SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

May 2026, Vol. 69, Iss. 5, 150105:1–150105:15
https://doi.org/10.1007/s11432-025-4809-7

Special Topic: Large Multimodal Models

Multimodal prior-augmented text-driven 3D human-object interaction generation

Yin WANG[1], Ziyao ZHANG[1], Zhiying LENG[1], Haitian LIU[1], Frederick W.B. LI[2], Mu LI[1] & Xiaohui LIANG[1,3*]

[1]*State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing 100191, China*
[2]*Department of Computer Science, University of Durham, Durham DH1 3LE, UK*
[3]*Zhongguancun Laboratory, Beijing 100191, China*

Received 19 September 2025/Revised 5 January 2026/Accepted 11 February 2026/Published online 16 April 2026

Abstract We address the challenging task of text-driven 3D human-object interaction (HOI) motion generation. Existing methods primarily rely on a direct text-to-HOI mapping, which suffers from three key limitations due to the significant cross-modality gap: (Q1) sub-optimal human motion, (Q2) unnatural object motion, and (Q3) weak interaction between humans and objects. To address these challenges, we propose MP-HOI, a novel framework grounded in four core insights. (1) Multimodal data priors: We leverage multimodal data (text, image, pose/object) from large multimodal models as priors to guide HOI generation, which tackles (Q1) and (Q2) in data modeling. (2) Enhanced object representation: We improve existing object representations by incorporating geometric keypoints, contact features, and dynamic properties, enabling expressive object representations, which tackle (Q2) in data representation. (3) Multimodal-aware mixture-of-experts (MoE) model: We propose a modality-aware MoE model for an effective multimodal feature fusion paradigm, which tackles (Q1) and (Q2) in feature fusion. (4) Cascaded diffusion with interaction supervision: We design a cascaded diffusion framework that progressively refines human-object interaction features under dedicated supervision, which tackles (Q3) in interaction refinement. Comprehensive experiments demonstrate that MP-HOI outperforms existing approaches in generating high-fidelity and fine-grained HOI motions.

Keywords text-driven motion generation, human-object interaction, multimodal models, diffusion model

Citation Wang Y, Zhang Z Y, Leng Z Y, et al. Multimodal prior-augmented text-driven 3D human-object interaction generation. Sci China Inf Sci, 2026, 69(5): 150105, https://doi.org/10.1007/s11432-025-4809-7

1 Introduction

Humans continuously interact with surrounding objects in daily life, e.g., moving monitors onto desks, pushing suitcases to desired locations, or washing apples and placing them on plates. Each task requires precise interaction between human actions and object movements. Exploring human-object interaction motion generation holds significant importance due to its broad downstream applications in character animation, VR/AR content creation, and robotics [1–6]. As language serves as a natural interface for expressing interaction intentions, text-driven HOI motion generation has emerged as a promising research direction, aiming to generate 3D human-object interaction sequences guided by text prompts.

Existing work has explored text-driven HOI motion generation using diffusion models [7] guided by text [8, 9], object trajectories [10, 11], contact maps [12, 13], and other cues [14–16]. Despite progress, current methods remain limited to coarse-grained interaction motions, falling short in three key aspects: (Q1) Sub-optimal human motion. Existing methods, such as HIMO-Gen [8], typically take text and object geometry as input and generate human motion through simple condition concatenation. This insufficient modeling of conditions leads to sub-optimal human motion, resulting in low-quality sequences that are misaligned with the intended interactions. (Q2) Unnatural object motion. Prior studies [8, 10] often adopt a simplified object representation, namely 3D translation and 6D rotation (only 9 dimensions), which disregards the geometric structure of objects. Consequently, the generated object motions often appear unnatural, exhibiting floating or sliding artifacts. (Q3) Weak interaction motion. Methods such as CHOIS [11] employ a one-step and direct text-to-HOI mapping. However, given the complexity of human-object interactions, this approach struggles to capture fine-grained HOI dynamics, leading to unrealistic contact, interpenetration, or implausible interactions in the generated motions.

* Corresponding author (email: liang_xiaohui@buaa.edu.cn)

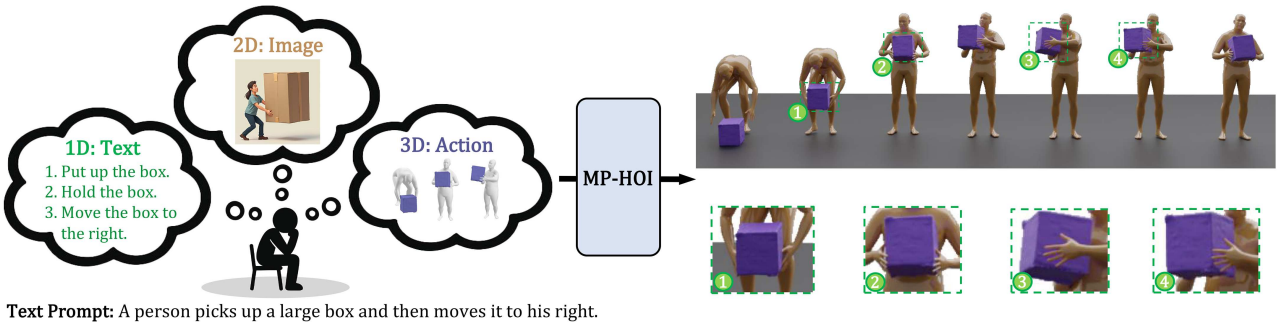

Figure 1 (Color online) MP-HOI excels in generating fine-grained human-object interaction motions from multimodal data priors, achieving both high-quality human-object interactions and precise text-motion alignment.

To address these challenges, our core insight for effective modeling of human-object interaction can be structured around four key innovations. **(1) Multimodal data priors—addressing (Q1) and (Q2) in data modeling.** We extract structured hierarchical priors from large multimodal models. Textual (1D): fine-grained descriptions derived from language parsing. Visual (2D): image-based motion references. Spatial (3D): human atomic actions and object geometric structure. These multimodal priors provide rich semantic guidance for generating both human and object motion. **(2) Enhanced object representation—addressing (Q2) in data representation.** We augment object representations with additional 3D mesh keypoints, contact information, and translational/angular velocities. By enriching objects with detailed geometric and dynamic attributes, we enable more precise and stable object motion generation. **(3) Multimodal-aware mixture-of-experts model—addressing (Q1) and (Q2) in feature fusion.** We propose a modality-aware MoE framework that selects specialized experts to enable effective multimodal interaction, which optimizes the multimodal feature fusion paradigm. **(4) Cascaded diffusion with interaction supervision—addressing (Q3) in interaction optimization.** We design a cascaded diffusion framework: human diffusion, object diffusion, and human-object interaction diffusion. Coarse-grained human and object motions are generated first to guide fine-grained interaction motion generation. Additional supervision losses further refine the interaction features.

We validate our approach through comprehensive experiments on two benchmark datasets: FullBodyManipulation (single-object interaction) [10] and HIMO (multi-object interaction) [8]. Results show that our method outperforms existing techniques, achieving a new state-of-the-art in text-driven HOI motion generation, as shown in Figure 1. Our contributions are summarized as follows.

- We introduce the leverage of multimodal priors—1D textual, 2D visual, and 3D spatial—to guide the fine-grained human-object interaction motion generation.
- We propose an enhanced object representation that incorporates geometric keypoints, contact features, and dynamic properties, enabling structurally rich and stable motion representation.
- We present a modality-aware mixture-of-experts model that optimizes the fusion paradigm for multimodal features.
- We design a cascaded diffusion framework that progressively refines human-object interaction features under dedicated supervision, achieving state-of-the-art performance on existing benchmarks.

2 Related work

2.1 Text-driven human motion generation

Three primary methodologies have emerged to tackle the challenge of text-driven human motion generation. (i) Latent space alignment [17–21] aims to learn a unified latent space between text and motion embeddings. (ii) Conditional autoregressive models [22–27] generate motion tokens sequentially by leveraging previous tokens and text. Recent advancements [28–30] utilize masked motion modeling to generate more natural movements. (iii) Conditional diffusion models [31–40], which learn probabilistic text-to-motion mappings within a conditional diffusion framework, have shown remarkable performance. While these advancements have propelled human motion generation forward, they predominantly center on individual motion generation, lacking the ability to generate interactive motions with external elements (e.g., objects).

2.2 Text-driven human-object interaction generation

The generation of human-object interactions has recently emerged as a promising research direction, attracting increasing attention from the community [12, 13, 41]. OMOMO [10] generates 3D human pose sequences based on the given motion of interacting 3D objects. GRAB [42] predicts 3D hand grasping poses for specific 3D object shapes, performing various grasping manipulations. InterDiff [43] develops a diffusion-based generative model that predicts future human-object collaborative motion from their 3D interaction history. CG-HOI [44] proposes a method to generate realistic 3D human-object interactions from text descriptions and given static object geometry. IMoS [45] synthesizes full-body human and 3D object motions from textual inputs, but focuses exclusively on small-object grasping. HIMO [8] introduces a large-scale motion capture dataset of humans interacting with multiple objects and develops a baseline model. However, current methods can only achieve coarse-grained human-object interactions, which primarily manifest in three limitations: sub-optimal human motion, unnatural object motion, and weak interaction motion. Therefore, exploring fine-grained human-object interaction generation remains a critical yet challenging problem.

2.3 Large model-assisted motion generation

In recent years, large language models (LLMs) such as BERT [46], GPT-4 [47], and T5 [48] have demonstrated remarkable capabilities in language tasks. Some studies have leveraged the strengths of LLMs to assist in motion generation tasks. For example, ActionGPT [49] utilizes GPT-3 to parse input text prompts into simple and long-form text prompts. SINC [50] also employs GPT-3.5 to establish relationships between motion and the human body. FineMoGen [38] uses LLMs to adjust text prompts according to user input requirements for motion editing tasks. Fg-T2M++ [39] leverages GPT-4 to parse text prompts into detailed prompts for body part joints and analyzes the keyword properties in the text prompts to enable fine-grained motion generation. Overall, current methods primarily utilize the text generation capability of LLMs to enhance the richness of textual prompts, thereby improving motion generation tasks. However, they only exploit the single-modality (text) prior knowledge of LLMs, lacking consideration for leveraging multimodal large models (e.g., incorporating image) to effectively guide motion generation tasks.

