MoEG-HOI: Mixture of Expert Groups for One-Stage Hand-Object Interaction Motion Generation with Hand-Finger-Joint Semantic Guidance

Hang Xu[1], Yang Xiao[1*], Changlong Jiang[1], Haohong Kuang[1], Kaidi Zhang[1], Min Du[2], Ran Wang[3,4]

[1]National Key Laboratory of Multispectral Information Intelligent Processing Technology, School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan 430074, China
[2]ByteDance Inc., Beijing 100089, China
[3]School of Journalism and Information Communication, Huazhong University of Science and Technology
[4]School of Future Technology, Huazhong University of Science and Technology, Wuhan 430074, China
{hang_xu, Yang_Xiao, cljiang, haohong_kuang, kd_zhang, rex_wang}@hust.edu.cn, bingwen.ai@bytedance.com

Abstract

In this paper, MoEG-HOI is proposed as a novel method for the challenging 3D hand-object interaction (HOI) motion generation task, by introducing Mixture-of-Experts (MoE) to this field for the first time. Almost all the mainstream approaches in HOI motion generation leverage diffusion model as its strong generative ability. Nevertheless, due to HOI's fine-grained property, well training diffusion in one-stage way is actually not trivial. Existing state-of-the-art (SOTA) methods (e.g., Text2HOI and MF-MDM) alleviate this mainly via a coarse-to-fine, multi-stage paradigm. Although effective and practical, this paradigm prevents end-to-end training for optimal performance. In contrast, MoEG-HOI applies MoE to address this in one-stage way, with end-to-end training ability. This allows each expert to specialize in certain distinct HOI patterns, which alleviates individual expert's training difficulty. However, intuitively applying MoE is not optimal due to the issues of: (1) towards expert design, original MoE cannot well characterize hand's articulated structure at the levels of hand, finger, and joint explicitly, and (2) for expert routing mechanism, the characteristics of variational HOI action classes and diffusion noise levels have not been concerned. Towards the first problem, MoE's experts are designed into groups that correspond to motion generation for hand, finger, and joint respectively, under the semantic guidance from global to local. To facilitate this, HOI's text description will be correspondingly refined at Hand-Finger-Joint levels using LLM. Secondly, during MoE routing, the information of HOI's action label and diffusion noise level is concerned to select experts jointly, to better reveal actions' inter-class variation and dynamics of diffusion generation. SOTA performance on ARCTIC, GRAB and H2O datasets demonstrates the effectiveness of our method.

1 Introduction

Hand–object interaction (HOI) motion generation aims to synthesize natural, realistic and physically consistent 3D hand motions based on conditions such as textual descriptions (Christen et al. 2024; Cha et al. 2024; Huang et al. 2025) and given object trajectories (Zhan et al. 2024; Zhang et al. 2025b). High-quality HOI generation holds significant value for downstream applications in AR/VR, robotics, human-computer interaction, and embodied intelligence (Srivastava et al. 2022). As generative models advance, diffusion models (Ho, Jain, and Abbeel 2020; Song, Meng, and Ermon 2022) with a transformer-based architecture (Vaswani et al. 2017) have emerged as the mainstream approach in this field, owing to their powerful generative capabilities and advantages in modeling long-term sequences.

Although recent works (Li, Wu, and Liu 2023; Jiang et al. 2023; Zeng et al. 2025; Jiang et al. 2025; Xue et al. 2025) have achieved significant progress in modeling human-object interactions, generating realistic hand–object inter-

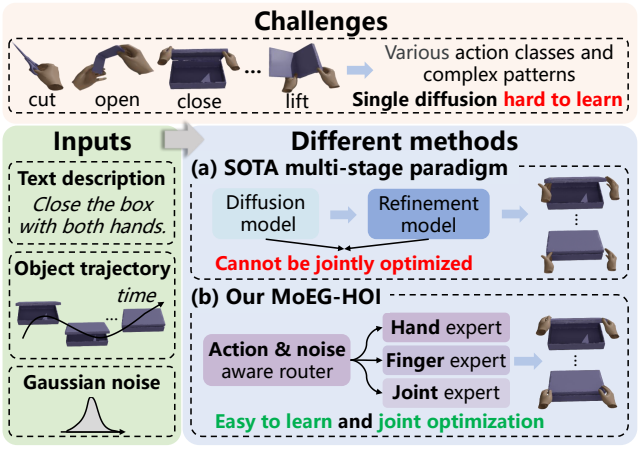

Figure 1: Our main research intuition and idea. Hand-object interaction (HOI) motion generation is challenging due to various action classes and complex patterns during interaction, making it difficult for a single diffusion model to learn. (a) SOTA methods adopt a multi-stage, coarse-to-fine refinement paradigm, but suffer from suboptimal optimization. (b) In contrast, our MoEG-HOI employs hierarchical expert groups with action and noise aware routers, alleviating the above problem with end-to-end training.

*Yang Xiao is corresponding author (Yang_Xiao@hust.edu.cn).
Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

action motions remains challenging due to the fine-grained and complex interaction patterns as shown in Fig. 1. This makes it non-trivial to properly train a single diffusion model in a one-stage manner to generate high-quality motions. To alleviate this, state-of-the-art (SOTA) methods, such as Text2HOI (Cha et al. 2024) and MF-MDM (Zhan et al. 2024), mainly adopt a coarse-to-fine, multi-stage paradigm, as shown in Fig. 1(a). These methods first generate a coarse motion via the diffusion model and subsequently use a separate refinement model to enhance realism. Despite its effectiveness, this multi-stage paradigm decouples the generation stage from the refinement stage. This separation prevents holistic, end-to-end optimization, as the generator receives no feedback from the subsequent refinement stage. This observation motivates us to explore an alternative direction: *is it possible to design a one-stage model powerful enough to directly generate high-fidelity HOI motions?*

Motivated by this consideration, we explore the potential of a one-stage paradigm. *We argue that a key limitation for one-stage methods is that the single denoising model struggles to learn complex HOI patterns.* To overcome this, we propose MoEG-HOI, a novel framework that introduces the Mixture-of-Experts (MoE) to this field for the first time, as shown in Fig. 1 (b). This allows specialized experts to collectively generate fine-grained motions, alleviating the learning burden on a single model and enabling end-to-end training. However, a naive application of existing MoE methods, such as that in (Fei et al. 2024), is not optimal for this task (e.g., FID of 0.971 vs. our 0.726). This is due to two key issues: (1) in terms of expert design, it cannot explicitly characterize the hand's articulated nature at the hand, finger, and joint levels; and (2) for the expert routing mechanism, the distinct properties of varying HOI action classes and different diffusion noise levels are not taken into account.

