composes complementary geometric reasoning operations during reconstruction. The proposed experts capture global human–object interaction, entity-specific topology, and evolving relative geometry, enabling the decoder to refine human and object representations according to the current reconstruction state. We further introduce a bidirectional geometric interaction mechanism that couples intermediate geometry with cross-entity feature reasoning. Experiments on BEHAVE and InterCap demonstrate consistent improvements in human and object reconstruction as well as contact recovery over state-of-the-art methods, while maintaining favorable model complexity and inference efficiency.

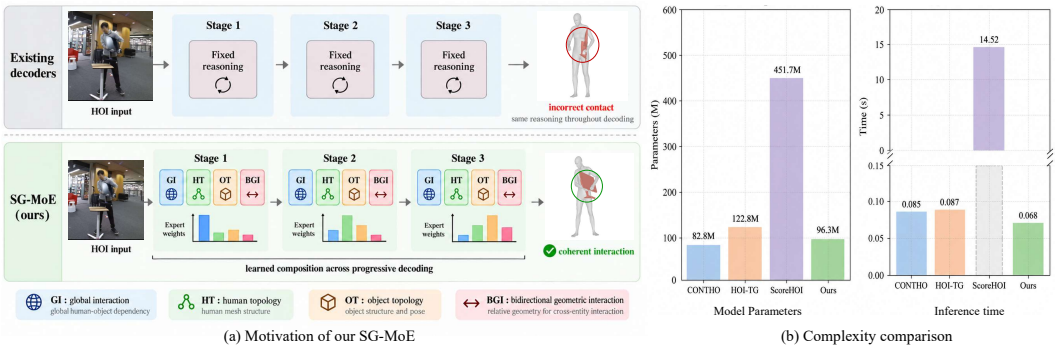
1 Introduction

Reconstructing a person and an interacting object from a single RGB image is a fundamental problem in human-centric 3D understanding. It supports applications in immersive environments and content creation (Xie et al., 2026; Wen et al., 2025; Cao et al., 2021), embodied perception (Gavryushin et al., 2024), and robot learning (Hu et al., 2024; Wang et al., 2025a). Unlike reconstructing the two entities independently, human–object interaction (HOI) reconstruction must recover not only plausible human and object surfaces, but also a geometrically coherent joint configuration. This is particularly challenging under monocular depth ambiguity and mutual occlusion, where substantially different 3D arrangements may explain similar image evidence.

Existing methods reduce this ambiguity by introducing different forms of interaction and geometric reasoning. Optimization-based approaches constrain the joint spatial arrangement of independently estimated human and object models (Zhang et al., 2020). Learning-based methods instead encode interaction through joint representations or learned priors (Xie et al., 2022; Huo et al., 2023; Xie et al., 2024). More recent systems incorporate contact cues (Nam et al., 2024), topology-aware reasoning (Wang et al., 2025b), diffusion-based refinement (Li et al., 2025), or cross-view representations (Geng et al., 2026). These studies highlight the complementary roles of interaction and geometric cues in HOI reconstruction, which suggests that global human–object interaction captures long-range dependencies and topology preserves within-entity structure.

However, most existing decoders commit to a predefined reasoning structure throughout reconstruction. This can be restrictive for progressive decoding because the dominant source of uncertainty changes as the geometry is refined. At early stages, the model must infer a coarse joint configuration from incomplete and ambiguous image evidence. At later stages, the challenge shifts toward correcting local human–object relations, such as hand–object, foot–object, and body–support configurations, where small errors in relative position or orientation can lead to implausible contact even when the individual human and object shapes appear reasonable. Such errors are especially

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) Motivation of our SG-MoE (b) Complexity comparison

Figure 1: **Motivation and overview of SG-MoE.** (a) Existing HOI decoders typically apply a predefined reasoning structure throughout progressive reconstruction, although different stages may benefit from different geometric operations. SG-MoE instead organizes complementary operators—global interaction (GI), human topology (HT), object topology (OT), and bidirectional geometric interaction (BGI)—and learns their composition across decoding stages. (b) Complexity comparison with representative HOI reconstruction methods. SG-MoE uses fewer parameters and lower measured inference time than the listed baselines while achieving stronger reconstruction results.

pronounced in mutually occluded regions and from unseen viewpoints, where the missing geometry cannot be directly recovered from image evidence alone. As illustrated in Figure 1(a), applying the same reasoning pattern throughout decoding forces a fixed structure to serve these different reconstruction regimes. This motivates our central question: can a progressive HOI decoder learn how complementary geometric reasoning operations should be composed as reconstruction evolves?

We address this question with SG-MoE, a Structured Geometric Mixture-of-Experts decoder for progressive HOI reconstruction. Rather than treating experts as interchangeable subnetworks for increasing model capacity, we construct a small set of operators with explicit and complementary geometric roles. Global Interaction (GI) (Vaswani et al., 2017) models unconstrained dependencies between the human and object surfaces, while Human Topology (HT) and Object Topology (OT) (Wang et al., 2025b) propagate information according to their respective mesh structures. To explicitly model the evolving spatial relation between the two entities, we further introduce Bidirectional Geometric Interaction (BGI). The decoder contains six interaction blocks organized into three stages, and each block learns a dense composition of these structured operators. Thus, SG-MoE changes how complementary geometric reasoning is combined along the reconstruction trajectory, rather than enforcing one fixed reasoning pattern at every stage.

A key challenge is that the relative human–object geometry itself evolves during reconstruction. BGI therefore converts intermediate human–object displacements and distances into learned attention biases. It first updates human features using object features and then updates the object using the refined human representation, allowing information to propagate sequentially across the two surfaces. Intermediate coordinate supervision is applied throughout the decoder so that the evolving representation remains geometrically meaningful. This allows BGI to repeatedly recompute relative relations from progressively refined coordinates, directly coupling representation refinement, geometric estimation, and cross-entity interaction without requiring an additional contact prediction loss.

Experiments on BEHAVE (Bhatnagar et al., 2022) and InterCap (Huang et al., 2022) validate both the reconstruction performance and the proposed decoder design. SG-MoE reduces object Chamfer distance by 18.6% and 29.2% relative to the strongest compared object-reconstruction baseline on the two datasets, while also improving human reconstruction and contact precision, recall, and F1. As summarized in Figure 1(b), these gains are also obtained without increasing model size or measured inference cost relative to the representative compared methods.

In summary, the contributions of this work include:

- We formulate HOI reconstruction as the learned composition of complementary geometric reasoning operations and introduce SG-MoE, which combines global interaction, human and object topology, and relative geometry across progressive decoding stages.

- We introduce a Bidirectional Geometric Interaction expert that converts evolving intermediate coordinates into geometric attention biases and performs sequential human–object refinement, coupling geometric estimation with cross-entity feature interaction.
- Experimental results on both BEHAVE and InterCap show that our model achieves more accurate human and object reconstruction while also better recovering human–object contact, outperforming state-of-the-art baselines.

2 Related Work

Monocular human–object reconstruction. Monocular HOI reconstruction aims to recover both the individual human and object geometry and their coherent relative configuration from a single image. Early approaches reduce monocular ambiguity through explicit interaction constraints. PHOSA (Zhang et al., 2020) jointly optimizes the spatial arrangement of independently estimated human and object models, while CHORE (Xie et al., 2022) represents human–object interaction implicitly and StackFLOW (Huo et al., 2023) introduces a learned interaction prior. More recent methods exploit additional structural cues, including contact information (Nam et al., 2024), topology-aware reasoning (Wang et al., 2025b), diffusion-based refinement (Li et al., 2025), and novel-view evidence (Geng et al., 2026). These methods demonstrate that interaction-aware geometric cues are essential for recovering plausible HOI configurations. Our work takes a complementary perspective: rather than introducing a single additional cue, we study how multiple geometric reasoning operations should be organized and combined throughout progressive reconstruction.