3 Preliminaries

Mixture-of-Experts [51,52] assigns specific tasks to specialized experts, each adept at handling a particular aspect of the problem. This approach is well-suited to our multimodal priors-augmented HOI task, where the fusion of multimodal content and motion features presents a dynamic and complex challenge. Mixture-of-Experts involves a gating network to activate distinct subsets of expert networks for different inputs, which mainly consists of two key components. (1) MoE layer: A MoE layer contains N experts (denoted as $e_i(\cdot), i = 1, 2, \ldots, N$). (2) Gating network: A router G routes the input token $\mathbf{x}$ to the most suitable top-k experts. Formally, given the input token $\mathbf{x}$, the output token $\mathbf{y}$ of the MoE layer is the weighted sum of outputs from the k activated experts:

$$\mathbf{y} = \sum_{i \in \mathcal{T}} g_i(\mathbf{x}) e_i(\mathbf{x}), \tag{1}$$

where $g(\mathbf{x}) = \sigma(\mathbf{W}\mathbf{x})$, in which $\mathbf{W}$ is the gate parameter matrix, σ is the softmax function, and $\mathcal{T}$ represents the set of the top-k indices.

Training MoE models directly often results in most tokens being assigned to a few experts, while others do not get sufficient training. Therefore, a load balancing loss [53, 54] is applied to ensure a balanced distribution of input tokens across the experts:

$$\mathcal{L}_{\mathrm{b}} = \sum_{i=1}^{N} f_i P_i, \tag{2}$$

where $f_i = \frac{1}{T_t} \sum_{t=1}^{T_t} \mathbf{1}$ (Token $\mathbf{x}_t$ selects Expert i), and f_i represents the fraction of tokens routed to expert i. P_i is the fraction of the router probability assigned to expert i, defined as $P_i = \frac{1}{T_t} \sum_{t=1}^{T_t} \mathrm{Softmax}\big(g(\mathbf{x}_t)\big)_i$, T_t denotes the number of tokens, N is the number of experts, and $\mathbf{1}(\cdot)$ denotes the indicator function.

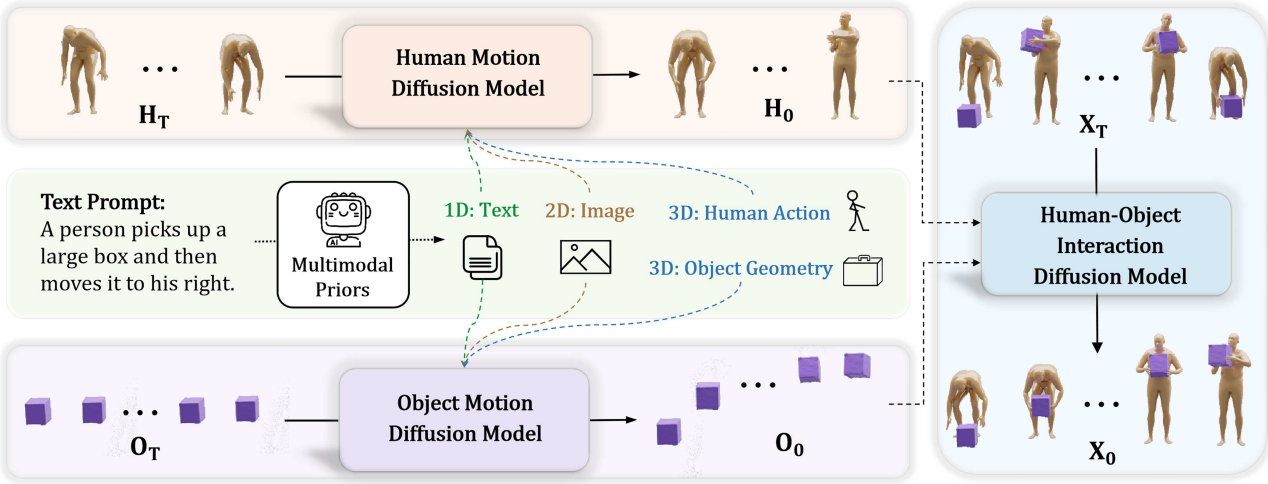

Figure 2 (Color online) Overview of MP-HOI. Given a text prompt and multimodal priors, the reverse denoising process of the human motion diffusion model and object motion diffusion model starts from noisy motion data H_T and O_T, generating clean human and object motion data (H_0 and O_0). Then, the human-object interaction diffusion model takes the text prompt and the clean human and object motion data (H_0 and O_0) as inputs, and generates the final clean human-object interaction motion data X_0.

4 Methodology

4.1 Overview

We first formulate the text-driven HOI generation task. Given a text prompt $\mathbf{T}_p$ and object geometry $\mathbf{G}_o$, our goal is to generate a HOI motion sequence $\mathbf{M} \in \mathbb{R}^{S\times D}$, where S indicates the length of the motion sequence and D represents the dimension of HOI motion representation, which comprises a human motion $\mathbf{H}$ and an object motion $\mathbf{O}$. The text prompt $\mathbf{T}_p$ is represented as $\mathbf{T}_p \in \mathbb{R}^{N\times L}$, where N denotes the number of words and L is the dimension of the word vector. The object geometry $\mathbf{G}_o$ is represented as $\mathbf{G}_o \in \mathbb{R}^{K\times 3}$, where K denotes the number of vertices on the object mesh.

As illustrated in Figure 2, we introduce MP-HOI, a diffusion model-based framework for text-driven HOI generation. We start from a human motion random noise and an object motion random noise, represented, respectively, as $\mathbf{H}_T \in \mathbb{R}^{S\times D_h}$ and $\mathbf{O}_T \in \mathbb{R}^{S\times D_o}$, where D_h denotes the dimension of human motion, and D_o represents the dimension of object motion. MP-HOI utilizes a human motion diffusion model and an object motion diffusion model to denoise them over t_h and t_o steps, respectively, generating $\mathbf{H}_0$ and $\mathbf{O}_0$. These are then fed as conditional inputs into a human-object interaction diffusion model to facilitate the denoising process of the HOI motion random noise, ultimately producing $\mathbf{X}_0$ over t_{hoi} steps.

4.2 Data representation

Human representation. Let $\mathbf{H} \in \mathbb{R}^{S\times D_h}$ denote the human motion. We adopt the SMPL-X [55] parametric model to represent the human motions. The representation D_h consists of the global joint position $\mathbf{D}_h^j \in \mathbb{R}^{52\times 3}$, joint rotation $\mathbf{D}_h^r \in \mathbb{R}^{52\times 6}$ represented in the continuous 6D rotation format [56] and global translation $\mathbf{D}_h^t \in \mathbb{R}^3$. Thus, the overall human motion representation is 471 dimensions. Notably, since the FullBodyManipulation [10] dataset does not provide hand parameters, we omit the hand component of the SMPL-X [55] model during processing.

Enhanced object representation. Let $\mathbf{O} = \{\mathbf{O}_j\}_{j=0}^{N_o} \in \mathbb{R}^{S\times D_o}$ denote object motion, where N_o represents the number of the objects. Typically, each D_o in $\mathbf{O}_j$ includes relative rotation $\mathbf{D}_o^r \in \mathbb{R}^6$ and global translation $\mathbf{D}_o^t \in \mathbb{R}^3$. Therefore, the object motion representation contains only nine dimensions in total, which is significantly fewer than the human motion representation. This dimension gap poses a substantial challenge for learning object motion. To address this limitation, we enhance the object representation with three additional informative features. First, we incorporate the object's translational velocity $\mathbf{D}_o^{v_t} \in \mathbb{R}^3$ and angular velocity $\mathbf{D}_o^{v_a} \in \mathbb{R}^3$, both derived from its translation and rotation. Second, we represent the object's geometric point cloud in a reduced form using 51 key points (50 sampled points plus 1 centroid). The global positions of these points $\mathbf{D}_o^p \in \mathbb{R}^{51\times 3}$ are computed via translation and rotation. Third, we include contact label information $\mathbf{D}_o^c \in \mathbb{R}^2$ by calculating the distance between the object and the human's left and right hands. As a result, our enhanced object representation D_o comprises a total of 170 dimensions per object.

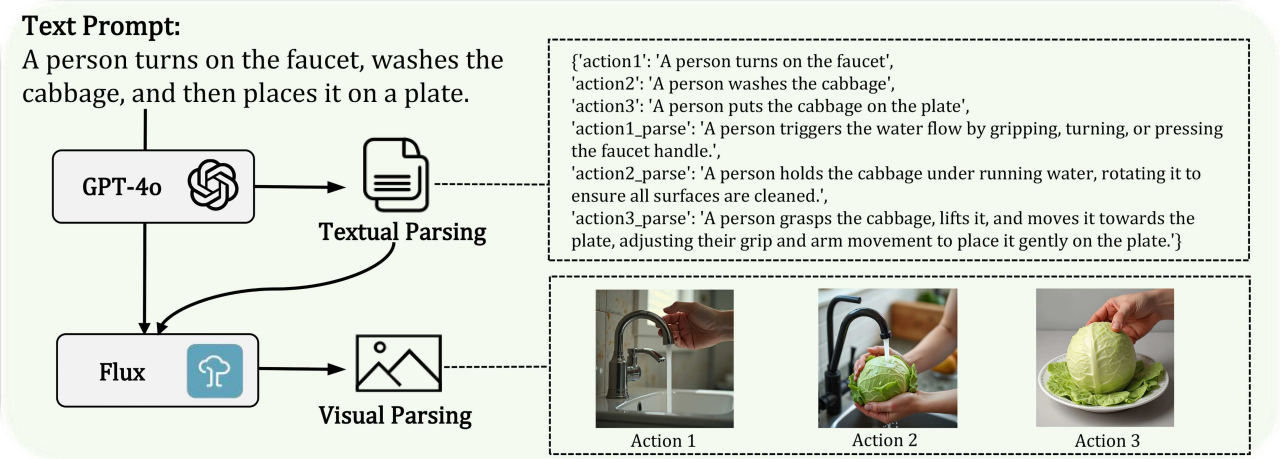

Figure 3 (Color online) The pipeline for large-model processing multimodal data (text and image).