To address the first issue, MoEG-HOI introduces Hierarchical Semantics-guided Experts (HSE), a novel expert design guided by a global-to-local principle, where the experts are organized into three hierarchical groups: Hand, Finger, and Joint. This design explicitly mirrors the hand's natural decomposition, allowing the model to process global hand poses and local finger joint movements separately. To facilitate this, a Large Language Model (LLM) (Liu et al. 2025) is leveraged to refine input text into corresponding fine-grained semantic guidance for each of the three levels. Furthermore, recognizing that different HOI actions demand varying levels of complexity (e.g., "hold" versus "screw"), we also design hierarchically sized experts within each group. This design allows the model to adaptively match expert capacity to the action's intrinsic complexity.

To tackle the second issue, MoEG-HOI introduces a novel Action and Noise aware Router (ANR), which jointly considers two critical pieces of information: the HOI action label and the diffusion noise level. This makes the routing both action-aware, enabling it to capture distinct inter-class variations, and noise-aware, allowing it to dynamically adapt to the state of the denoising process at different timesteps.

MoEG-HOI is evaluated on three public datasets: ARCTIC (Fan et al. 2023), GRAB (Taheri et al. 2020), and H2O (Kwon et al. 2021). Extensive experimental results demonstrate the effectiveness of our proposed method.

Main contributions of this paper are outlined as follows:

- MoEG-HOI is proposed as a novel one-stage method that introduces Mixture-of-Experts to hand-object interaction motion generation task for the first time.
- A novel expert design comprising Hand-Finger-Joint expert groups with hierarchical sizing is proposed, allowing for specialized processing of both global hand poses and local finger joint movements.
- A joint routing mechanism based on action classes and noise levels is proposed, enabling experts to specialize in handling diverse samples, capturing both inter-class variations and dynamics of diffusion generation.

2 Related Work

Hand-Object Interaction Generation. Existing methods (Zhang et al. 2024b; Xu et al. 2024; Wei et al. 2024; Li et al. 2025, 2024; Wang et al. 2024; Zhong et al. 2025) explore static grasp generation based on object point clouds. Meanwhile, some works aim to generate dynamic hand-object interaction motions based on conditions such as text descriptions (Christen et al. 2024; Cha et al. 2024; Huang et al. 2025) and object motion trajectories (Zhang et al. 2025b,a), which mainly employ a diffusion-based model due to its strong generative ability. To generate high-quality interaction motions, a coarse-to-fine and multi-stage paradigm is generally employed. Text2HOI (Cha et al. 2024) decomposes the task into three stages: contact map generation, motion generation and refinement. Similarly, MF-MDM (Zhan et al. 2024) also employs a two-stage process of generation and refinement. Different from them, our work focuses on the one-stage paradigm with end-to-end training.

Mixture-of-Experts. Mixture-of-Experts (MoE) models utilize multiple specialized expert networks to adaptively select and process input with sparse activation (Shazeer et al. 2017), thereby enhancing performance across various tasks. Notably, the integration of MoE with the Transformer architecture (Fedus, Zoph, and Shazeer 2022) has yielded superior results, particularly in the field of LLMs such as DeepSeek-MoE (Dai et al. 2024). Some studies (Fei et al. 2024; Zhao et al. 2024; Yuan et al. 2025) also employ MoE in the diffusion model for image generation and show considerable effectiveness. However, intuitively applying them to the HOI motion generation task is not optimal due to that they fail to consider characteristics of this task. Our work leverages the explicit structure of the hand to divide MoE's experts into three groups at hand, finger, and joint level. Within each group, we employ hierarchically sized experts to adaptively process different interaction actions. During experts routing, our work uses action and noise aware routers to select the experts jointly, capturing the inter-class variations and dynamics of diffusion generation.

3 Method

3.1 Preliminarily

Task Setting and Data Representation. Given a textual task description T, the involved object point cloud P and its