Topology- and geometry-aware reasoning. Attention and graph-based operators provide complementary ways to model 3D structure. METRO (Lin et al., 2021b) captures long-range dependencies among mesh tokens using self-attention, while MeshGraphormer (Lin et al., 2021a) combines global attention with graph convolution to preserve local mesh connectivity. HOI-TG (Wang et al., 2025b) extends topology-aware graph reasoning to joint human–object reconstruction. These approaches motivate the use of both global interaction and entity-specific structural priors. Contact-based methods such as CONTHO (Nam et al., 2024) explicitly model interaction regions, while ScoreHOI (Li et al., 2025) incorporates physical constraints during diffusion-based refinement and CrossHOI (Geng et al., 2026) leverages cross-view representations to improve geometric consistency. In contrast, BGI operates directly within the progressive decoder by incorporating evolving relative geometry into the feature interaction process.

Mixture-of-experts and adaptive composition. Mixture-of-experts models combine multiple expert transformations using learned routing weights (Yuksel et al., 2012). Sparse MoE architectures are commonly used to increase model capacity while restricting the number of active experts (Jiang et al., 2024; Dai et al., 2024), and have also been explored in vision (Riquelme et al., 2021), 3D, and multimodal learning (Gao et al., 2026; Wang et al., 2024; Lin et al., 2026; Yang et al., 2026; Han et al., 2025). In these settings, experts are often architecturally similar and routing mainly determines which subnetworks are activated. Different from these approaches. SG-MoE uses a small set of heterogeneous experts with explicit geometric roles: global human–object interaction, human topology, object topology, and relative geometry. All experts remain active, and their outputs are densely combined across progressive decoding blocks. Thus, our goal is not sparse computation, but learning how complementary geometric operations should be composed as reconstruction evolves.

3 Method

3.1 Overview and Representation

Given an RGB image depicting a human–object interaction (HOI) event, the goal of HOI reconstruction is to recover the human mesh and estimate the object's rotation and translation relative to its template. As illustrated in Figure 2, we propose SG-MoE, a progressive reconstruction framework that learns to compose complementary geometric reasoning operations throughout decoding. We first extract image features and initialize joint, human-vertex, and object queries from predefined templates. These tokens are then progressively refined by a multi-stage decoder composed of a sequence of SG-MoE blocks. Within each block, global human–object interaction, human topology, object topology, and bidirectional relative geometry are modeled by four structured experts and

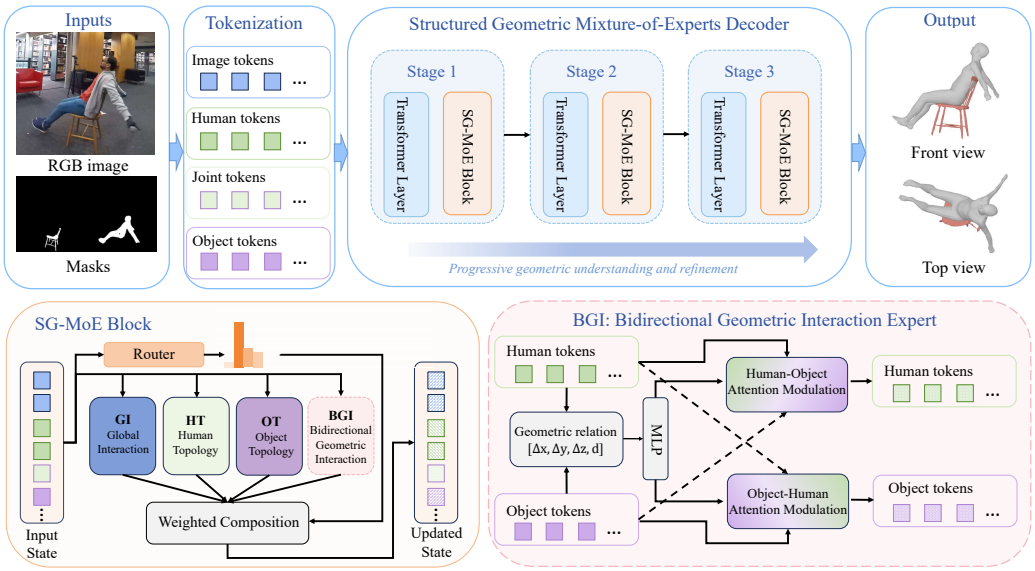


Figure 2: Overview of SG-MoE. The framework progressively refines human and object representations through structured geometric experts, with learned expert fusion and bidirectional geometric interaction.

combined through learned dense fusion. The final refined tokens are used to predict the human mesh and recover the object pose.

Input tokens. We follow the image and mask preprocessing of previous works (Nam et al., 2024; Wang et al., 2025b; Li et al., 2025) and use a ResNet-50 encoder (He et al., 2016; Moon et al., 2022) to extract visual features. These features are converted into image tokens and also provide global visual context for the geometric queries. The geometric queries are initialized from predefined human and object templates (Nam et al., 2024; Wang et al., 2025b; Li et al., 2025). Specifically, the joint queries are taken from the canonical SMPL-H joint template (Loper et al., 2015), while the human-vertex queries are obtained by downsampling the SMPL-H mesh to 431 vertices. The object queries are similarly initialized by downsampling the corresponding canonical object template to 64 template points. Together with the image tokens, the input sequence therefore contains 73 joint tokens, 431 human-vertex tokens, and 64 object-point tokens. Each geometric token is initialized with its canonical 3D coordinate and associated visual feature. An initial linear projection maps all tokens to a 1024-dimensional representation, after which learned positional embeddings are added to encode token identity and position. We denote the resulting initial sequence by T^0.

Geometric representation. During progressive decoding, the human and object tokens jointly encode image features and intermediate 3D geometry. We reserve the last three channels of each human and object token as coordinate estimates, which are supervised at the output of every interaction block. These coordinates are also used by BGI to construct relative geometric relations within each block. In this way, the evolving token representation is continuously tied to explicit 3D geometry, allowing geometric reasoning to be updated together with feature refinement throughout the decoder.

3.2 Progressive Composition of Structured Geometric Experts

Progressive HOI reconstruction requires geometric reasoning at different structural scopes. Global interaction is needed to establish dependencies between distant human and object regions, while topology-aware propagation preserves the intrinsic structure of each entity. As reconstruction proceeds, the evolving relative geometry further provides explicit cues about how the two surfaces should be arranged. Rather than hard-wiring these operations into a fixed decoding sequence, we treat them as complementary geometric experts and learn how their updates are composed throughout reconstruction.

For each interaction block b, a shared Transformer first contextualizes the complete token sequence, allowing image, joint, human, and object tokens to exchange information. We denote the resulting representation by S^b. All geometric experts operate on this common state and return outputs with identical token organization, which provides a shared interface for combining heterogeneous geometric transformations.

Global Interaction (GI). GI (Vaswani et al., 2017) global interaction between the two surfaces without imposing predefined mesh connectivity. It applies self-attention to the concatenated human and object tokens, allowing any human region to communicate with any object region according to feature affinity. This provides global cross-entity reasoning without imposing predefined mesh connectivity.