4.3 Multimodal priors

Existing methods predominantly adopt a direct text-to-HOI mapping approach for generation. However, due to the significant cross-modal gap between text and motion modalities, such methods often exhibit limitations in modeling fine-grained HOI interactions. Recently, large-scale models across various modalities (e.g., GPT-4 [47], DeepSeek [57], DALL-E 3 [58], Sora [59]) have advanced the field of multimodal learning through their powerful modeling capabilities. Our key insight is to leverage multimodal priors from diverse modalities (1D, 2D, and 3D) to enhance the model's comprehension of interaction concepts, providing rich semantic guidance for generating human/object motion.

For the textual 1D prior, we leverage the strong priors of GPT-4o [47] to accurately capture the fine-grained relationship between natural language and human motion. Given a text prompt, we parse the sequence of actions being performed and provide detailed semantic explanations for each action. Specifically, for an input text such as "A person turns on the faucet, washes the cabbage, and then places it on a plate," we first analyze the execution order of motions to obtain action-1, action-2, and action-3. Then, we provide fine-grained descriptions for action-i to support a detailed understanding of interactive concepts at the textual level. The complete prompt is provided in the supplementary materials.

For the visual 2D prior, we leverage the powerful text-to-image capabilities of Flux [60] to provide fine-grained visual guidance for HOI. In the previous stage, we have already extracted a sequence of interaction actions—for example, action-1 is "A person turns on the faucet." Each action-i is then used as a textual prompt, augmented with style keywords such as "realism, photographic, detailed hand" and fed into Flux to generate the corresponding HOI images. These visual cues, generated for each interaction step, further strengthen the understanding of human-object interactions at the visual level. A complete example is provided in Figure 3.

For the spatial 3D prior, we process the human and object components separately. For the object, we regard its geometric structure as a rich source of spatial information and use its point cloud data $\mathbf{G}_o \in \mathbb{R}^{1024\times 3}$ as the 3D prior. For the human, we observe that HOI types are not infinitely diverse but instead fall into a limited set. Through all datasets [8, 10] analysis, we identified approximately 55 common HOI actions, such as pick up, place, eat, and wash. For each category, we select the most representative frame that captures the essence of the interaction as an atomic motion, and pair it with a corresponding textual annotation. To associate a given action-i with the most relevant atomic motion, we perform text-based retrieval based on semantic similarity. Specifically, we use CLIP [61] to extract text features for both the action-i text and all annotated atomic motion texts, and compute cosine similarity to find the closest match. By integrating the object's spatial geometry with the human's representative atomic motion, we further reinforce the understanding of HOI at the spatial level.

In summary, given a text prompt, we extract fine-grained action-parsed textual descriptions $\mathbf{C}_t$, visual images $\mathbf{C}_v$ generated based on the interaction order text, atomic motions $\mathbf{C}_a$, and object point clouds $\mathbf{G}_o$. For feature extraction, we employ CLIP to encode the text $\mathbf{C}_t$ and the images $\mathbf{C}_v$, yielding feature representations $\mathbf{C}_t^f \in \mathbb{R}^{N_a\times D_t}$ and $\mathbf{C}_v^f \in \mathbb{R}^{N_a\times D_v}$, respectively, where N_a denotes the number of text or image. The atomic motions $\mathbf{C}_a$ are encoded using an MLP to obtain $\mathbf{C}_a^f \in \mathbb{R}^{N_a\times D_a}$, while the object point clouds $\mathbf{G}_o$ are processed with a PointNet [62] architecture to obtain $\mathbf{C}_p^f \in \mathbb{R}^{N_b\times D_p}$, where N_b denotes the number of object geometry.

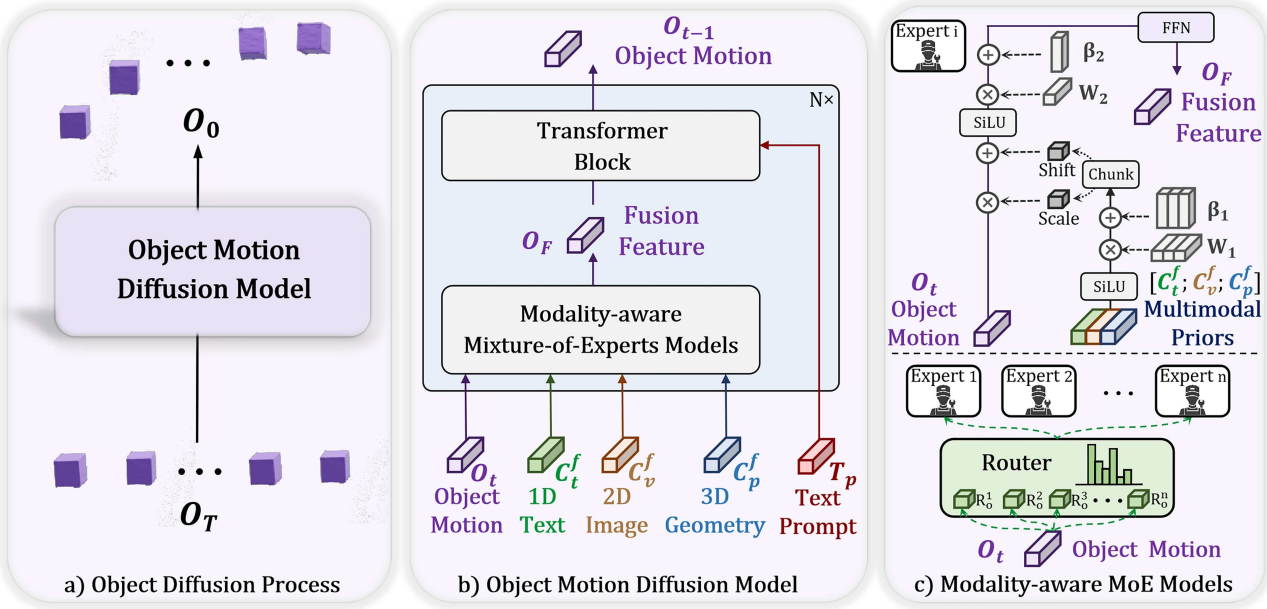

Figure 4 (Color online) Illustration of the overall object motion diffusion pipeline. (a) Object diffusion process; (b) object motion diffusion model; (c) architecture of modality-aware MoE models. Notably, the human motion diffusion model adopts this identical architecture, with the object geometry feature C_p^f replaced by the atomic motion feature C_a^f.

4.4 Human/object motion diffusion process

Existing methods typically adopt a one-step generation approach (e.g., only the human-object interaction diffusion in Figure 2) to generate human-object interaction motions. However, due to the inherent complexity of HOI dynamics, such direct text-to-HOI mapping often yields suboptimal results, including imprecise human motion, unnatural object trajectories, and coarse-grained interaction patterns. To address this issue, we propose a multi-step generation paradigm to enhance interaction quality. As illustrated in Figure 2, our framework first employs separate human motion diffusion and object motion diffusion models to generate preliminary human and object motions, conditioned on text prompts and multimodal priors. These intermediate motions serve as bridging representations, mitigating the feature gap between textual descriptions and HOI sequences. Subsequently, the preliminary human and object motions are integrated into the human-object interaction diffusion model as reference features to guide the HOI generation process, ultimately yielding refined HOI motion sequences.

Human/object motion diffusion model. Figure 4 presents the complete object motion diffusion pipeline. It is worth noting that the human motion diffusion pipeline follows a similar process, with the only difference being the replacement of the object geometry feature $\mathbf{C}_p^f$ with the atomic motion feature $\mathbf{C}_a^f$. Therefore, we take the object motion diffusion model as an example for a detailed explanation.

The goal of the object motion diffusion model is to generate a plausible object motion sequence based on text prompts and multimodal priors. We take the denoising process at time step t as an example, as illustrated in Figure 4(b). Specifically, the object motion features $\mathbf{O}_t$ and multimodal prior features $(\mathbf{C}_t^f, \mathbf{C}_v^f, \mathbf{C}_p^f)$ are fed into a modality-aware mixture-of-experts model which serves as an information fusion module. This module injects the multimodal knowledge into the object motion features, producing a fused representation. The fused object motion features are then combined with the text prompt features $\mathbf{T}_p$ and passed through multiple Transformer blocks, each consisting of a self-attention layer, a cross-attention layer, and a feedforward layer. This process ultimately outputs the refined object motion features $\mathbf{O}_{t-1}$ at time step $t-1$.

Modality-aware mixture-of-experts models. To generate fine-grained human-object interaction motion using data priors that provide rich semantic guidance, the core challenge lies in effectively fusing multi-modal condition features with HOI motion features. On a deeper level, the fundamental heterogeneity across different modalities (e.g., textual, visual, and spatial) introduces additional significant challenges for feature fusion. General fusion methods, such as simple concatenation or attention-based fusion, are typically static and inflexible. They may prioritize certain modalities while neglecting the information gains from others, failing to preserve fine-grained semantic features within some modalities' features that are crucial for capturing the nuances of human-object interaction modeling.

Mixture of experts [63, 64] offers a promising alternative by dynamically routing inputs to specialized experts, each handling distinct modality patterns. Thus, we argue that MoEs provide a highly effective and flexible solution to this challenge. However, previous MoE models typically adopt simple FFNs as expert layers [65, 66]. Such architectures are insufficient for effectively integrating motion and multi-modal condition features, as they fail to account for how each modality influences motion representation in a fine-grained manner. To address this limitation, we propose the modality-aware mixture-of-experts model, as shown in Figure 4(c).

Specifically, the MoEs include a routing function that directs the input features to the appropriate expert models, and several experts specializing in different tasks. Taking the object motion feature $\mathbf{O}_t$ as an example, the routing function contains a routing parameter matrix $\mathbf{R}_o$ that is used to assign tasks:

$$\mathbf{l}_o = \tau_t(\mathbf{W}_o\mathbf{O}_t)\mathbf{R}_o, \tag{3}$$

where τ_t is the trainable temperature hyper-parameter, $\mathbf{W}_o$ is the trainable matrix, $\mathbf{l}_o \in \mathbb{R}^{N_t \times N_e}$ represents the logits for selecting experts in object motion. Here, N_t denotes the number of tokens input to the MoE, and N_e denotes the number of expert models. By considering which expert model best matches the input tokens, the Router dynamically identifies and selects the most suitable expert.