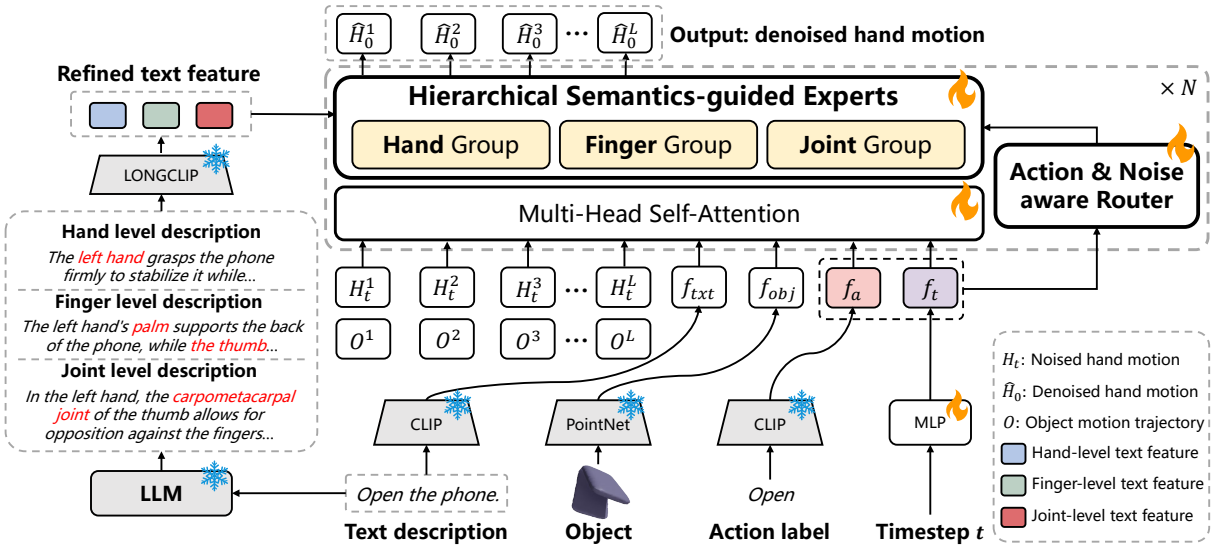

Figure 2: Overview of MoEG-HOI. MoEG-HOI is a diffusion-based model comprising N identical blocks, each integrating a multi-head self-attention layer and a Hierarchical Semantics-guided Experts (HSE) Module with Action and Noise aware Routers (ANR). The HSE module organizes experts into three hierarchical levels—Hand, Finger, and Joint —which are guided by corresponding text descriptions refined by an LLM. The ANR uses the action label and diffusion timestep to dynamically select the most suitable experts for each expert group.

motion trajectory O, our goal is to generate a corresponding 3D hand motion sequence H of length L, which can accurately reflect the intent of text description and exhibit precise coordination with the dynamics of the object.

The hand motion sequence is represented as $H \in \mathbb{R}^{L\times 99}$, where L=150 denotes the sequence length and each 99-dimensional vector is formed by flattening and concatenating the 3D hand translation parameters $t_h \in \mathbb{R}^3$ and MANO (Romero, Tzionas, and Black 2017) pose parameters $\theta_h \in \mathbb{R}^{16\times 6}$. The object trajectory is denoted as $O \in \mathbb{R}^{L\times d_o}$, where each d_o-dimensional vector comprises the concatenated 3D object translation $t_o \in \mathbb{R}^3$ and rotation $r_o \in \mathbb{R}^6$. For articulated objects, d_o is extended to 10 by incorporating an additional object articulation angle $\alpha \in \mathbb{R}^1$.

Diffusion Models. We adopt a transformer-based diffusion model for generating HOI motions. In the forward process, Gaussian noise is progressively added to the hand motion until reaching pure noise at the timestep T:

$$H_t = \sqrt{\bar{\alpha}_t} H_0 + \sqrt{1-\bar{\alpha}_t}\epsilon_t, \tag{1}$$

where t is the diffusion timestep, H_0 is the clean hand motion and H_t is the noised hand motion at timestep t. $\bar{\alpha}_t = \prod_{s=1}^{t}(1-\beta_s)$ is the cumulative noise schedule. In the reverse process, a denoise network iteratively removes noise from the data, resulting in the generation of HOI motions.

3.2 Overview of Our MoEG-HOI

An overview of our method is illustrated in Fig. 2. MoEG-HOI is a transformer-based diffusion model with mixture of experts groups that predicts denoised hand motion $\hat{H}_0$ from noised input H_t. The model takes several conditioning inputs: object geometry features f_{obj} from a pointcloud encoder (provided by Text2HOI), text features f_{txt}, action label features f_a from pre-trained CLIP (Radford et al. 2021), and timestep embeddings f_t from an MLP. The noised hand motion H_t and object trajectory O are processed through linear layers and concatenated by frames. All of the above are projected to D dimensions via linear layers and concatenated as input to MoEG-HOI.

MoEG-HOI consists of N identical blocks, each containing a multi-head self-attention and a Hierarchical Semantics-guided Experts (HSE) Module with Action and Noise aware Routers (ANR). The HSE organizes experts into three independent groups—Hand, Finger, and Joint—to directly mirror the hand's inherent kinematic structure. To provide semantic guidance for these hierarchical expert groups, we introduce level-specific semantic features generated by an LLM refiner, which uses Deepseek-V3 (Liu et al. 2025) to refine coarse-grained input text into three fine-grained descriptions targeting hand, finger, and joint levels respectively. A pre-trained Long-CLIP model (Zhang et al. 2024a) then encodes these descriptions into features f_{hand}, f_{finger} and f_{joint} for the HSE. Within each block, the ANR uses action labels and noise timesteps as routing conditions to generate expert activation weights, computing the final output as a weighted sum of all activated expert outputs.

3.3 LLM Refiner

To provide semantic guidance for the HSE, we utilize a carefully designed prompt to guide LLM to refine coarse text into detailed hierarchical descriptions. Our prompt design follows three core principles as follows. First, the hierarchical decomposition principle requires refining the description into Hand, Finger, and Joint levels to directly match the ex-