Human and Object Topology (HT/OT). HT propagates information over the downsampled SMPL-H graph, while OT operates on the graph induced by the corresponding object template, following the graph operators of HOI-TG (Wang et al., 2025b). The two experts encode entity-specific structural priors and complement the topology-free interaction.

Bidirectional Geometric Interaction (BGI). The above experts reason from features and intrinsic connectivity, but neither explicitly models the evolving relative 3D configuration between the two entities. BGI fills this role by using intermediate human–object geometry to guide cross-entity interaction. We describe its geometry-conditioned bidirectional update in Sec. 3.3.

Learned geometric composition. The four experts correspond to complementary forms of reasoning: global cross-entity interaction, human topology, object topology, and relative human–object geometry. Instead of imposing a fixed ordering or averaging their outputs uniformly, each interaction block predicts four normalized mixture weights from the pooled contextualized representation.

Let $\Delta_k^b = E_k^b(S^b) - S^b$ denote the update produced by expert k relative to the shared state. The fused representation is

$$T^b = S^b + \sum_{k \in \mathcal{K}} g_k^b \Delta_k^b, \tag{1}$$

where $\mathcal{K} = \{\mathrm{GI}, \mathrm{HT}, \mathrm{OT}, \mathrm{BGI}\}$ and $\sum_k g_k^b = 1$.

This formulation emphasizes that SG-MoE learns how to combine complementary geometric updates from a common reconstruction state, rather than following a predefined reasoning structure.

3.3 Bidirectional Geometric Interaction

While GI captures feature-based global interaction, it does not explicitly encode the evolving spatial relation between the reconstructed human and object. BGI addresses this limitation by turning intermediate relative geometry into a bias for cross-entity attention, allowing feature interaction to adapt to the current 3D configuration.

Let $H \in \mathbb{R}^{431 \times C}$ and $O \in \mathbb{R}^{64 \times C}$ denote the human and object tokens from the shared representation of a block. Their coordinate channels provide intermediate positions $\mathbf{x}_{h,i}$ and $\mathbf{x}_{o,j}$. For each human–object pair, we encode both relative direction and separation:

$$B_{ij} = \mathrm{MLP}([\,\mathbf{x}_{h,i} - \mathbf{x}_{o,j}; \|\mathbf{x}_{h,i} - \mathbf{x}_{o,j}\|_2\,]), \tag{2}$$

where $B \in \mathbb{R}^{431 \times 64}$ serves as a learned geometric bias. The displacement captures directional relation, while the distance explicitly measures spatial proximity.

BGI performs the interaction sequentially. It first updates the human representation using object features under the geometry-aware attention:

$$A_{h \leftarrow o} = \mathrm{softmax}_{\mathrm{row}}\left(\frac{Q_h K_o^\top}{\sqrt{C}} + B\right), \qquad H' = H + A_{h \leftarrow o} V_o. \tag{3}$$

The refined human representation is then used to update the object:

$$A_{o \leftarrow h} = \mathrm{softmax}_{\mathrm{row}}\left(\frac{Q_o K_{h'}^\top}{\sqrt{C}} + B^\top\right), \qquad O' = O + A_{o \leftarrow h} V_{h'}. \tag{4}$$

This sequential design differs from two independent cross-attention operations: the object is refined using a human representation that has already incorporated object information. As reconstruction progresses, each block recomputes the geometric bias from its updated intermediate coordinates, allowing relative geometry and feature interaction to evolve together. BGI therefore provides an explicit feedback mechanism between geometric estimation and cross-entity reasoning, without requiring contact supervision during training.

3.4 Reconstruction and Training

Prediction heads. The final decoder output is projected to 3D coordinates. Human tokens produce a coarse mesh with 431 vertices, which is progressively upsampled to 1723 and 6890 vertices using two multilayer perceptron layers. Object tokens predict 64 template-aligned 3D points, from which the object rotation and translation are recovered by correspondence-based alignment (Nam et al., 2024; Wang et al., 2025b; Li et al., 2025).

Progressive geometric supervision. A key role of the intermediate predictions is to keep the evolving token representation geometrically meaningful throughout decoding. Let X_h^b and X_o^b denote the coordinate channels of the human and object tokens after block b. We supervise both the final reconstruction and all intermediate block outputs:

$$\mathcal{L}_h = D_1(\widehat{V}_h, V_h^*) + D_1(\widehat{V}_{h,c}, V_{h,c}^*) + \sum_{b=1}^{N} D_1(X_h^b, V_{h,c}^*), \tag{5}$$

$$\mathcal{L}_o = D_1(\widehat{V}_o, V_o^*) + \sum_{b=1}^{N} D_1(X_o^b, V_o^*), \tag{6}$$

where $D_1(\cdot,\cdot)$ denotes the mean L1 coordinate error and $N = 6$ means the number of SG-MoE blocks. The intermediate terms directly constrain each refinement stage, ensuring that the geometry used by BGI remains tied to progressively improving 3D estimates rather than unconstrained latent features.

We additionally supervise the predicted vertices and regularize the reconstructed human mesh with an edge-length loss. The complete objective is

$$\mathcal{L} = \mathcal{L}_h + \mathcal{L}_o + \mathcal{L}_{\text{edge}}. \tag{7}$$

All terms are equally weighted. The full model is trained end-to-end using only reconstruction and surface supervision.

4 Experiments

4.1 Experimental Setup

Datasets and inputs. We evaluate on BEHAVE (Bhatnagar et al., 2022) and InterCap (Huang et al., 2022), following the train/test protocols used in template-based HOI reconstruction (Nam et al., 2024; Wang et al., 2025b; Li et al., 2025; Geng et al., 2026). We use the same dataset splits, input modalities, object templates, and evaluation protocol as these reference methods.

Metrics. Following prior HOI reconstruction work (Nam et al., 2024; Wang et al., 2025b; Li et al., 2025), we evaluate human and object geometry using Chamfer distance. After jointly aligning the predicted human and object meshes with the corresponding ground truth, we compute Chamfer distance separately for the reconstructed human surface and the object surface. All distances are reported in centimeters, with lower values indicating more accurate reconstruction.

Beyond surface reconstruction accuracy, we evaluate the reconstructed human–object interaction using contact precision, recall, and F1. Following prior work (Wang et al., 2025b), we construct the predicted contact map by marking human vertices whose minimum distance to the reconstructed object mesh is below $5\,$cm. We then compare the predicted contact map with the ground-truth annotations to compute contact precision (C_p) and recall (C_r), and report their harmonic mean as contact F1, i.e., $C_f = 2C_pC_r/(C_p + C_r)$. Higher values indicate more accurate recovery of the human–object contact configuration.