As for expert architecture, drawing inspiration from FiLM [67], we design the condition modulation module to adaptively model the influence of multi-modal conditions on motion features:

$$[\mathbf{C}_f^1, \mathbf{C}_f^2] = \mathbf{W}_1\phi([\mathbf{C}_t^f; \mathbf{C}_v^f; \mathbf{C}_a^f]) + \boldsymbol{\beta}_1, \tag{4}$$

where $\mathbf{W}_1$ and $\boldsymbol{\beta}_1$ are the trainable matrices, $[\cdot;\cdot]$ indicates a concatenation of input tensors, ϕ denotes the activation function, $\mathbf{C}_f^1 \in \mathbb{R}^{1\times D_f}$ and $\mathbf{C}_f^2 \in \mathbb{R}^{1\times D_f}$ represent the scale and shift parameters, respectively, which are then adaptively fused with the motion features:

$$\mathbf{O}_F = \mathrm{FFN}\left[\mathbf{W}_2\phi\left[(1+\mathbf{C}_f^1)\mathbf{O}_t + \mathbf{C}_f^2\right] + \boldsymbol{\beta}_2\right], \tag{5}$$

where $\mathbf{W}_2$ and $\boldsymbol{\beta}_2$ are the trainable matrices, $\mathbf{O}_F$ is then reshaped to $\mathbb{R}^{T\times D_o}$ as the final fusion output of the MoE Layer.

Human/object motion diffusion training objective. To supervise the human/object motion generation process, we minimize the L2 loss between predicted and ground truth motions, denoted by

$$\mathcal{L}_{h/o}^{l2} = \mathbb{E}[\| \mathbf{x}_0 - \epsilon_\theta(\mathbf{x}_t, t, \mathbf{T}_p, \mathbf{C}_t^f, \mathbf{C}_v^f, \mathbf{C}_{a/p}^f) \|_2^2], \tag{6}$$

where $\epsilon_\theta(\mathbf{x}_t, t, \mathbf{T}_p, \mathbf{C}_t^f, \mathbf{C}_v^f, \mathbf{C}_{a/p}^f)$ denotes the model prediction, $\mathcal{L}_h^{l2}$ and $\mathcal{L}_o^{l2}$ denote the human motion diffusion loss and the object motion diffusion loss, respectively. In addition, it is necessary to consider the load balancing loss of the MoE, which serves to optimize the expert assignment mechanism. Therefore, the total training loss in this section includes the $\mathcal{L}_1$ loss and load balancing loss: $\mathcal{L}_{h/o} = \mathcal{L}_{h/o}^{l2} + \lambda_b\mathcal{L}_b$, where λ_b is the trade-off hyperparameter.

4.5 Human-object interaction diffusion process

Human-object interaction diffusion model. The human-object interaction diffusion model generates a fine-grained human-object interaction motion sequence based on the text prompt and the preliminary human and object motions produced in the previous stage, as illustrated in Figure 5. We take the denoising process at time step t and human motion $\mathbf{X}_t^h$ as an example. Specifically, the human motion features $\mathbf{X}_t^h$ are passed through multiple HOI Transformer blocks, each consisting of a self-attention layer, a mixed-attention layer, and a feedforward layer. The mixed-attention layer facilitates information exchange between the human motion $\mathbf{X}_t^h$ and various contextual features $(\mathbf{H}_0, \mathbf{T}_p, \mathbf{X}_t^o)$. Formally, the Query, Key, and Value matrices in this layer are computed as follows:

$$\mathbf{Q} = \mathbf{Q}_m^h\mathbf{X}_t^h, \quad \mathbf{K} = [\mathbf{K}_h\mathbf{H}_0; \mathbf{K}_p\mathbf{T}_p; \mathbf{K}_m^o\mathbf{X}_t^o], \quad \mathbf{V} = [\mathbf{V}_h\mathbf{H}_0; \mathbf{V}_p\mathbf{T}_p; \mathbf{V}_m^o\mathbf{X}_t^o], \tag{7}$$

where $\mathbf{Q}_m^h$, $\mathbf{K}_h$, $\mathbf{K}_p$, $\mathbf{K}_m^o$, $\mathbf{V}_h$, $\mathbf{V}_p$ and $\mathbf{V}_m^o$ are trainable matrices. Then, the global templates $\mathbf{G}_g$ in the attention are computed to yield the output $\mathbf{Y}$:

$$\mathbf{G}_g = \mathrm{softmax}(\mathbf{K})\mathbf{V}, \quad \mathbf{Y} = \mathrm{softmax}(\mathbf{Q})\mathbf{G}_g. \tag{8}$$

By passing through multiple HOI Transformer blocks, the human motion representation is progressively refined to produce the optimized motion feature $\mathbf{X}_{t-1}^h$. Similarly, the object motion $\mathbf{X}_t^o$ undergoes a parallel process, resulting in the optimized object motion feature $\mathbf{X}_{t-1}^o$.

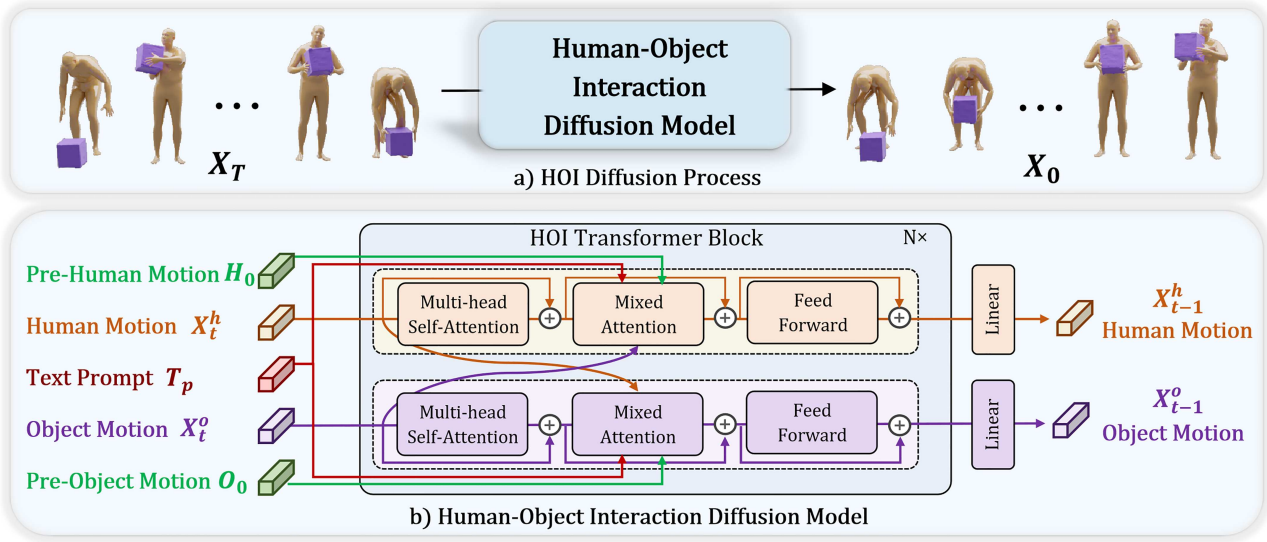

Figure 5 (Color online) Illustration of the overall human-object interaction motion diffusion pipeline. (a) Human-object interaction motion diffusion process; (b) architecture of the human-object interaction diffusion model. Pre-Human Motion and Pre-Object Motion represent the human motion and object motion generated in the human/object motion diffusion process, respectively.

Human-object interaction diffusion training objective. To supervise the human-object interaction motion generation process, we employ a comprehensive set of loss functions to ensure the plausibility of the generated motion. First, we apply an L2 loss $\mathcal{L}_{l2}$, similar to (6), to encourage accurate reconstruction of HOI motion. The HOI motion representation $\mathbf{X}$ consists of both human and object motion components. Notably, the human motion representation includes global joint positions, joint rotations, and global translation, all of which are effectively constrained by the L2 loss to ensure high-quality human motion. More importantly, the object motion representation is enriched with additional features such as relative rotation, global translation, angular velocity, translational velocity, the global positions of keypoints, and contact information. Consequently, even this simple L2 loss implicitly supervises these diverse and informative object features.

Second, we introduce a velocity-level constraint $\mathcal{L}_{vel}$ by computing the velocities of the human body and both hands, ensuring that the generated motion exhibits realistic dynamics. Third, to maintain reasonable spatial relationships among objects, we incorporate a distance-based loss $\mathcal{L}_{dis}$ that penalizes physically implausible inter-object distances. Fourth, to prevent unnatural penetration between the human and the object, we apply an interaction loss $\mathcal{L}_{inter}$, which computes the distance between the human's left/right hands and the object's centroid, thereby guiding appropriate proximity in human-object interactions. In summary, we formulate the final optimization objective at this stage as

$$\mathcal{L}_{\text{hoi}} = \lambda_{l2}\mathcal{L}_{l2} + \lambda_{vel}\mathcal{L}_{vel} + \lambda_{dis}\mathcal{L}_{dis} + \lambda_{inter}\mathcal{L}_{inter}, \tag{9}$$

where $\lambda_{l2}, \lambda_{vel}, \lambda_{dis}, \lambda_{inter}$ are hyperparameters. Since our MP-HOI is trained in an end-to-end manner, the overall training objective is defined as the sum of the losses from the preceding stages, formulated as $\mathcal{L} = \mathcal{L}_h + \mathcal{L}_o + \mathcal{L}_{\text{hoi}}$.

5 Experiments

5.1 Datasets, metrics and implementation details

Datasets. FullBodyManipulation [10] contains 10 h of motion data involving human interactions with a single object, comprising a total of 4838 HOI sequences. Each HOI sequence is accompanied by a textual description that guides the volunteers during the motion recording. A total of 17 participants were involved in the data collection process. They interacted with each object according to the given textual instructions. The dataset includes 15 commonly used objects in daily tasks, such as clothes stand, suitcase, table, trashcan, monitor, and others.