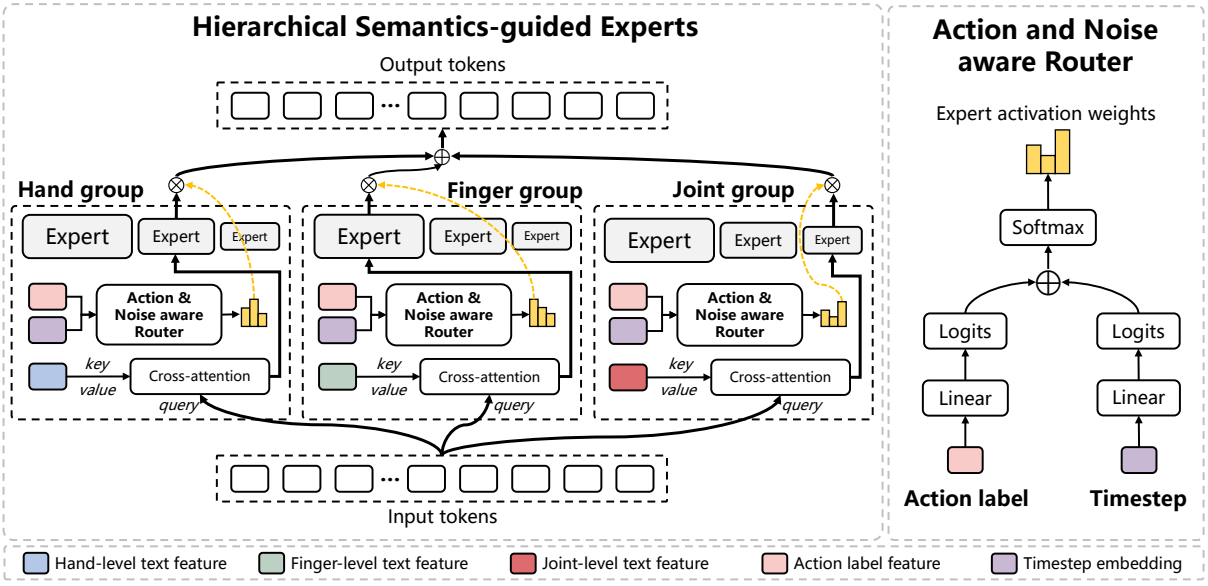

Figure 3: Structure of the Hierarchical Semantics-guided Experts Module (left) and the Action and Noise aware Router Module (right). HSE (left) consists of three expert groups (Hand, Finger, and Joint), each using a cross-attention block to fuse its corresponding level-specific text feature into the input tokens. Each expert group comprises multiple hierarchically sized experts, activated jointly by the action label feature and timestep embedding via the ANR module. ANR (right) combines signals from the action label and timestep to generate the expert activation weights.

pert group structure of the HSE module. Then, the coarse-to-fine principle ensures that the description progresses from overall poses to local details, following a natural motion logic. Finally, the interaction specificity principle demands clarification of the specific object parts being interacted with and the coordination between hands, providing key information for generating physically accurate motions. More details are provided in the Supplementary Materials.

3.4 Hierarchical Semantics-guided Experts

To address the failure of prior flat MoE architectures to explicitly capture the hand's kinematic hierarchy, we propose the Hierarchical Semantics-guided Experts (HSE) module. By organizing experts into Hand, Finger, and Joint groups that mirror this natural decomposition, HSE enables more effective and specialized modeling of complex HOI motion.

As illustrated in Fig. 3 (left), the HSE consists of three expert groups, each containing a cross-attention block to inject the semantic guidance and M heterogeneous experts E_m ($m \in [1, M]$), implemented as two-layer MLPs with progressively decreasing hidden dimensions, i.e., $d_m = d_{max}/2^{(m-1)}$, to promote diversity in capacity across the hierarchy. HSE processes two inputs: a token sequence $z \in \mathbb{R}^{(L+L_c)\times D}$ where L_c denotes the number of condition tokens, and three level-specific text features f_i, where $i \in \{\text{hand}, \text{finger}, \text{joint}\}$, provided by the LLM refiner.

The internal process of HSE involves two main steps. First, within each expert group, the cross-attention uses input tokens z as queries to interact with corresponding text features f_i, producing semantically guided tokens z_i':

$$z_i' = \text{CrossAttn}(q = z, k = f_i, v = f_i), \tag{2}$$

Secondly, the Action and Noise Aware Router (ANR) (detailed in Sec. 3.5) generates activation weights $w_{i,m}$ for expert E_m within each expert group based on the action label feature f_a and the timestep embedding f_t:

$$w_{i,m} = \text{ANR}_i(f_a, f_t). \tag{3}$$

To ensure sparse expert activation and that only the most relevant Top-K experts contribute to the output, the weights of experts outside the Top-K are set to zero:

$$w_{i,k} = \begin{cases} w_{i,k}, & \text{if } w_{i,k} \in TopK(w_i) \\ 0, & \text{otherwise} \end{cases}, \tag{4}$$

with $k \in [1, K]$. Subsequently, these non-zero weights are normalized across all activated experts from three groups:

$$w'_{i,k} = \frac{w_{i,k}}{\sum_{i=1}^{3}\sum_{k=1}^{K} w_{i,k}}. \tag{5}$$

Finally, the module's output y is computed as the weighted sum over all activated experts' outputs

$$y = \sum_{i=1}^{3}\sum_{k=1}^{K} w'_{i,k} \cdot E_{i,k}(z'_i). \tag{6}$$

3.5 Action and Noise Aware Router

To effectively model the diverse and fine-grained patterns of HOI, we consider two key factors: the variations across action classes and the dynamics of diffusion generation. Based on this, we design the Action and Noise Aware Router (ANR), shown in Fig. 3 (right), which leverages both action label and diffusion timestep to guide the expert activation.

To facilitate this, a naive approach is to directly concatenate the action label feature f_a and the timestep embedding f_t as input to a shared routing layer. However, the disparity in magnitude, structure, and informativeness often causes the routing layer to over-rely on one feature when concatenated directly. To mitigate this issue, we adopt a decoupled encoding strategy, independently projecting f_a and f_t to expert logits via separate linear layers:

$$w_a = \text{Linear}(f_a), \quad w_t = \text{Linear}(f_t). \tag{7}$$

The outputs of the two routing branches are then summed and passed through a softmax function to compute the expert activation weight $w_{i,m}$, which is subsequently used in Eq. 3:

$$w_{i,m} = Softmax(w_a + w_t). \tag{8}$$

This design ensures that both features contribute independently to expert selection, enabling the router to account for both inter-class variations and generation dynamics.

In addition, to encourage balanced expert utilization and avoid expert collapse (i.e., some experts never being activated), we adopt the load balance loss $\mathcal{L}_{balance}$ on the softmax-normalized logits from both routing branches:

$$\mathcal{L}_{balance} = \mathcal{L}(Softmax(w_a)) + \mathcal{L}(Softmax(w_t)), \tag{9}$$

where $\mathcal{L}(\cdot)$ denotes the prior balance loss (Fei et al. 2024).

3.6 Training Loss

Our model is trained using the following composite loss function, which jointly supervises motion fidelity, physical plausibility, interaction quality, and expert utilization.