Method	BEHAVE					InterCap				
	$\mathrm{CD}_h \downarrow$	$\mathrm{CD}_o \downarrow$	$C_p \uparrow$	$C_r \uparrow$	$C_f \uparrow$	$\mathrm{CD}_h \downarrow$	$\mathrm{CD}_o \downarrow$	$C_p \uparrow$	$C_r \uparrow$	$C_f \uparrow$
METRO (Lin et al., 2021b)	46.82	59.13	0.084	0.520	0.145	42.83	82.12	0.049	0.641	0.091
MeshGraphormer (Lin et al., 2021a)	42.40	36.60	0.047	0.348	0.083	38.82	46.18	0.031	0.332	0.057
PHOSA (Zhang et al., 2020)	12.17	26.62	0.393	0.266	0.317	11.20	20.57	0.228	0.159	0.187
CHORE (Xie et al., 2022)	5.58	10.66	0.587	0.472	0.523	7.01	12.81	0.339	0.253	0.290
CONTHO (Nam et al., 2024)	4.99	8.42	0.628	0.496	0.554	5.96	9.50	0.661	0.432	0.523
HOI-TG (Wang et al., 2025b)	4.59	8.00	0.662	0.554	0.603	5.43	8.68	0.700	0.473	0.565
ScoreHOI (Li et al., 2025)	4.85	7.86	0.634	<u>0.586</u>	0.609	5.56	8.75	0.627	<u>0.590</u>	<u>0.608</u>
CrossHOI (Geng et al., 2026)	<u>4.27</u>	<u>7.68</u>	<u>0.687</u>	0.576	<u>0.627</u>	<u>5.17</u>	<u>8.38</u>	<u>0.724</u>	0.491	0.585
SG-MoE (Ours)	**4.14**	**6.25**	**0.717**	**0.608**	**0.658**	**4.27**	**5.93**	**0.738**	**0.624**	**0.676**

Table 1: Quantitative comparison on BEHAVE and InterCap. Best and second-best values among the compared methods are bold and underlined, respectively. Chamfer distances are in centimeters. Lower Chamfer distance and higher contact metrics are better.

Baselines. We report both general human-mesh reconstruction models and HOI-specific methods. METRO (Lin et al., 2021b) and MeshGraphormer (Lin et al., 2021a) represent global mesh-token reasoning and topology-aware mesh reasoning, respectively. PHOSA (Zhang et al., 2020) and CHORE (Xie et al., 2022) provide optimization-based and implicit interaction formulations, while CONTHO (Nam et al., 2024), HOI-TG (Wang et al., 2025b), ScoreHOI (Li et al., 2025), and CrossHOI (Geng et al., 2026) represent more recent interaction-aware systems.

4.2 Comparison with Prior Methods

Quantitative results. Table 1 compares SG-MoE with representative human and HOI reconstruction methods. SG-MoE achieves the best performance across all reported metrics on both BEHAVE and InterCap. In particular, it reduces object Chamfer distance by 18.6% on BEHAVE and 29.2% on InterCap relative to the strongest compared baselines, while also consistently improving human reconstruction. Importantly, the geometric improvements are accompanied by higher contact precision, recall, and F1 on both datasets. This indicates that SG-MoE not only reconstructs the individual human and object surfaces more accurately, but also better recovers their relative spatial configuration and contact relationships.

Notably, the largest improvements are observed in object reconstruction, together with consistent gains in contact quality. This suggests that combining global interaction, entity topology, and relative geometry is particularly beneficial for recovering the object pose and its spatial relation to the human. These contact improvements are obtained without explicit contact supervision during training.

Qualitative results. Figure 3 presents qualitative comparisons from both front and back views. The additional viewpoint helps expose errors in object orientation, relative placement, and contact that may be difficult to assess from the input image alone. Since checkpoints or executable code for HOI-TG and CrossHOI are not publicly available, we provide qualitative comparisons with CONTHO and ScoreHOI. Although the compared methods can recover individually plausible human and object shapes, they may still exhibit local misalignment around interaction regions, as highlighted by the red circles. In contrast, SG-MoE produces more coherent human–object configurations, particularly in occluded or unseen regions, as illustrated in (b), (d), and (f), which is consistent with its improved reconstruction and contact metrics. More results and dynamic results can be found in Appendix and supplementary video.

Model complexity and inference efficiency. We additionally compare model size and inference time with representative HOI reconstruction methods using the same machine, as summarized in Figure 1(b). SG-MoE contains 96.3M parameters, fewer than HOI-TG (122.8M) and substantially fewer than ScoreHOI (451.7M). It also achieves the lowest measured inference time among the compared methods, requiring 0.068 s per sample, compared with 0.085 s for CONTHO, 0.087 s for HOI-TG, and 14.52 s for ScoreHOI. These results indicate that the reconstruction gains of SG-MoE are not obtained by simply increasing model capacity or inference cost.

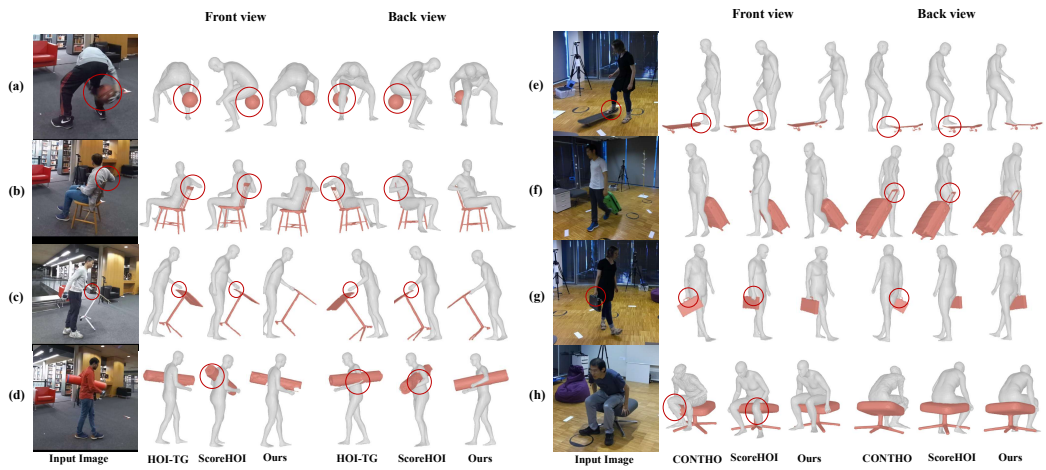


Figure 3: Qualitative comparison with HOI-TG, CONTHO, and ScoreHOI. Front and back views reveal differences in relative human–object placement and contact regions that may be less apparent from the input viewpoint.

Variant	BEHAVE					InterCap				
	$CD_h \downarrow$	$CD_o \downarrow$	$C_p \uparrow$	$C_r \uparrow$	$C_f \uparrow$	$CD_h \downarrow$	$CD_o \downarrow$	$C_p \uparrow$	$C_r \uparrow$	$C_f \uparrow$
Composition variants										
Uniform averaging	4.39	7.35	0.687	0.545	0.608	4.71	7.52	0.724	0.553	0.627
Top-2 routing	4.15	6.38	0.693	**0.612**	0.650	4.53	7.17	0.732	0.565	0.638
Static learned weights	4.26	6.93	0.679	0.588	0.630	4.59	7.62	0.712	0.547	0.619
Removing the mixture block										
w/o SG-MoE	4.52	8.06	0.665	0.487	0.562	5.52	8.60	0.695	0.501	0.582
Removing individual experts										
w/o GI	4.15	7.11	0.679	0.563	0.616	4.49	7.09	0.713	0.574	0.636
w/o HT	4.20	7.43	0.668	0.557	0.607	4.60	7.05	0.696	0.584	0.635
w/o OT	4.22	7.12	0.683	0.564	0.618	4.48	7.65	0.729	0.549	0.626
w/o BGI	4.25	7.20	0.673	0.579	0.622	4.56	6.82	0.704	0.608	0.652
Full SG-MoE	**4.14**	**6.25**	**0.717**	0.608	**0.658**	**4.27**	**5.93**	**0.738**	**0.624**	**0.676**

Table 2: Composition and expert ablations. Bold indicates the best value in each column. Chamfer distances are in centimeters. All variants are independently trained using the same training function and hyperparameters.