HIMO [8] includes 9.44 h of motion data depicting human interactions with two or three objects, comprising 3376 HOI sequences. Each sequence is paired with a textual description. A total of 34 participants contributed to the data collection. The dataset covers 53 everyday household objects, such as plate, laptop, bottle, apple, bowl, and more. It also encompasses many interaction types, including: Put A (and B) into C, Wash A (and B) under faucet, Use A and B, Place A on B, among others.

Metrics. To evaluate HOI motion generation, we first employ general metrics to assess the quality of the generation, such as R-TOP, FID, MM-Dist, and Diversity. R-TOP reflects the semantic consistency between generated HOIs and the given textual prompts. FID measures the similarity between the feature distributions extracted from the generated motions and the ground truth motions. MM-Dist computes the average Euclidean distance between the feature of generated motions and the text prompt feature. Diversity evaluates the dissimilarity among all generated motions across all descriptions. Furthermore, we follow [10, 11] evaluation metrics to assess the quality of human-object interactions, including Interaction Distance, Contact Percentage, Precision, Recall and F1 score. We first compute the distance between hand positions and the object centroid point. The interaction distance D_I is defined as the difference between the distances of the generated motions and those of the ground truth motions. Additionally, a contact threshold is empirically set to determine contact labels for each frame. We then count true positives, false positives, and false negatives to compute precision C_{prec}, recall C_{rec}, and F1 score C_{F1}. Moreover, Contact Percentage $C_{\%}$ reflects frame-level contact inference accuracy and is defined as the proportion of frames where contact is detected.

Implementation details. Regarding the multimodal priors, we utilize GPT-4o [47] to process text data and Flux [60] to generate image data. The hyperparameter N_a is set to 2 on the FullBodyManipulation [10] dataset and 3 on the HIMO [8] dataset. N_b is set to 1 on the FullBodyManipulation (1 object), 2 on HIMO (2 objects), and 3 on HIMO (3 objects). Regarding the motion diffusion model, we employ a 4-layer Transformer in human motion diffusion model, a 4-layer Transformer in object motion diffusion model and an 8-layer Transformer in the HOI motion diffusion model. As for the MoE models, the number of experts is set to 16, with top-k set to 2. The dimension R_0 is set to 256. As for the text encoder, a frozen text encoder from CLIP ViT-B/32 is utilized, complemented by two additional Transformer encoder layers. Regarding some hyperparameters, D_a is set to 512, D_p is set to 256, λ_b is set to 10, λ_{l2} is set to 1, λ_{vel} is set to 2, λ_{dis} is set to 0.3, λ_{inter} is set to 0.01, and the guidance scale is set to 2.0. In terms of the diffusion model, the variances β_t are predefined to linearly spread from 0.0001 to 0.02, and the total number of noising steps is set at $T = 1000$. We use the Adam optimizer to train the model with an initial learning rate of 0.0001, gradually decreasing to 0.00001 through a cosine learning rate scheduler. The training process is conducted on four NVIDIA GeForce RTX 3090 GPUs, with a batch size of 16 on a single GPU.

5.2 Evaluation of human-object interaction generation

General evaluation. The results in Table 1 compare MP-HOI against state-of-the-art (SOTA) methods including IMoS [45], MDM [32], OMOMO [10], PriorMDM [68], MotionDiffuse [31], CHOIS [11], and HIMO-Gen [8] on the FullBodyManipulation [10] and HIMO [8] datasets. In terms of motion quality evaluation, compared to other methods, MP-HOI achieves significantly higher scores in R-TOP3, FID, and MM-Dist. These results highlight our method's proficiency in generating high-quality HOI motion sequences that seamlessly align with the textual prompts. Notably, compared with the HIMO-Gen [8], our method reduces the FID by 23.92% on the FullBodyManipulation single-object interaction dataset, by 27.75% on HIMO 2-object dataset, and by an impressive 66.02% on the 3-object dataset. We further analyze the underlying reasons. Specifically, MotionDiffuse [31] and MDM [32] rely primarily on single-modality inputs and do not incorporate object-related information. Without explicit modeling of object conditions, their generated motions often lack accurate and physically consistent human-object interactions. In contrast, IMoS [45], CHOIS [11], and HIMO-Gen [8] consider object conditions, but they fail to accurately model interactions. Because these methods lack prior knowledge about human-object interactions, they cannot achieve a fine-grained understanding of interaction concepts. Furthermore, their object motion representations rely on a direct 9-dimensional encoding, which makes it difficult for the models to fully interpret the embedded information, resulting in lower motion quality. This limitation restricts their ability to model complex human-object interactions, which explains the significant performance gap compared to MP-HOI.

Human-object interaction evaluation. We perform a human-object interaction evaluation to assess the interaction performance based on five HOI interaction metrics. As shown in Table 2, compared to SOTA methods, MP-HOI achieves significant improvements in Precision C_{prec}, Recall C_{rec}, and F1 score C_{F1}. These results reveal that our method has better hand-object contact accuracy and enhanced physical reasonableness. For example, compared with CHOIS [11] in the 2-object setting of the HIMO dataset, C_{prec} increases by 2.25%, C_{rec} increases by 4.10%, and C_{F1} increases by 4.12%. In addition, our method also performs well on the Contact Percentage ($C_{\%}$) metric. For example, it improves by 15.27% compared to HIMO-Gen [8], indicating that the percentage of contact frames in the MP-HOI generated human-object interaction motions is closest to that of the ground truth motions. Finally, regarding the interaction distance (D_I) metric, our method achieves a 19.04% reduction compared to MotionDiffuse [31]. Therefore, the human-object distances in our generated motions are the most reasonable, effectively reducing interpenetration and other unrealistic artifacts.

Table 1 Comparisons to current state-of-the-art methods on the FullBodyManipulation [10] and HIMO [8] test sets. "↑" denotes that higher is better. "↓" denotes that lower is better. We repeat all the evaluations 20 times and report the average with a 95% confidence interval. We report the best and the second-best results in **bold** and underline, respectively.

Dataset	Method	Publication	Motion quality evaluation			Diversity evaluation
			R-TOP 3 ↑	FID ↓	MM Dist ↓	Diversity ↑
FullBodyManipulation (1 Object)	OMOMO [10]	TOG 2023	$0.773^{\pm 0.009}$	$1.276^{\pm 0.016}$	$2.468^{\pm 0.034}$	$9.719^{\pm 0.087}$
	MotionDiffuse [31]	TPAMI 2024	$0.830^{\pm 0.003}$	$0.892^{\pm 0.019}$	$2.220^{\pm 0.021}$	$\underline{10.02}^{\pm 0.066}$
	CHOIS [11]	ECCV 2024	$0.791^{\pm 0.005}$	$\underline{0.823}^{\pm 0.012}$	$\underline{2.177}^{\pm 0.011}$	$9.998^{\pm 0.037}$
	HIMO-Gen [8]	ECCV 2024	$\underline{0.851}^{\pm 0.008}$	$0.924^{\pm 0.026}$	$2.346^{\pm 0.035}$	$9.877^{\pm 0.089}$
	Ours (MP-HOI)	–	$\mathbf{0.872}^{\pm \mathbf{0.005}}$	$\mathbf{0.703}^{\pm \mathbf{0.042}}$	$\mathbf{1.948}^{\pm \mathbf{0.016}}$	$\mathbf{10.38}^{\pm \mathbf{0.059}}$
HIMO (2 Objects)	IMoS [45]	CGF 2023	$0.501^{\pm 0.012}$	$7.589^{\pm 0.112}$	$8.740^{\pm 0.031}$	$7.003^{\pm 0.320}$
	MDM [32]	ICLR 2023	$0.605^{\pm 0.009}$	$6.845^{\pm 0.331}$	$8.018^{\pm 0.050}$	$11.38^{\pm 0.234}$
	OMOMO [10]	TOG 2023	$0.592^{\pm 0.012}$	$6.132^{\pm 0.271}$	$7.921^{\pm 0.065}$	$\underline{12.73}^{\pm 0.196}$
	PriorMDM [68]	ICLR 2024	$0.589^{\pm 0.003}$	$7.851^{\pm 0.251}$	$7.250^{\pm 0.006}$	$12.57^{\pm 0.146}$
	MotionDiffuse [31]	TPAMI 2024	$0.576^{\pm 0.009}$	$4.364^{\pm 0.039}$	$5.190^{\pm 0.039}$	$10.79^{\pm 0.106}$
	CHOIS [11]	ECCV 2024	$0.567^{\pm 0.041}$	$3.996^{\pm 0.587}$	$5.986^{\pm 0.693}$	$12.44^{\pm 0.514}$
	HIMO-Gen [8]	ECCV 2024	$\underline{0.636}^{\pm 0.003}$	$\underline{1.481}^{\pm 0.042}$	$\underline{3.649}^{\pm 0.010}$	$11.66^{\pm 0.204}$
	Ours (MP-HOI)	–	$\mathbf{0.842}^{\pm \mathbf{0.007}}$	$\mathbf{1.070}^{\pm \mathbf{0.021}}$	$\mathbf{2.968}^{\pm \mathbf{0.029}}$	$\mathbf{12.83}^{\pm \mathbf{0.079}}$
HIMO (3 Objects)	IMoS [45]	CGF 2023	$0.466^{\pm 0.101}$	$4.990^{\pm 0.177}$	$7.770^{\pm 0.058}$	$9.231^{\pm 0.113}$
	MDM [32]	ICLR 2023	$0.502^{\pm 0.013}$	$4.571^{\pm 0.110}$	$6.314^{\pm 0.026}$	$8.895^{\pm 0.285}$
	OMOMO [10]	TOG 2023	$0.553^{\pm 0.037}$	$4.561^{\pm 0.039}$	$5.463^{\pm 0.049}$	$9.169^{\pm 0.073}$
	PriorMDM [68]	ICLR 2024	$0.513^{\pm 0.025}$	$4.821^{\pm 0.203}$	$5.890^{\pm 0.023}$	$\underline{9.340}^{\pm 0.023}$
	MotionDiffuse [31]	TPAMI 2024	$0.515^{\pm 0.013}$	$4.719^{\pm 0.059}$	$5.673^{\pm 0.046}$	$8.993^{\pm 0.097}$
	CHOIS [11]	ECCV 2024	$\underline{0.602}^{\pm 0.007}$	$\underline{3.653}^{\pm 0.046}$	$\underline{4.763}^{\pm 0.046}$	$9.135^{\pm 0.091}$
	HIMO-Gen [8]	ECCV 2024	$0.535^{\pm 0.018}$	$4.771^{\pm 0.110}$	$5.086^{\pm 0.041}$	$8.946^{\pm 0.137}$
	Ours (MP-HOI)	–	$\mathbf{0.729}^{\pm \mathbf{0.009}}$	$\mathbf{1.621}^{\pm \mathbf{0.030}}$	$\mathbf{3.392}^{\pm \mathbf{0.019}}$	$\mathbf{10.02}^{\pm \mathbf{0.103}}$

Table 2 Human-object interaction evaluation on the FullBodyManipulation [10] and HIMO [8] test sets. '→' means the closer to GT the better.