For diffusion training, following (Tevet et al. 2022), our model directly estimates the clean signal. The training process optimizes from the reconstruction loss:

$$\mathcal{L}_{diff} = \mathbb{E}_{H_0,t} \left\| \hat{H}_\theta \left(H_t, t, c \right) - H_0 \right\|_2^2, \tag{10}$$

where H_0 and $\hat{H}_\theta$ denote the GT and predicted hand motion parameters, respectively. Our model also optimizes the positions of hand joints J and vertices V, which can be computed by the hand motion parameters via the MANO model:

$$\mathcal{L}_{rec} = \left\| \hat{J} - J \right\|_2^2 + \left\| \hat{V} - V \right\|_2^2. \tag{11}$$

To enhance the spatial alignment and physical plausibility of HOI, our model leverages a distance-based loss:

$$\begin{aligned} \mathcal{L}_{dist} = & \sum_{p \in P} \left\| d(p, \hat{v}) - d(p, v) \right\| + \\ & \lambda \sum_{v \in V} \left\| d(\hat{v}, p) - d(v, p) \right\|, \end{aligned} \tag{12}$$

where P are the object points and $d(\cdot)$ measures the distance from hand vertices to their nearest object points and from object points to their nearest hand vertices.

In addition, we optimize the denoised hand motion parameters to enforce contact with the object:

$$\mathcal{L}_{con} = \sum_{p \in P_{con}} \min_{v \in V} \left\| p - v \right\| + \sum_{v \in V_{con}} \min_{p \in P} \left\| v - p \right\|, \tag{13}$$

where P_{con} and V_{con} are the object points and hand vertices classified as in contact based on a distance threshold τ mm. Furthermore, we add a constraint to explicitly prevent hand-object penetration during interaction:

$$\mathcal{L}_{pen} = \sum_{p \in P_{pen}} \min_{v \in V} \left\| p - v \right\|, \tag{14}$$

where P_{pen} denotes object points inside the hand. We penalize the acceleration of hand vertices for smoothness:

$$\mathcal{L}_{acc} = \sum_{v \in V} \left\| v_i - 2 \cdot v_{i-1} + v_{i-2} \right\|. \tag{15}$$

Finally, considering the balance loss presented in Eq. 9, our total loss can be summarized as follows:

$$\begin{aligned} \mathcal{L}_{total} = & \mathcal{L}_{diff} + \lambda_1 \mathcal{L}_{rec} + \lambda_2 \mathcal{L}_{dist} + \lambda_3 \mathcal{L}_{con} \\ & + \lambda_4 \mathcal{L}_{pen} + \lambda_5 \mathcal{L}_{acc} + \lambda_6 \mathcal{L}_{balance}. \end{aligned} \tag{16}$$

4 Experiments

4.1 Experiment Settings

Datasets. Experiments are conducted on three public 3D hand-object interaction datasets: ARCTIC (Fan et al. 2023), GRAB (Taheri et al. 2020), and H2O (Kwon et al. 2021). For three datasets, the experimental setup follows the protocol of Text2HOI (Cha et al. 2024). We utilize subjects s1-s9 (excluding s3) for the training set and s10 for the test set on ARCTIC dataset. For the GRAB dataset, subjects s1-s9 are used for training and s10 for testing. For the H2O dataset, the training set comprises subjects s1-s2, while s3 is used for testing. For all three datasets, the corresponding text descriptions annotated by Text2HOI (Cha et al. 2024) are adopted.

Evaluation Metrics. We evaluate the methods on various metrics, each targeting a distinct aspect of generation quality. **Contact Ratio (CR)** measures the proportion of frames where the minimum hand-object distance is within a 5 mm threshold. **Penetration (Pen)** measures the percentage of frames exhibiting hand vertices penetration beyond a 1 cm threshold. **Power Spectrum KL divergence of Joints (PSKL-J)** reflects the smoothness of the generated motion, by comparing the acceleration distributions of predicted and GT motions, reporting results in both directions. **Fréchet Inception Distance (FID)** measures the distance of feature distributions between generated and ground-truth based on a feature extractor. **Mean Per Joint/Vertex Position Error (MPJPE and MPVPE)** calculate the average positional error (mm) for hand joints and vertices, respectively.

Baseline. Our experiments are performed on the following open-source methods. **Text2HOI** (Cha et al. 2024) generates HOI motions from text description, which employ three stages including contact map generation, motion generation and refinement. To adapt for our task setting, we change its conditions to both text description and object trajectory and retrain its generation and refinement stages. **MF-MDM** (Zhan et al. 2024) is a two-stage diffusion model designed to generate HOI motions from text description and object trajectory, which includes generation and refinement stages. **OMOMO** (Li, Wu, and Liu 2023) is a whole-body

Dataset	Method	CR↑	Pen↓	PSKL-J(gt.p.)↓	PSKL-J(p.gt.)↓	FID↓	MPJPE↓	MPVPE↓
ARCTIC	OMOMO (TOG'23)	0.221	**0.042**	0.4497	0.3970	0.824	318.60	318.32
	Text2HOI (CVPR'24)	0.665	0.163	0.3477	**0.3012**	0.962	116.37	116.13
	MF-MDM (CVPR'24)	0.848	0.170	**0.3020**	0.3140	0.903	101.44	101.26
	MoEG-HOI (ours)	**0.920**	0.253	0.3334	0.3640	**0.726**	**96.13**	**96.11**
GRAB	OMOMO (TOG'23)	0.102	**0.044**	2.4935	2.4563	0.655	420.40	419.41
	Text2HOI (CVPR'24)	0.554	0.258	0.7990	1.2046	0.690	167.84	166.92
	MF-MDM (CVPR'24)	0.434	0.209	0.8956	1.0573	1.362	225.23	225.03
	MoEG-HOI (ours)	**0.566**	0.264	**0.2243**	**0.2447**	**0.599**	**153.15**	**152.56**
H2O	OMOMO (TOG'23)	0.596	0.447	4.9839	4.9809	2.377	111.02	110.60
	Text2HOI (CVPR'24)	0.577	**0.412**	1.9346	1.8466	1.847	107.21	106.73
	MF-MDM (CVPR'24)	0.592	0.453	1.4101	1.5260	10.732	121.26	119.09
	MoEG-HOI (ours)	**0.777**	0.631	**0.8130**	**0.9807**	**1.648**	**101.05**	**100.40**

Table 1: Quantitative results on ARCTIC, GRAB and H2O datasets. MoEG-HOI achieves SOTA or highly competitive performance across all three datasets. The best results are in bold and the second best are underlined.