4.3 Ablation Studies

Table 2 evaluates the expert composition strategy and the contribution of each geometric expert. All variants are trained independently using the same training procedure and hyperparameters.

Effect of expert composition. We compare four strategies for combining the structured experts. Uniform averaging assigns equal weights to all experts, while static learned weights learn a fixed mixture for each decoding block without conditioning on its current representation. Top-2 routing performs token-wise sparse selection by retaining the two experts with the largest routing weights. The full SG-MoE instead uses dense block-wise composition predicted from the contextualized representation.

The full model consistently outperforms uniform averaging and static learned weights on both datasets. In particular, compared with uniform averaging, object Chamfer distance decreases from 7.35 to 6.25 on BEHAVE and from 7.52 to 5.93 on InterCap, together with clear gains in contact quality. Its consistent advantage over static learned weights further indicates that simply assigning

a fixed expert preference to each decoder block is insufficient; conditioning the composition on the evolving reconstruction representation provides additional benefit.

Dense composition also performs better overall than Top-2 routing. Although Top-2 yields slightly higher contact recall on BEHAVE, the full model achieves lower reconstruction error and higher contact F1 on both datasets, with a particularly clear improvement on InterCap. These results suggest that the four structured experts are better treated as complementary geometric operations to be jointly composed, rather than sparsely selected as interchangeable subnetworks.

Effect of the SG-MoE block. Removing the entire structured expert mixture consistently degrades both reconstruction and contact quality on BEHAVE and InterCap. The degradation is especially pronounced for object reconstruction and contact F1, showing that the gains cannot be attributed solely to the shared Transformer backbone. Together with the composition study above, this result supports both the use of heterogeneous geometric experts and learning how to combine their updates during progressive reconstruction.

Contribution of individual experts. Removing any individual expert increases both human and object reconstruction error on the two datasets, confirming that the four reasoning mechanisms provide complementary information. Without GI, object reconstruction degrades substantially, highlighting the importance of topology-free global interaction between the two surfaces. Removing HT or OT also affects both entities rather than only the surface on which the corresponding graph operator acts, indicating that entity-specific structural reasoning contributes to the joint HOI configuration. Removing BGI consistently degrades human and object reconstruction as well as contact quality. In particular, object Chamfer distance increases from 6.25 to 7.20 on BEHAVE and from 5.93 to 6.82 on InterCap, while contact F1 also drops on both datasets. This supports explicitly modeling the evolving relative geometry in addition to feature-based interaction and intrinsic topology.

Overall, the ablations show that no single geometric operation is sufficient by itself. Global interaction, entity topology, and relative geometry contribute complementary inductive biases, while learned dense composition provides an effective way to integrate them throughout progressive reconstruction.

Routing across decoder stages. Figure 4 shows mean weights for the six blocks grouped into three stages. The earlier blocks distribute more weight across the experts, including a larger object-topology contribution, whereas the later blocks emphasize human topology. This pattern shows that the learned composition differs with decoder depth. The weights describe the mixture coefficients; the size of an expert's geometric update also depends on its output features. The removal experiments provide the corresponding evidence for its contribution to reconstruction accuracy.

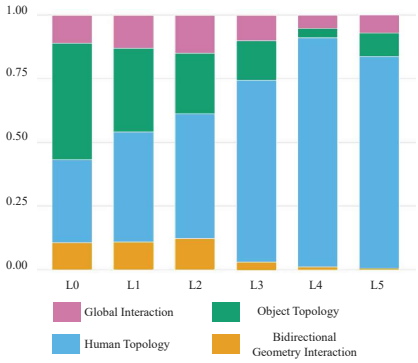


Figure 4: Mean expert weights across the six interaction blocks on BEHAVE. L0–L1, L2–L3, and L4–L5 belong to the three decoding stages, respectively.

5 Conclusion

We presented SG-MoE, a progressive HOI decoder that learns how to combine complementary geometric operations. Its structured experts cover global interaction, entity topology, and relative human–object geometry. BGI connects intermediate coordinates to sequential cross-entity attention, and supervision at every block ties the evolving representation to reconstruction. Across BEHAVE and InterCap, learned dense fusion improves over uniform averaging and yields lower reconstruction errors and higher contact scores than the compared methods. These results support learned composition of geometric operators as a useful design for joint human–object reconstruction.

AI USE STATEMENT

Generative AI assistance was used to edit the manuscript for clarity and organization, check consistency between the supplied implementation and method description, check experimental tables against their discussion, and provide feedback on paper pre-review. The authors are responsible for verifying the final text, equations, citations, and reported results.

ETHICS STATEMENT

This work uses existing human–object interaction datasets, including BEHAVE and InterCap, and does not involve the collection of new human-subject data. We follow the original dataset protocols and use the data solely for research purposes. Since monocular human–object reconstruction operates on images containing people and estimates detailed 3D human geometry, such techniques may raise privacy concerns if applied to personal imagery without appropriate consent.

REPRODUCIBILITY STATEMENT

Section 3 describes the progressive decoder, structured geometric experts, learned expert composition, bidirectional geometric interaction, and training objective. Section 4.1 provides the dataset, evaluation, and training settings, while Section 4.3 details the evaluated composition and expert variants. The appendix includes additional architectural specifications, tensor dimensions, loss definitions, and evaluation details.

REFERENCES

Bharat Lal Bhatnagar, Xianghui Xie, Ilya A Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Behave: Dataset and method for tracking human object interactions. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 15935–15946, 2022.

Zhe Cao, Ilija Radosavovic, Angjoo Kanazawa, and Jitendra Malik. Reconstructing hand-object interactions in the wild. In *Int. Conf. Comput. Vis.*, pp. 12397–12406. IEEE, 2021.

Damai Dai, Chengqi Deng, Chenggang Zhao, RX Xu, Huazuo Gao, Deli Chen, Jiashi Li, Wangding Zeng, Xingkai Yu, Yu Wu, et al. Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 1280–1297, 2024.

Jingnan Gao, Zhe Wang, Xianze Fang, Xingyu Ren, Zhuo Chen, Shengqi Liu, Yuhao Cheng, Jiangjing Lyu, Xiaokang Yang, and Yichao Yan. More: 3d visual geometry reconstruction meets mixture-of-experts. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 14680–14691, 2026.

Alexey Gavryushin, Yifei Liu, Daoji Huang, Yen-Ling Kuo, Julien Valentin, Luc Van Gool, Otmar Hilliges, and Xi Wang. Romeo: Revisiting optimization methods for reconstructing 3d human-object interaction models from images. In *Eur. Conf. Comput. Vis.*, pp. 71–87. Springer, 2024.

Pei Geng, Shanshan Zhang, and Jian Yang. Crosshoi: Learning cross-view representations for monocular 3d human-object interaction reconstruction. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 7121–7130, 2026.

Xumeng Han, Longhui Wei, Zhiyang Dou, Yingfei Sun, Zhenjun Han, and Qi Tian. Vimoe: An empirical study of designing vision mixture-of-experts. *IEEE Trans. Image Process.*, 34:7209–7221, 2025.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 770–778, 2016.

Haoyu Hu, Xinyu Yi, Zhe Cao, Jun-Hai Yong, and Feng Xu. Hand-object interaction controller (hoic): Deep reinforcement learning for reconstructing interactions with physics. In *Acm siggraph 2024 conference papers*, pp. 1–10, 2024.