Dataset	Method	Publication	Human-object interaction evaluation				
			C_{prec} ↑	C_{rec} ↑	C_{F1} ↑	$C_{\%}$ →	D_I →
FullBodyManipulation (1 Object)	GT	–	–	–	–	$0.445^{\pm 0.000}$	$0.501^{\pm 0.000}$
	MotionDiffuse [31]	TPAMI 2024	$0.358^{\pm 0.015}$	$0.272^{\pm 0.013}$	$0.278^{\pm 0.013}$	$\underline{0.321}^{\pm 0.017}$	$0.615^{\pm 0.019}$
	CHOIS [11]	ECCV 2024	$\underline{0.372}^{\pm 0.009}$	$\underline{0.305}^{\pm 0.009}$	$\underline{0.303}^{\pm 0.008}$	$0.318^{\pm 0.019}$	$0.637^{\pm 0.019}$
	HIMO-Gen [8]	ECCV 2024	$0.364^{\pm 0.008}$	$0.286^{\pm 0.008}$	$0.289^{\pm 0.009}$	$0.297^{\pm 0.011}$	$\underline{0.613}^{\pm 0.013}$
	Ours (MP-HOI)	–	$\mathbf{0.396}^{\pm \mathbf{0.007}}$	$\mathbf{0.343}^{\pm \mathbf{0.006}}$	$\mathbf{0.342}^{\pm \mathbf{0.007}}$	$\mathbf{0.371}^{\pm \mathbf{0.005}}$	$\mathbf{0.589}^{\pm \mathbf{0.007}}$
HIMO (2 Objects)	GT	–	–	–	–	$0.833^{\pm 0.000}$	$0.215^{\pm 0.000}$
	MotionDiffuse [31]	TPAMI 2024	$\underline{0.845}^{\pm 0.013}$	$0.764^{\pm 0.013}$	$0.770^{\pm 0.012}$	$0.747^{\pm 0.017}$	$0.315^{\pm 0.017}$
	CHOIS [11]	ECCV 2024	$0.829^{\pm 0.011}$	$0.789^{\pm 0.012}$	$0.791^{\pm 0.011}$	$\underline{0.756}^{\pm 0.010}$	$\underline{0.294}^{\pm 0.009}$
	HIMO-Gen [8]	ECCV 2024	$0.844^{\pm 0.007}$	$\underline{0.804}^{\pm 0.008}$	$\underline{0.802}^{\pm 0.007}$	$0.707^{\pm 0.010}$	$0.302^{\pm 0.009}$
	Ours (MP-HOI)	–	$\mathbf{0.863}^{\pm \mathbf{0.010}}$	$\mathbf{0.837}^{\pm \mathbf{0.011}}$	$\mathbf{0.835}^{\pm \mathbf{0.009}}$	$\mathbf{0.815}^{\pm \mathbf{0.007}}$	$\mathbf{0.255}^{\pm \mathbf{0.008}}$
HIMO (3 Objects)	GT	–	–	–	–	$0.843^{\pm 0.000}$	$0.222^{\pm 0.000}$
	MotionDiffuse [31]	TPAMI 2024	$0.841^{\pm 0.015}$	$0.795^{\pm 0.015}$	$0.794^{\pm 0.016}$	$\underline{0.784}^{\pm 0.011}$	$\underline{0.303}^{\pm 0.014}$
	CHOIS [11]	ECCV 2024	$\underline{0.850}^{\pm 0.010}$	$\underline{0.803}^{\pm 0.011}$	$\underline{0.804}^{\pm 0.011}$	$0.771^{\pm 0.015}$	$0.332^{\pm 0.012}$
	HIMO-Gen [8]	ECCV 2024	$0.844^{\pm 0.010}$	$0.772^{\pm 0.009}$	$0.779^{\pm 0.010}$	$0.758^{\pm 0.011}$	$0.315^{\pm 0.014}$
	Ours (MP-HOI)	–	$\mathbf{0.859}^{\pm \mathbf{0.009}}$	$\mathbf{0.824}^{\pm \mathbf{0.009}}$	$\mathbf{0.825}^{\pm \mathbf{0.008}}$	$\mathbf{0.815}^{\pm \mathbf{0.008}}$	$\mathbf{0.274}^{\pm \mathbf{0.007}}$

User study evaluation. We conduct a user study, in which we compare our method with HIMO-Gen [8] and MotionDiffuse [31]. This user study engaged 30 participants to evaluate 10 motion sequences generated by each method. The designed questionnaire consisted of two questions: (1) "Which method generates motion that best aligns with the textual prompt?" and (2) "Which method best captures the fine-grained details of human-object interactions?" Participants rated the methods on a 1-to-3 scale (indicating bad, good, and best). As depicted in Figure 6, MP-HOI significantly outperforms the competing methods in both semantic alignment and interaction quality. These results demonstrate that MP-HOI not only produces motions that closely follow textual prompts but also excels in modeling human-object interactions, validating the effectiveness and superiority of our approach.

Qualitative analysis. Figure 7 qualitatively compares PerMoGen against HIMO-Gen [8] and MotionDiffuse [31].

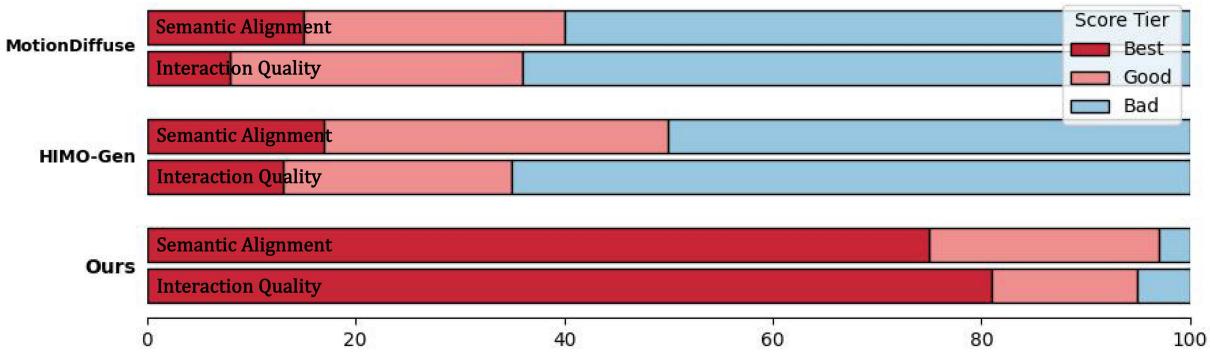

Figure 6 (Color online) User study results. The color bars indicate the percentage distribution of scores for each evaluation criterion.

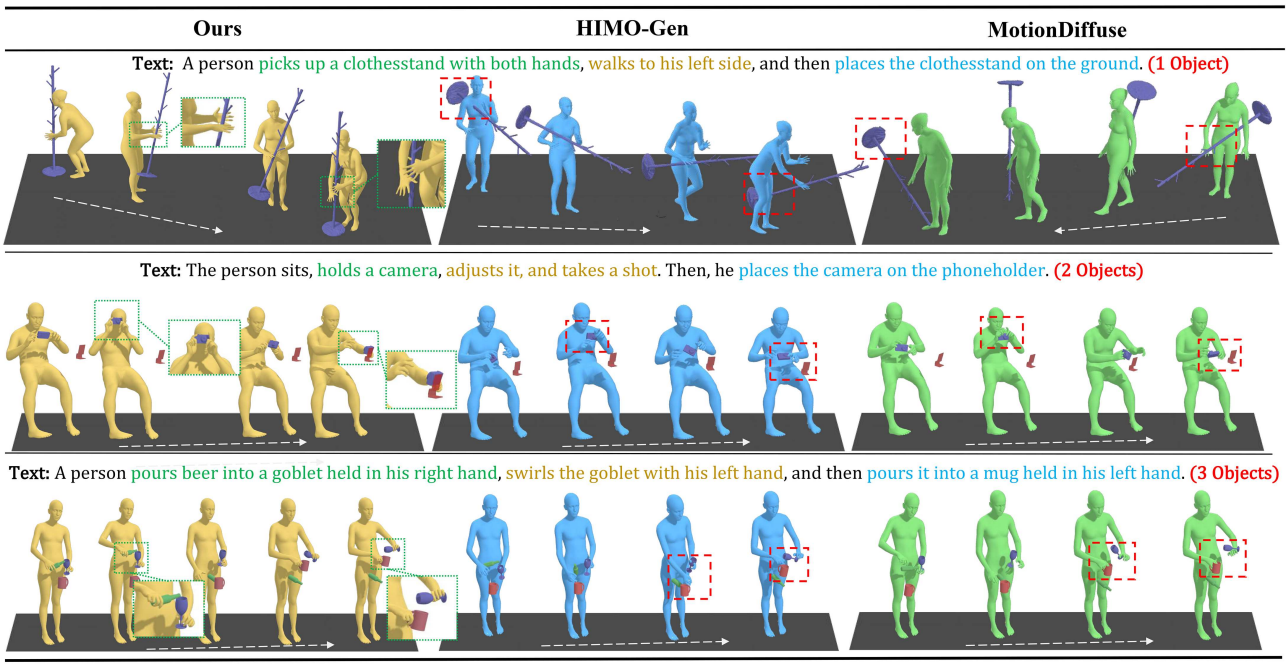

Figure 7 (Color online) Visual results compared with existing methods. The arrow represents the time axis. The green box zooms in on the detailed interactions demonstrated by our approach. The red boxes highlight the errors in other methods.