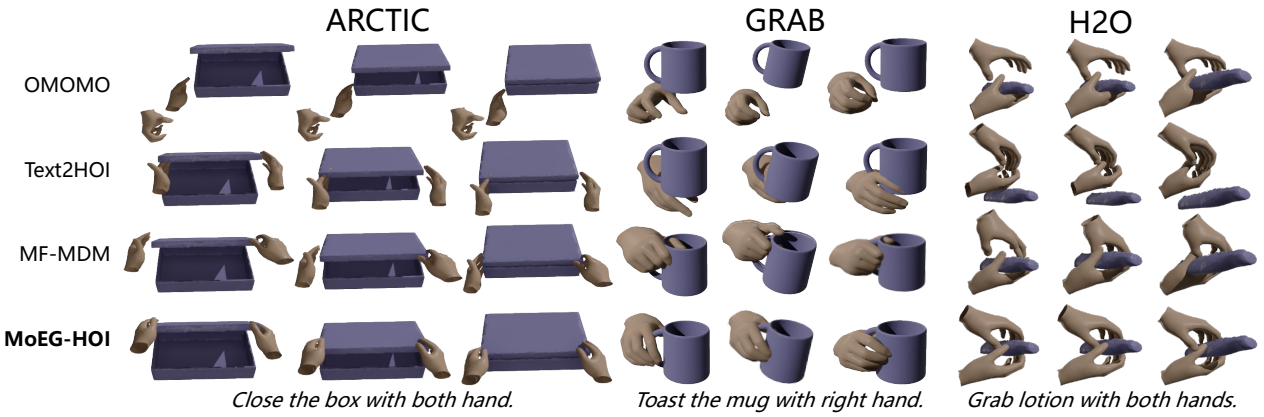

Figure 4: Qualitative Comparison on ARCTIC, GRAB and H2O datasets. Compared to three baseline methods, MoEG-HOI generates more physically plausible and natural motions that successfully fulfill the text description and object trajectory.

generation method conditioned on the object trajectory. We tailor its conditions to both text description and object trajectory. Following (Zhang et al. 2025b), we generate hand joints in its first stage and then generate hand parameters based on joints with contact constraints in its second stage.

Implementation Details. For MOEG-HOI, the number of blocks N is 8 and the latent dimension D is set to 512. Each expert group contains M=3 experts with a maximum dimension d_{max}=1024, and only $K = 1$ expert is activated. For the loss function, λ is set to 0.1 and τ is 10. λ_1, λ_2, λ_3, λ_4 and λ_5 are set to 20 and λ_6 is 0.01. The maximum diffusion step T is set to 1000 and the variances β_t vary from 0.0001 to 0.2. All models are trained on a single NVIDIA GeForce RTX 3090 GPU with the random seed 0. MOEG-HOI is trained on the ARCTIC, GRAB, and H2O datasets for 100, 100, and 200 epochs, respectively, using the AdamW optimizer with a learning rate of 0.0001 and a batchsize of 32.

4.2 Quantitative Results

Experimental results on ARCTIC, GRAB and H2O datasets are presented in Tab. 1. The findings demonstrate that the proposed MoEG-HOI achieves SOTA or highly competitive performance across all three datasets. On ARCTIC dataset, MoEG-HOI attains SOTA results in metrics such as CR (0.920), FID (0.726), and MPJPE (96.13). Although OMOMO achieves the lowest Pen (0.042), its CR is concurrently the lowest by a large margin (0.221). This indicates that its low penetration is achieved at the cost of not making effective contact with the object. The dominance of MoEG-HOI is further underscored on the GRAB and H2O datasets, where it surpasses all methods in nearly every metric. For instance, on GRAB it improves motion smoothness (PSKL-J) by nearly a factor of five (0.2447 vs. 1.0573), while on H2O it substantially boosts the Contact Ratio (CR) to 0.777 and drastically reduces FID from 10.732 (MF-MDM) to 1.648.

4.3 Qualitative Results

The qualitative comparison on the three datasets is provided in Fig. 4. In contrast to baseline methods, which suffer from clear artifacts such as failed contact (OMOMO), mesh penetration (MF-MDM in GRAB dataset), or unnatural poses (Text2HOI in ARCTIC dataset), our MoEG-HOI generates more physically plausible and natural motions that successfully fulfill the text description and object trajectory.

Method	CR↑	Pen↓	PSKL-J(gt.p.)↓	PSKL-J(p.gt.)↓	FID↓	MPJPE↓	MPVPE↓
w/o refine text	0.871	0.262	0.4206	0.3808	0.867	96.43	96.36
+ w/o expert group	0.879	0.252	0.4512	0.4678	0.893	97.87	97.65
+ w/o hierarchical size	0.874	0.248	0.3932	0.4205	0.937	99.54	99.36
w/o independent balance	0.833	0.250	0.4322	0.3801	0.793	103.13	103.07
w/o action router	0.903	0.280	0.4551	0.4710	0.974	103.62	103.44
w/o timestep router	0.820	**0.204**	0.4130	0.4456	0.820	104.39	104.25
DiT-MoE (Fei et al. 2024)	0.874	0.312	0.3455	0.3822	0.971	101.34	101.25
Ours	**0.920**	0.253	**0.3334**	**0.3640**	**0.726**	**96.13**	**96.11**

Table 2: Component ablation on ARCTIC dataset. For DiT-MoE, the expert number is 10, including 1 shared expert.