Yinghao Huang, Omid Taheri, Michael J Black, and Dimitrios Tzionas. Intercap: Joint markerless 3d tracking of humans and objects in interaction. In *DAGM German Conference on Pattern Recognition*, pp. 281–299. Springer, 2022.

Chaofan Huo, Ye Shi, Yuexin Ma, Lan Xu, Jingyi Yu, and Jingya Wang. Stackflow: monocular human-object reconstruction by stacked normalizing flow with offset. In *IJCAI*, pp. 902–910, 2023.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.

Keller Jordan, Yuchen Jin, Vlado Boza, You Jiacheng, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks, 2024. URL `https://kellerjordan.github.io/posts/muon/`.

Ao Li, Jinpeng Liu, Yixuan Zhu, and Yansong Tang. Scorehoi: Physically plausible reconstruction of human-object interaction via score-guided diffusion. In *Int. Conf. Comput. Vis.*, pp. 7592–7602, 2025.

Bin Lin, Zhenyu Tang, Yang Ye, Jinfa Huang, Junwu Zhang, Yatian Pang, Peng Jin, Munan Ning, Jiebo Luo, and Li Yuan. Moe-llava: Mixture of experts for large vision-language models. *IEEE Trans. Multimedia*, 2026.

Kevin Lin, Lijuan Wang, and Zicheng Liu. Mesh graphormer. In *Int. Conf. Comput. Vis.*, pp. 12939–12948, 2021a.

Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 1954–1963, 2021b.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, October 2015.

Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. Accurate 3d hand pose estimation for whole-body 3d human mesh estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2308–2317, 2022.

Hyeongjin Nam, Daniel Sungho Jung, Gyeongsik Moon, and Kyoung Mu Lee. Joint reconstruction of 3d human and object via contact-based refinement transformer. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 10218–10227, 2024.

Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel Keysers, and Neil Houlsby. Scaling vision with sparse mixture of experts. *Adv. Neural Inform. Process. Syst.*, 34:8583–8595, 2021.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Adv. Neural Inform. Process. Syst.*, 30, 2017.

Guoan Wang, Jin Ye, Junlong Cheng, Tianbin Li, Zhaolin Chen, Jianfei Cai, Junjun He, and Bohan Zhuang. Sam-med3d-moe: Towards a non-forgetting segment anything model via mixture of experts for 3d medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 552–561. Springer, 2024.

Ruiyan Wang, Lin Zuo, Zonghao Lin, Qiang Wang, Zhengxue Cheng, Rong Xie, Jun Ling, and Li Song. Pa-hoi: A physics-aware human and object interaction dataset. In *ACM Int. Conf. Multimedia*, pp. 12769–12775, 2025a.

Zhenrong Wang, Qi Zheng, Sihan Ma, Maosheng Ye, Yibing Zhan, and Dongjiang Li. End-to-end hoi reconstruction transformer with graph-based encoding. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 27706–27715, 2025b.

Boran Wen, Dingbang Huang, Zichen Zhang, Jiahong Zhou, Jianbin Deng, Jingyu Gong, Yulong Chen, Lizhuang Ma, and Yong-Lu Li. Reconstructing in-the-wild open-vocabulary human-object interactions. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 17426–17436, 2025.

Xianghui Xie, Bharat Lal Bhatnagar, and Gerard Pons-Moll. Chore: Contact, human and object reconstruction from a single rgb image. In *Eur. Conf. Comput. Vis.*, pp. 125–145. Springer, 2022.

Xianghui Xie, Bharat Lal Bhatnagar, Jan Eric Lenssen, and Gerard Pons-Moll. Template free reconstruction of human-object interaction with procedural interaction generation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 10003–10015, 2024.

Xianghui Xie, Bowen Wen, Yan Chang, Hesam Rabeti, Jiefeng Li, Ye Yuan, Gerard Pons-Moll, and Stan Birchfield. Cari4d: Category agnostic 4d reconstruction of human-object interaction. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 14006–14016, 2026.

Zhenjie Yang, Yilin Chai, Xiaosong Jia, Qifeng Li, Yuqian Shao, Xuekai Zhu, Haisheng Su, and Junchi Yan. Drivemoe: Mixture-of-experts for vision-language-action model in end-to-end autonomous driving. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 10678–10688, 2026.

Seniha Esen Yuksel, Joseph N Wilson, and Paul D Gader. Twenty years of mixture of experts. *IEEE transactions on neural networks and learning systems*, 23(8):1177–1193, 2012.

Jason Y Zhang, Sam Pepose, Hanbyul Joo, Deva Ramanan, Jitendra Malik, and Angjoo Kanazawa. Perceiving 3d human-object spatial arrangements from a single image in the wild. In *Eur. Conf. Comput. Vis.*, pp. 34–51. Springer, 2020.

A Implementation Details

A.1 Decoder Configuration

The decoder contains two independently parameterized interaction blocks at each of three feature widths. Each block consists of a shared Transformer encoder layer, four structured geometric experts, and a lightweight router for dense expert composition. Table 3 summarizes the main architectural settings. The human and object graph residual operators follow HOI-TG (Wang et al., 2025b), together with its image and mask preprocessing protocol.

Component	Configuration
Visual encoder	ResNet-50; Hand4Whole initialization
Decoder widths	$1024, 512, 256$
Interaction blocks	Two per stage; six in total
Shared Transformer	Four heads; FFN width equal to feature width
Transformer dropout	0
Human tokens	431 coarse vertices and 73 joints
Object tokens	64 template-aligned points
Router	Mean pooling, LayerNorm, affine projection, softmax
Final coordinate projection	Linear map from 256 to 3
Human upsampling	Linear maps: $431 \to 1723 \to 6890$

Table 3: Decoder configuration used on both benchmarks.

Coordinate representation. The initial feature–coordinate representation is projected to 1024 dimensions before adding learned positional embeddings. At stage transitions, linear projections map the complete token representation to the next feature width. Within each interaction block, BGI reads the last three channels of the human and object tokens from the shared Transformer output as intermediate coordinates. Intermediate geometric supervision is instead applied to the corresponding channels after expert composition. Thus, the geometry used for interaction is produced by the current contextualized representation, while supervision constrains the refined block output. The final prediction head separately projects all 256 feature channels to 3D coordinates.

Expert interfaces. All four experts receive the same shared representation and return sequences with identical shape and token ordering. GI updates the concatenated human and object tokens through self-attention. HT and OT apply the corresponding graph residual operators to the human and object subsets, respectively. BGI sequentially updates the human and object tokens as described in Section 3.3. Tokens outside an expert's scope are retained unchanged, allowing the four outputs to be directly combined by the learned mixture weights.

Router implementation. For block b, the shared token representation S^b is mean-pooled over the token dimension. The pooled feature is normalized by LayerNorm and projected to four logits, followed by softmax:

$$g^b = \mathrm{softmax}\big(W_g^b \,\mathrm{LN}\left(\,\mathrm{Pool}(S^b)\right) + a_g^b\big)\,. \tag{8}$$

Each block has an independent router, and the resulting four weights are shared across token positions within that block. All four experts are evaluated in the full model.

Geometric-bias network. For each human–object pair, BGI concatenates the 3D relative displacement $\mathbf{r}_{ij}$ and its Euclidean distance d_{ij}. A two-layer MLP with GELU activation maps the resulting four-dimensional relation to a scalar bias:

$$B_{ij} = W_2\,\mathrm{GELU}\left(W_1[\mathbf{r}_{ij}; d_{ij}] + a_1\right) + a_2. \tag{9}$$

The hidden dimensions are 256, 128, and 64 for decoder widths 1024, 512, and 256, respectively. The resulting bias matrix has size 431×64 for each example. Query, key, and value projections retain the full feature width C, and both cross-attention directions use the standard $1/\sqrt{C}$ scaling. The reverse update uses the refined human representation together with the transpose of the geometric bias computed before the two sequential updates.