In the single-object scenario, HIMO-Gen [8] exhibits significant human-object penetration and generates unrealistic object motion. Similarly, MotionDiffuse [31] fails to accurately follow the text prompts, often producing motion in the opposite direction of what is described. In the two-object scenario, HIMO-Gen [8] generates low-quality motion where the camera is not properly held in the human's hand as instructed, while MotionDiffuse [31] fails to place the camera back onto the phoneholder as required by the text. In the three-object scenario, both HIMO-Gen [8] and MotionDiffuse [31] demonstrate weak and imprecise interactions, struggling to engage with each object in a smooth and meaningful way. In contrast, MP-HOI consistently produces motion that closely aligns with the textual prompts across all scenarios. It demonstrates fine-grained modeling of human-object interactions, highlighting the superiority of our approach.

Quantitative ablation study. We conduct a comprehensive set of ablation studies to evaluate the contribution of each key component in MP-HOI, as summarized in Table 3. First, we examine the impact of multimodal priors. Removing the multimodal priors leads to a noticeable decline of 69.27% in FID and 9.45% in motion diversity. This result highlights the importance of incorporating multimodal cues, as they not only guide the generation of high-quality interactive motions but also significantly enhance motion diversity. We further investigated the impact of each modality on the final performance and found that different modalities affect motion quality to varying degrees. For example, under the MM-Dist metric, removing the 1D prior leads to a 10.37% drop in performance, removing the 2D prior results in a 13.29% drop, and removing the 3D prior causes a 17.27% drop. These results suggest that, since the task involves 3D motion generation, knowledge that is closer to the 3D domain more effectively guides the

Table 3 Ablation study results on the FullBodyManipulation [10] test set.

Method	Motion quality evaluation			Diversity evaluation	Human-object interaction evaluation				
	R-TOP 3 ↑	FID ↓	MM Dist ↓	Diversity ↑	C_{prec} ↑	C_{rec} ↑	C_{F1} ↑	$C_{\%}$ →	D_I →
w/o Multimodal Priors (1D, 2D, 3D)	$0.761^{\pm0.010}$	$1.190^{\pm0.022}$	$2.321^{\pm0.009}$	$9.399^{\pm0.019}$	$0.369^{\pm0.009}$	$0.300^{\pm0.008}$	$0.304^{\pm0.009}$	$0.319^{\pm0.031}$	$0.602^{\pm0.027}$
w/o Multimodal Priors (1D)	$0.784^{\pm0.015}$	$0.995^{\pm0.019}$	$2.150^{\pm0.017}$	$9.431^{\pm0.045}$	$0.386^{\pm0.024}$	$0.319^{\pm0.026}$	$0.321^{\pm0.024}$	$0.351^{\pm0.035}$	$0.582^{\pm0.010}$
w/o Multimodal Priors (2D)	$0.772^{\pm0.009}$	$1.054^{\pm0.011}$	$2.207^{\pm0.026}$	$9.543^{\pm0.027}$	$0.381^{\pm0.019}$	$0.310^{\pm0.020}$	$0.309^{\pm0.022}$	$0.334^{\pm0.025}$	$0.597^{\pm0.034}$
w/o Multimodal Priors (3D)	$0.769^{\pm0.011}$	$1.086^{\pm0.026}$	$2.226^{\pm0.010}$	$9.561^{\pm0.029}$	$0.373^{\pm0.013}$	$0.303^{\pm0.012}$	$0.301^{\pm0.012}$	$0.323^{\pm0.019}$	$0.599^{\pm0.021}$
GPT-4o + Flux	$0.783^{\pm0.008}$	$1.516^{\pm0.030}$	$2.442^{\pm0.019}$	$9.576^{\pm0.059}$	$0.345^{\pm0.018}$	$0.312^{\pm0.009}$	$0.316^{\pm0.010}$	$0.311^{\pm0.022}$	$0.553^{\pm0.044}$
GPT-3.5 + Stable Diffusion	$0.762^{\pm0.010}$	$1.811^{\pm0.022}$	$2.679^{\pm0.046}$	$9.421^{\pm0.089}$	$0.312^{\pm0.025}$	$0.298^{\pm0.012}$	$0.301^{\pm0.016}$	$0.307^{\pm0.031}$	$0.539^{\pm0.024}$
w/o Enhanced Object Representation	$0.759^{\pm0.005}$	$1.346^{\pm0.011}$	$2.475^{\pm0.020}$	$9.416^{\pm0.069}$	$0.359^{\pm0.007}$	$0.289^{\pm0.008}$	$0.290^{\pm0.008}$	$0.315^{\pm0.022}$	$0.615^{\pm0.017}$
w/o Cascaded Diffusion Framework	$0.755^{\pm0.005}$	$1.368^{\pm0.016}$	$2.599^{\pm0.030}$	$9.711^{\pm0.105}$	$0.352^{\pm0.012}$	$0.310^{\pm0.012}$	$0.315^{\pm0.012}$	$0.309^{\pm0.017}$	$0.586^{\pm0.030}$
w/o Modality-aware MoE Models	$0.836^{\pm0.007}$	$0.931^{\pm0.020}$	$2.080^{\pm0.041}$	$9.404^{\pm0.085}$	$0.379^{\pm0.006}$	$0.327^{\pm0.007}$	$0.329^{\pm0.008}$	$0.330^{\pm0.009}$	$0.597^{\pm0.016}$
w/o HOI Supervision Loss	$0.841^{\pm0.009}$	$0.890^{\pm0.019}$	$2.174^{\pm0.009}$	$9.835^{\pm0.052}$	$0.364^{\pm0.009}$	$0.317^{\pm0.009}$	$0.319^{\pm0.008}$	$0.325^{\pm0.011}$	$0.569^{\pm0.010}$
Ours (MP-HOI)	$\mathbf{0.872}^{\pm\mathbf{0.005}}$	$\mathbf{0.703}^{\pm\mathbf{0.042}}$	$\mathbf{1.948}^{\pm\mathbf{0.016}}$	$\mathbf{10.38}^{\pm\mathbf{0.059}}$	$\mathbf{0.396}^{\pm\mathbf{0.007}}$	$\mathbf{0.343}^{\pm\mathbf{0.006}}$	$\mathbf{0.342}^{\pm\mathbf{0.007}}$	$\mathbf{0.371}^{\pm\mathbf{0.005}}$	$\mathbf{0.589}^{\pm\mathbf{0.007}}$

generation process. Additionally, we examined the effect of different large model combinations on motion generation performance. We randomly sampled 100 examples and processed them using various combinations of large models. The results show that the combination of GPT-4o and Flux leads to more significant improvements in metrics, primarily because it provides more detailed and comprehensive text analysis and image-based prompts. Second, when replacing our enhanced object encoding with the original 9-dimensional format (translation + rotation), we observe a substantial decline in HOI motion quality. Specifically, the FID increases by 91.46%, and R-TOP 3 drops by 12.95%, demonstrating that our enriched object representation is critical for generating high-quality HOI motions. Third, eliminating the cascaded diffusion framework reduces MP-HOI to a single-step generation paradigm, similar to other baseline methods. This results in coarse and less realistic HOI motions, as reflected in the degradation of both motion quality and interaction accuracy. For instance, MM-Dist decreases by 33.41%, and C_{prec} drops by 11.11%. Fourth, when the modality-aware MoE module is removed, the model struggles to effectively fuse multimodal priors with HOI motion features. Thus, it fails to fully exploit the rich information across modalities, leading to noticeable drops in both motion quality and diversity, for example, a 32.43% degradation in FID and a 9.40% reduction in diversity. Finally, removing the HOI supervision constraints results in poor human-object interaction quality. In particular, the D_I metric experiences a significant performance drop (3.39%), indicating less plausible spatial relationships and a higher frequency of interpenetration artifacts.

Qualitative ablation study. To thoroughly evaluate the individual contribution of each module, we employed motion visualization for an in-depth comparative analysis and examined the impact of removing key components, as shown in Figure 8(a). When the enhanced object representation was removed, the quality of object motion significantly deteriorated, leading to sliding and floating artifacts of the chair. Without the multimodal priors, the human demonstrated poor interaction with the object, erroneously grasping unreasonable parts of the chair. In the absence of the cascaded diffusion framework, both the human and the motion quality degraded drastically, resulting in severe human-object penetration. In contrast, the complete MP-HOI model successfully performed the motion dictated by the text prompt and exhibited fine-grained human-object interaction, thereby validating the effectiveness of our approach.

Moreover, we provide ablation and visualization experiments for each modality, as shown in Figure 8(b). From these results, we observe that when the textual modality is removed, the generated motions may fail to fully reflect the multiple steps described in the input text. When the visual modality is removed, the quality of hand-object contact information is weakened. When the motion modality is removed, the generated motions tend to have lower overall motion quality. These observations demonstrate the specific contribution of each modality in supporting fine-grained HOI generation.

Generalization. To evaluate the generalization ability of MP-HOI to unseen objects, we qualitatively assess the effectiveness of our method. We feed their geometries as input conditions into our model and generate corresponding human-object motions, as illustrated in Figure 8(c). In the first example, the human correctly grasps the end of the pliers and uses the tool appropriately. In the second example, the human interacts with a reasonable region of the yoga ball and successfully completes a sequence of actions, including picking it up, walking, and putting it down. These examples demonstrate that MP-HOI can still generate high-quality human-object interaction motions even when interacting with previously unseen object geometries, such as a plier or a yoga ball, which are not present in the dataset. This highlights the effectiveness of our method and its strong generalization capability to unseen objects.