Method	CR↑	Pen↓	PSKL-J(gt.p.)↓	PSKL-J(p.gt.)↓	FID↓	MPJPE↓	MPVPE↓
d_{max}=512	0.879	**0.218**	0.4268	0.4491	0.910	105.34	105.42
d_{max}=2048	0.835	0.240	0.4727	0.4841	0.916	102.40	102.53
M=2	0.819	0.232	0.5061	0.5266	0.812	99.63	99.59
M=4	0.814	0.250	0.6624	0.6475	0.853	99.97	99.89
Ours (d_{max}=1024, M=3)	**0.920**	0.253	**0.3334**	**0.3640**	**0.726**	**96.13**	**96.11**

Table 3: Hyperparameters analysis on ARCTIC dataset. d_{max} and M respectively denote the maximum dimension of experts and the experts number in each expert group.

4.4 Ablation Study and Hyperparameter Analysis

As shown in Tab. 2, a series of ablation studies on the ARCTIC dataset systematically validates the effectiveness of the core designs in MoEG-HOI. First, the design of HSE is validated via a cumulative ablation strategy. The results demonstrate that sequentially removing the hierarchical semantic guidance (**w/o refine text**), the expert group structure (**+ w/o expert group**), and the hierarchical size (**+ w/o hierarchical size**) causes a continuous and steady decrease in key metrics such as FID. This outcome indicates that each design element within the HSE is crucial for the final performance.

Next, for the effectiveness of ANR, replacing the proposed decoupled encoding strategy with a naive feature concatenation (**w/o independent balance**) degrades performance, especially the CR (0.920 vs. 0.833). Furthermore, removing either the action-aware (**w/o action router**) or noise-aware (**w/o timestep router**) components results in a performance drop, causing the FID to worsen from 0.726 to 0.974 and 0.820, respectively. These results confirm the necessity of each component in the router design. In addition, Fig. 5 visualizes the learned expert activation on ARCTIC dataset. It clearly shows that different actions (rows) activate diverse combinations of experts, confirming action-aware specialization. Concurrently, the activation dynamically evolves as the diffusion timestep decreases, indicating noise-aware adaptation to the generation process.

Moreover, MoEG-HOI is compared against a generic baseline, DiT-MoE (Fei et al. 2024). The results show that our method achieves advantages in all metrics, especially in FID (0.726 vs. 0.971), which confirms MoEG-HOI is a better design for the HOI task compared to a generic MoE.

Finally, we analyze the key hyperparameters for the expert design, as shown in Tab. 3. The results indicate that varying either the number (M=2 or 4) or the maximum dimension (d_{max}=512 or 2048) of experts per group from our final setting (M=3, d_{max}=1024) leads to a general degradation in performance in almost all metrics, which validates that our current parameter is an effective configuration.

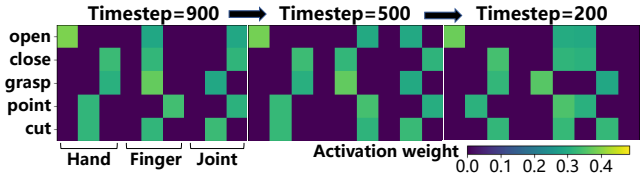

Figure 5: Visualization of expert activations on ARCTIC dataset. The activation patterns across different actions (rows) and their dynamics across diffusion timesteps provide clear visual evidence for the effectiveness of our ANR in learning specialized expert roles.

5 Conclusion and Future Work

In this paper, we propose MoEG-HOI, a novel one-stage framework that, for the first time, introduces the MoE architecture to 3D HOI motion generation. The proposed method features a Hierarchical Semantics-guided Experts that mirrors the hand's articulated structure and a dynamic Action and Noise Aware Router for specialized expert selection. Extensive experiments on three datasets demonstrate that our task-specific design can effectively tackle the fine-grained complexity of end-to-end HOI generation.

A limitation of the current method is its focus on generating single-action interactions. Future work will explore scaling our MoE approach to generate complex, long-horizon motion sequences, bridging the gap between single-action generation and compositional task execution.

Acknowledgements

This work is jointly supported by the National Natural Science Foundation of China under Grant No. 62271221, the National Social Science Foundation of China under Grant No. 25BXW041, and the Taihu Lake Innovation Fund for Future Technology, Huazhong University of Science and Technology (HUST) under Grant 2023-B-8.

References

Cha, J.; Kim, J.; Yoon, J. S.; and Baek, S. 2024. Text2hoi: Text-guided 3d motion generation for hand-object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1577–1585.

Christen, S.; Hampali, S.; Sener, F.; Remelli, E.; Hodan, T.; Sauser, E.; Ma, S.; and Tekin, B. 2024. Diffh2o: Diffusion-based synthesis of hand-object interactions from textual descriptions. In *Proceedings of the SIGGRAPH Asia Conference*, 1–11.

Dai, D.; Deng, C.; Zhao, C.; Xu, R. X.; Gao, H.; Chen, D.; Li, J.; Zeng, W.; Yu, X.; Wu, Y.; Xie, Z.; Li, Y. K.; Huang, P.; Luo, F.; Ruan, C.; Sui, Z.; and Liang, W. 2024. DeepSeekMoE: Towards Ultimate Expert Specialization in Mixture-of-Experts Language Models. arXiv:2401.06066.

Fan, Z.; Taheri, O.; Tzionas, D.; Kocabas, M.; Kaufmann, M.; Black, M. J.; and Hilliges, O. 2023. ARCTIC: A dataset for dexterous bimanual hand-object manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12943–12954.

Fedus, W.; Zoph, B.; and Shazeer, N. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120): 1–39.

Fei, Z.; Fan, M.; Yu, C.; Li, D.; and Huang, J. 2024. Scaling Diffusion Transformers to 16 Billion Parameters. arXiv:2407.11633.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. In *Proceedings of Advances in Neural Information Processing Systems*, 6840–6851.

Huang, M.; Chu, F.-J.; Tekin, B.; Liang, K. J.; Ma, H.; Wang, W.; Chen, X.; Gleize, P.; Xue, H.; Lyu, S.; et al. 2025. HOIGPT: Learning Long-Sequence Hand-Object Interaction with Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7136–7146.

Jiang, C.; Xiao, Y.; Wu, C.; Zhang, M.; Zheng, J.; Cao, Z.; and Zhou, J. T. 2023. A2j-transformer: Anchor-to-joint transformer network for 3d interacting hand pose estimation from a single rgb image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8846–8855.