B Training and Evaluation Details

B.1 Loss Reduction

All coordinate losses use mean L1 reduction. For two corresponding point sets with n points,

$$D_1(A, B) = \frac{1}{3n} \sum_{i=1}^{n} \sum_{c=1}^{3} |A_{ic} - B_{ic}|. \tag{10}$$

The loss is further averaged across the minibatch. When joint annotations contain validity indicators, both the predicted and target coordinates are masked before reduction; the denominator remains the total number of tensor elements, matching the implementation.

The full-resolution human mesh, coarse human mesh, object points, and each intermediate coordinate prediction are assigned unit weight. Intermediate losses are summed over the six interaction blocks. The joint-coordinate and human edge-length losses also use unit weights. For the edge regularizer, the absolute error of the three edge lengths of each triangle is averaged over all faces and minibatch examples.

The complete objective is given in Eq. 7. Intermediate coordinate supervision is included through the human and object reconstruction terms and directly constrains the geometry produced at each progressive refinement block.

B.2 Implementation

We use Moun Jordan et al. (2024) and train for 50 epochs with batch size 16 on one NVIDIA RTX 4090D GPU. The learning rate starts at 10^{-4} and decreases by a factor of 0.1 after epoch 30. The three stages contain two blocks each, with independent expert and router parameters. All ablation variants are trained independently with the same training function and hyperparameters.

B.3 Geometric Evaluation

Joint alignment and Chamfer distance. For Chamfer-distance evaluation, the predicted human and object vertices are first concatenated and jointly aligned with their corresponding ground-truth vertices using a single similarity transform. The least-squares alignment estimates scale, rotation, and translation jointly for the complete human–object configuration. The aligned meshes are then separated, and 6000 surface points are sampled independently from the predicted and ground-truth human and object meshes.

For sampled point sets P and Q, we compute

$$\mathrm{CD}(P, Q) = \frac{1}{|P|} \sum_{p \in P} \min_{q \in Q} \|p - q\|_2 + \frac{1}{|Q|} \sum_{q \in Q} \min_{p \in P} \|q - p\|_2. \tag{11}$$

The two directional terms are summed without an additional factor of $1/2$. Distances are multiplied by 100 and reported in centimeters. This evaluation alignment is applied only for Chamfer-distance computation and is independent of the correspondence-based object-pose recovery used by the model.

Contact reconstruction. The reconstructed object mesh is obtained by applying the predicted pose to the complete canonical object template. A human vertex is classified as being in contact when its minimum distance to the reconstructed object mesh is below the threshold τ:

$$\widehat{c}_i = \mathbf{1}\left[\min_{v \in \widehat{V}_{o,\mathrm{mesh}}} \|\widehat{\mathbf{v}}_{h,i} - v\|_2 < \tau \right]. \tag{12}$$

where $\tau = 5\,\mathrm{cm}$ follows the HOI-TG evaluation protocol (Wang et al., 2025b). Contact metrics are computed on the reconstructed meshes before the similarity alignment used for Chamfer distance.

For each example, precision is computed over predicted contact vertices and recall over ground-truth contact vertices. Examples with no predicted contacts are excluded from precision aggregation,

while examples with no ground-truth contacts are excluded from recall aggregation. We report contact F1 as the harmonic mean of the aggregate precision and recall values, $C_f = 2C_pC_r/(C_p + C_r)$.

C Composition Variants

We provide implementation details for the expert-composition variants evaluated in Table 2.

Uniform averaging. All four expert outputs are retained and assigned the fixed weight $1/4$ in every interaction block.

Static learned weights. Each interaction block contains four learnable logits that are independent of the input representation. A softmax produces one normalized expert-weight vector per block. This variant preserves block-specific expert preferences while removing representation-conditioned composition.

Top-2 routing. The independently trained Top-2 variant predicts expert logits separately for every token. Let p_{nk} denote the softmax weight of expert k for token n, and let $\mathcal{I}_n$ denote the two experts with the largest weights. The retained weights are

$$\widetilde{p}_{nk} = \begin{cases} p_{nk} / \max\left(\sum_{j\in\mathcal{I}_n} p_{nj}, 10^{-6}\right), & k \in \mathcal{I}_n, \\ 0, & k \notin \mathcal{I}_n. \end{cases} \tag{13}$$

Thus, Top-2 differs from the full model in both routing granularity and the number of active expert contributions.

Full dense composition. The full SG-MoE predicts one normalized expert-weight vector for each interaction block from the mean-pooled shared representation, as described above. All four structured experts remain active and their outputs are densely combined.

Expert-removal variants. The *w/o SG-MoE* variant removes all four expert transformations and their fusion while retaining the shared Transformer blocks and the remaining reconstruction pipeline. Individual-expert ablations remove only the specified expert and renormalize the learned mixture over the remaining experts. All variants are independently trained using the same optimization settings and hyperparameters. Removing an expert changes both the available geometric operation and the corresponding model capacity.

D More comparison results

Figure 5 and Figure 6 present qualitative comparisons with state-of-the-art methods on BEHAVE and InterCap, respectively. The results show that SG-MoE can recover accurate human and object geometry in both visible and occluded regions, while also producing more coherent contact configurations across diverse HOI scenarios. In particular, the reconstructed interactions remain plausible even in regions that are weakly observed or fully occluded from the input view. These qualitative observations are consistent with the design of our structured geometric experts and BGI, which jointly model complementary geometric cues and relative human–object interactions. Additional dynamic results are provided in the supplementary video.

E Limitations

Figure 7 illustrates representative failure cases of the current method. SG-MoE may struggle with unusual human poses or rare human–object interaction patterns that are underrepresented in the training data, leading to inaccurate geometry or implausible relative configurations. This suggests that the model still has limited generalization to uncommon interaction patterns. Future work could incorporate stronger human and object priors, or broader interaction priors, to improve robustness in such cases.

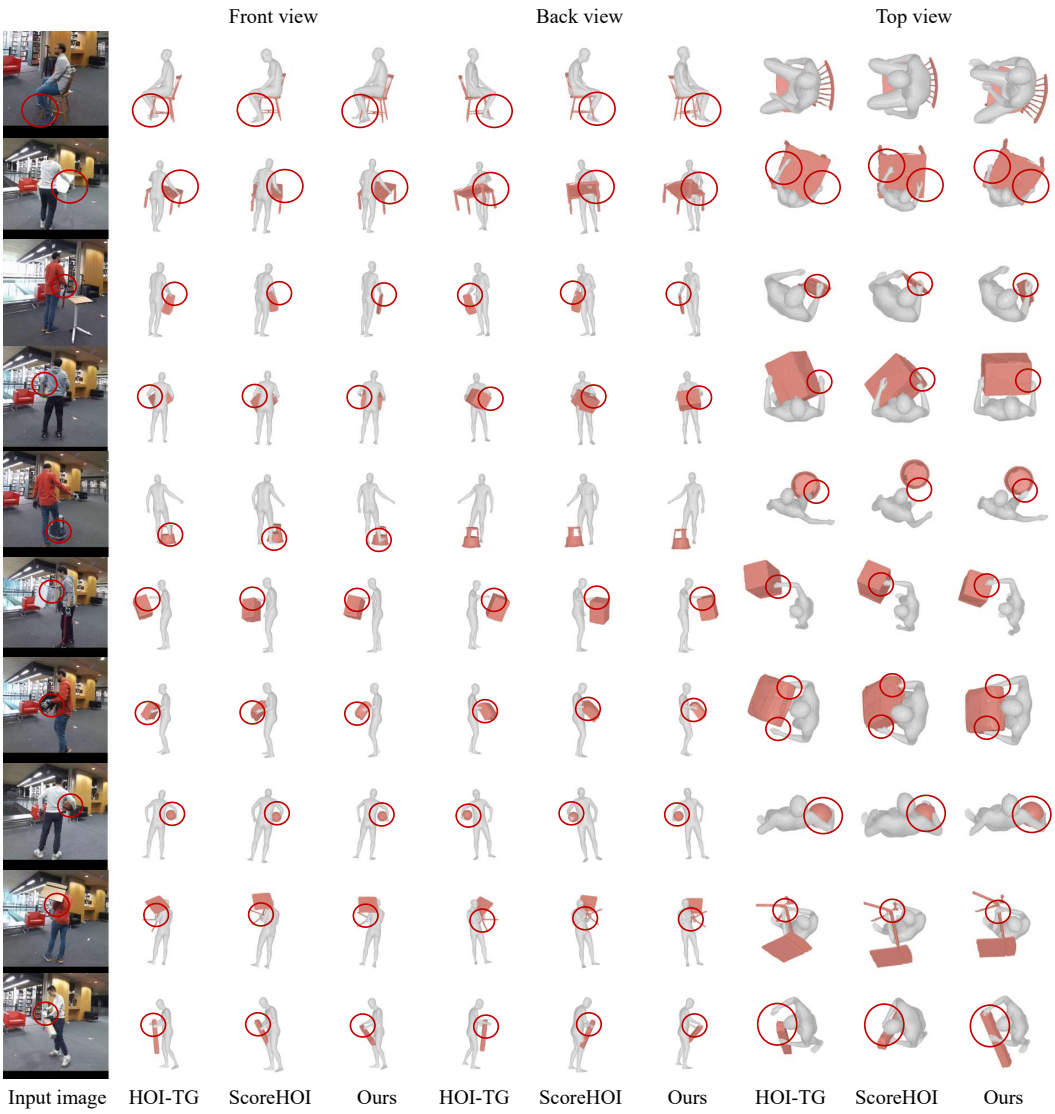


Figure 5: Qualitative results compared with HOI-TG and ScoreHOI on BEHAVE dataset. Please zoom in for details.

In addition, following prior methods (Nam et al., 2024; Wang et al., 2025b; Li et al., 2025), our approach relies on segmentation masks and a known object template. This requirement limits its applicability to unconstrained real-world settings, particularly for previously unseen or arbitrary objects. Future work will explore reducing these input assumptions and extending the framework to more general HOI scenarios.

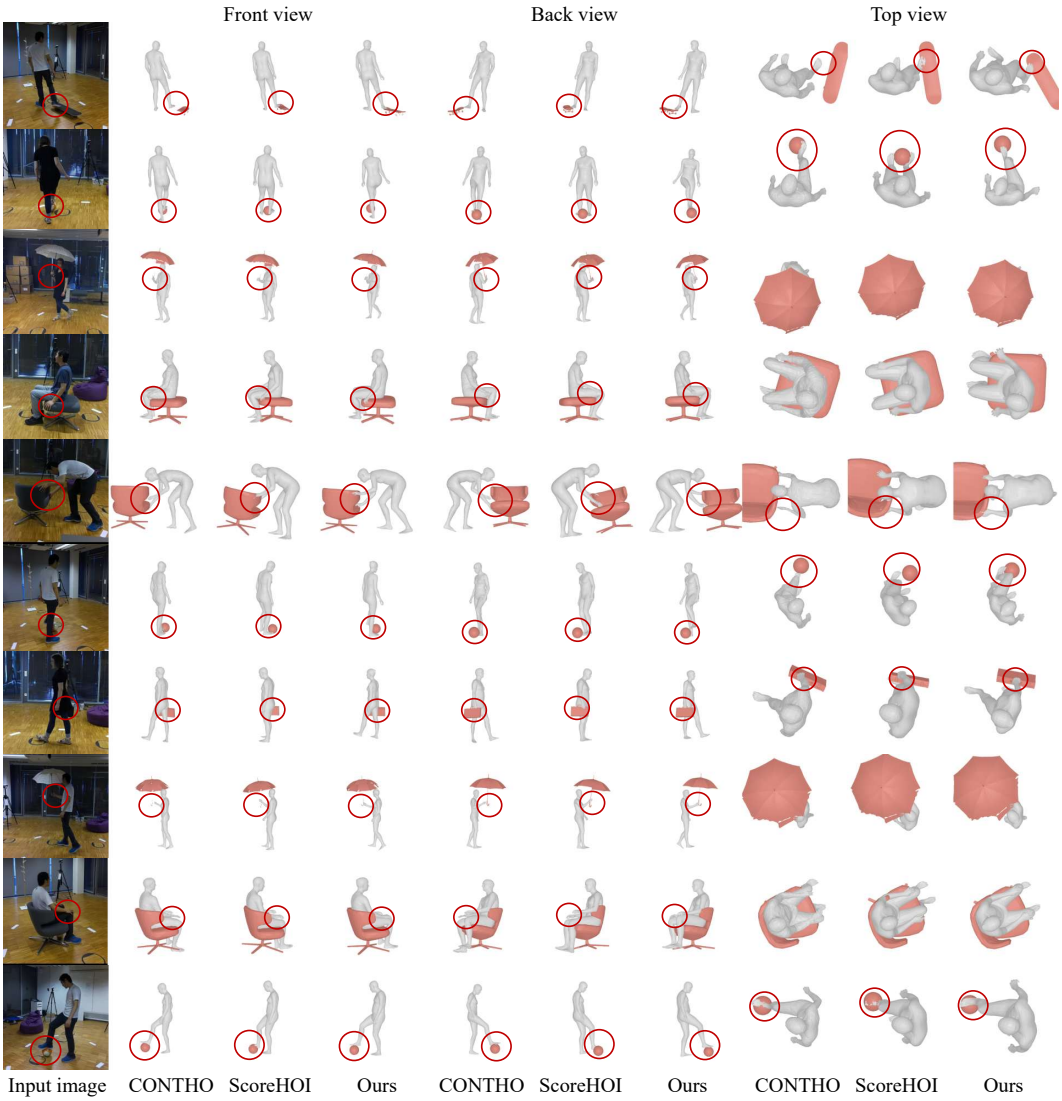


Figure 6: Qualitative results compared with CONTHO and ScoreHOI on InterCap dataset. Please zoom in for details.

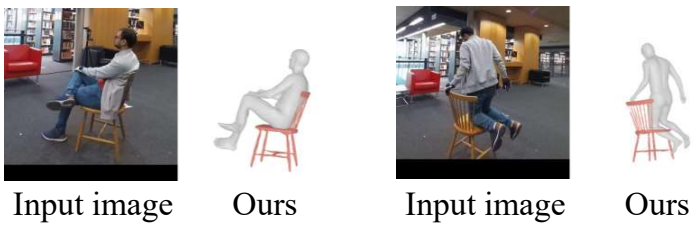


Figure 7: Failure cases of our method.