Jiang, C.; Xiao, Y.; Zheng, J.; Kuang, H.; Wu, C.; Zhang, M.; Cao, Z.; Du, M.; Zhou, J. T.; and Yuan, J. 2025. 3D Hand Pose Estimation via Articulated Anchor-to-Joint 3D Local Regressors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. In press, doi: 10.1109/TPAMI.2025.3609907.

Kwon, T.; Tekin, B.; Stühmer, J.; Bogo, F.; and Pollefeys, M. 2021. H2o: Two hands manipulating objects for first person interaction recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10138–10148.

Li, H.; Mao, W.; Deng, W.; Meng, C.; Fan, H.; Wang, T.; Osamu, Y.; Tan, P.; Wang, H.; and Deng, X. 2025. MultiGraspLLM: A Multimodal LLM for Multi-Hand Semantic Guided Grasp Generation. arXiv:2412.08468.

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

Li, K.; Wang, J.; Yang, L.; Lu, C.; and Dai, B. 2024. Semgrasp: Semantic grasp generation via language aligned discretization. In *Proceedings of the European Conference on Computer Vision*, 109–127.

Liu, A.; Feng, B.; Xue, B.; Wang, B.; Wu, B.; Lu, C.; Zhao, C.; Deng, C.; Zhang, C.; Ruan, C.; et al. 2025. DeepSeek-V3 Technical Report. arXiv:2412.19437.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*, 8748–8763.

Romero, J.; Tzionas, D.; and Black, M. J. 2017. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics (TOG)*, 36(6): 1–17.

Shazeer, N.; Mirhoseini, A.; Maziarz, K.; Davis, A.; Le, Q.; Hinton, G.; and Dean, J. 2017. Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer. arXiv:1701.06538.

Song, J.; Meng, C.; and Ermon, S. 2022. Denoising Diffusion Implicit Models. arXiv:2010.02502.

Srivastava, S.; Li, C.; Lingelbach, M.; Martín-Martín, R.; Xia, F.; Vainio, K. E.; Lian, Z.; Gokmen, C.; Buch, S.; Liu, K.; et al. 2022. Behavior: Benchmark for everyday household activities in virtual, interactive, and ecological environments. In *Proceedings of the Conference on Robot Learning*, 477–490.

Taheri, O.; Ghorbani, N.; Black, M. J.; and Tzionas, D. 2020. GRAB: A dataset of whole-body human grasping of objects. In *Proceedings of the European Conference on Computer Vision*, 581–600.

Tevet, G.; Raab, S.; Gordon, B.; Shafir, Y.; Cohen-Or, D.; and Bermano, A. H. 2022. Human Motion Diffusion Model. arXiv:2209.14916.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Proceedings of Advances in Neural Information Processing Systems*, volume 30.

Wang, Y.-K.; Xing, C.; Wei, Y.-L.; Wu, X.-M.; and Zheng, W.-S. 2024. Single-view scene point cloud human grasp generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 831–841.

Wei, Y.-L.; Jiang, J.-J.; Xing, C.; Tan, X.-T.; Wu, X.-M.; Li, H.; Cutkosky, M.; and Zheng, W.-S. 2024. Grasp as you say:

Language-guided dexterous grasp generation. In *Proceedings of Advances in Neural Information Processing Systems*, 46881–46907.

Xu, G.-H.; Wei, Y.-L.; Zheng, D.; Wu, X.-M.; and Zheng, W.-S. 2024. Dexterous grasp transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17933–17942.

Xue, M.; Liu, Y.; Guo, L.; Huang, S.; and Ding, C. 2025. Guiding Human-Object Interactions with Rich Geometry and Relations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22714–22723.

Yuan, Y.; Wang, Z.; Huang, Z.; Zhu, D.; Zhou, X.; Yu, J.; and Min, Q. 2025. Expert Race: A Flexible Routing Strategy for Scaling Diffusion Transformer with Mixture of Experts. arXiv:2503.16057.

Zeng, L.-A.; Huang, G.; Wei, Y.-L.; Gu, S.; Tang, Y.-M.; Meng, J.; and Zheng, W.-S. 2025. Chainhoi: Joint-based kinematic chain modeling for human-object interaction generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12358–12369.

Zhan, X.; Yang, L.; Zhao, Y.; Mao, K.; Xu, H.; Lin, Z.; Li, K.; and Lu, C. 2024. Oakink2: A dataset of bimanual hands-object manipulation in complex task completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 445–456.

Zhang, B.; Zhang, P.; Dong, X.; Zang, Y.; and Wang, J. 2024a. Long-clip: Unlocking the long-text capability of clip. In *Proceedings of the European Conference on Computer Vision*, 310–325.

Zhang, J.; Zhang, Y.; An, L.; Li, M.; Zhang, H.; Hu, Z.; and Liu, Y. 2025a. Manidext: Hand-object manipulation synthesis via continuous correspondence embeddings and residual-guided diffusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. In press, doi: 10.1109/TPAMI.2025.3588302.

Zhang, W.; Dabral, R.; Golyanik, V.; Choutas, V.; Alvarado, E.; Beeler, T.; Habermann, M.; and Theobalt, C. 2025b. Bimart: A unified approach for the synthesis of 3d bimanual interaction with articulated objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 27694–27705.

Zhang, Z.; Wang, H.; Yu, Z.; Cheng, Y.; Yao, A.; and Chang, H. J. 2024b. Nl2contact: Natural language guided 3d hand-object contact modeling with diffusion model. In *Proceedings of the European Conference on Computer Vision*, 284–300.

Zhao, W.; Han, Y.; Tang, J.; Wang, K.; Song, Y.; Huang, G.; Wang, F.; and You, Y. 2024. Dynamic Diffusion Transformer. arXiv:2410.03456.

Zhong, Y.; Jiang, Q.; Yu, J.; and Ma, Y. 2025. Dexgrasp anything: Towards universal robotic dexterous grasping with physics awareness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22584–22594.