Additional visualization results. Figure 8(d) presents four additional visualizations showcasing human-object interactions. From the first and second examples on the left side, it is evident that MP-HOI can stably grasp appropriate parts of the monitor and the edge of the table, thereby accurately performing the actions described in the text. In the third example, MP-HOI effectively models the interaction between the human and two objects,

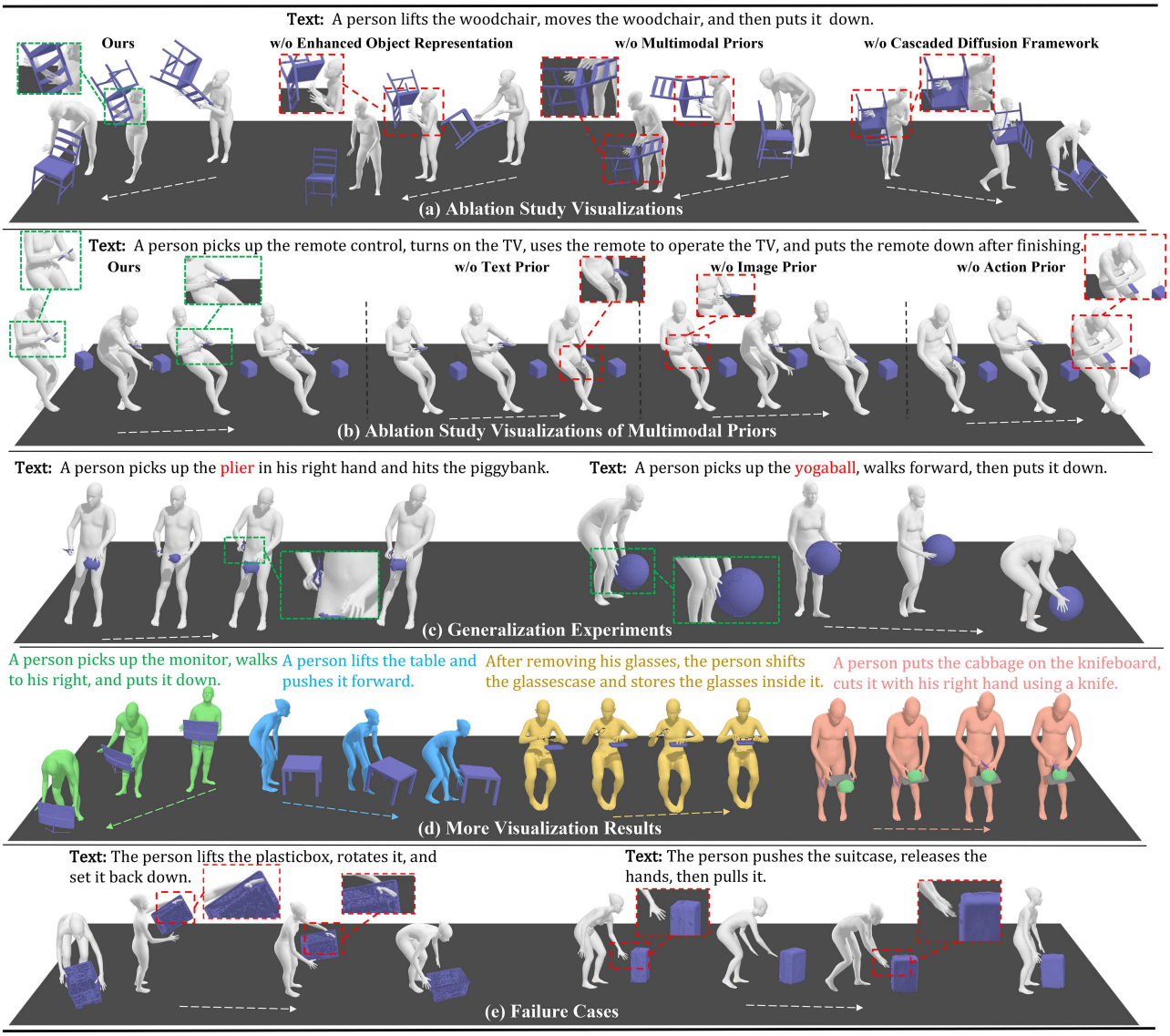

Figure 8 (Color online) Additional visualization results. (a) Ablation study visualizations; (b) ablation study visualizations of multimodal priors; (c) generalization experiments; (d) more visualization results; (e) failure cases.

successfully completing the task of picking up the glasses and placing them back into the glasses case. Even in the fourth example, which involves interaction with three objects, the human is able to reasonably manage the relationships among multiple objects and sequentially carry out the instructions from the text, demonstrating strong multi-object interaction modeling capabilities. Further comparisons and visualization examples are provided in the supplementary video.

Failure cases. While MP-HOI effectively generates human-object interaction motion based on textual prompts, it encounters two failure cases shown in Figure 8(e). The primary failures occur on the FullBodyManipulation [10] dataset, which does not provide hand pose parameters. As a result, the generated motions on this dataset sometimes cannot accurately model finger positions, leading to potential hand-object penetrations. Additionally, during close-range interactions with objects, such as pushing a suitcase, the lack of detailed hand parameters may result in overly large contact distances between the hand and the object.

6 Limitation and future work

Here, we discuss the limitations of MP-HOI and suggest directions for future work. First, although the generated motion sequences generally align well with textual prompts, some artifacts, such as foot sliding, may occasionally

occur. These issues could be alleviated in the future by incorporating additional physical constraints and enhancing physical simulations. Second, our current framework is limited to interactions with rigid objects. However, in the real world, humans often interact with deformable or fluid-like substances, such as gases or liquids. Exploring human interactions with such dynamic materials presents an interesting and promising direction for future research.

7 Conclusion

In this paper, we address the challenging task of text-driven human-object interaction generation by introducing MP-HOI, a novel framework built upon four key insights. (1) We leverage multimodal priors, including text, images, and pose/object information, to effectively guide the HOI generation process. (2) We enhance object representations by incorporating geometric keypoints, contact features, and dynamic properties, enabling more stable and informative representations. (3) We propose a modality-aware Mixture-of-Experts model to facilitate feature fusion of multimodal data. (4) We design a cascaded diffusion framework that progressively refines human-object interaction features under dedicated supervision. We conduct extensive quantitative and qualitative experiments, demonstrating that MP-HOI significantly outperforms existing methods by generating motions that are not only well-aligned with textual prompts but also capable of modeling fine-grained human-object interactions.

Acknowledgements This work was supported by Academic Excellence Foundation of BUAA for PhD Students, National Natural Science Foundation of China (Grant No. 62272019), China Postdoctoral Science Foundation (Grant No. 2025M774236), and Postdoctoral Fellowship Program of CPSF (Grant No. GZC20242159).

Supporting information Appendixes A–E. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 Fan S, Huang W, Cai X, et al. 3d human interaction generation: a survey. 2025. ArXiv:2503.13120
2 Guo S, Southern R, Chang J, et al. Adaptive motion synthesis for virtual characters: a survey. Vis Comput, 2015, 31: 497–512
3 Sui K, Ghosh A, Hwang I, et al. A survey on human interaction motion generation. 2025. ArXiv:2503.12763
4 Zhu W, Ma X, Ro D, et al. Human motion generation: a survey. IEEE Trans Pattern Anal Mach Intell, 2023, 46: 2430–2449
5 Li M, Wang Y, Leng Z, et al. Fine-grained text-driven dual-human motion generation via dynamic hierarchical interaction. 2025. ArXiv:2510.08260
6 Wang Y, Leng Z, Liu H, et al. Dynamic worlds, dynamic humans: generating virtual human-scene interaction motion in dynamic scenes. 2026. ArXiv:2601.19484
7 Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models. In: Proceedings of Advances in Neural Information Processing Systems, 2020. 33: 6840–6851
8 Lv X, Xu L, Yan Y, et al. Himo: a new benchmark for full-body human interacting with multiple objects. In: Proceedings of European Conference on Computer Vision, 2024. 300–318
9 Zeng L-A, Huang G, Wei Y-L, et al. Chainhoi: joint-based kinematic chain modeling for human-object interaction generation. 2025. ArXiv:2503.13130
10 Li J, Wu J, Liu C K. Object motion guided human motion synthesis. ACM Transactions on Graphics, 2023, 42: 1–11
11 Li J, Clegg A, Mottaghi R, et al. Controllable human-object interaction synthesis. In: Proceedings of European Conference on Computer Vision, 2024. 54–72
12 Peng X, Xie Y, Wu Z, et al. Hoi-diff: text-driven synthesis of 3d human-object interactions using diffusion models. 2023. ArXiv:2312.06553
13 Song W, Zhang X, Li S, et al. Hoianimator: generating text-prompt human-object animations using novel perceptive diffusion models. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024. 811–820
14 Wu Q, Shi Y, Huang X, et al. Thor: text to human-object interaction diffusion via relation intervention. 2024. ArXiv:2403.11208
15 Cha J, Kim J, Yoon J S, et al. Text2hoi: text-guided 3d motion generation for hand-object interaction. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024. 1577–1585
16 Zhang X, Bhatnagar B L, Starke S, et al. Force: dataset and method for intuitive physics guided human-object interaction. 2024. ArXiv:2403.11237v1
17 Ahuja C, Morency L-P. Language2pose: natural language grounded pose forecasting. In: Proceedings of International Conference on 3D Vision, 2019. 719–728
18 Ghosh A, Cheema N, Oguz C, et al. Synthesis of compositional animations from textual descriptions. In: Proceedings of IEEE/CVF International Conference on Computer Vision, 2021. 1396–1406
19 Tevet G, Gordon B, Hertz A, et al. Motionclip: exposing human motion generation to clip space. In: Proceedings of European Conference on Computer Vision, 2022. 358–374
20 Petrovich M, Black M J, Varol G. Temos: generating diverse human motions from textual descriptions. In: Proceedings of European Conference on Computer Vision, 2022. 480–497
21 Guo C, Zou S, Zuo X, et al. Generating diverse and natural 3d human motions from text. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022. 5152–5161
22 Guo C, Zuo X, Wang S, et al. Tm2t: stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts. In: Proceedings of European Conference on Computer Vision, 2022. 580–597
23 Zhang J, Zhang Y, Cun X, et al. Generating human motion from textual descriptions with discrete representations. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023. 14730–14740
24 Zhong C, Hu L, Zhang Z, et al. Attt2m: text-driven human motion generation with multi-perspective attention mechanism. In: Proceedings of IEEE/CVF International Conference on Computer Vision, 2023. 509–519
25 Lucas T, Baradel F, Weinzaepfel P, et al. Posegpt: quantization-based 3d human motion generation and forecasting. In: Proceedings of European Conference on Computer Vision, 2022. 417–435
26 Jiang B, Chen X, Liu W, et al. Motiongpt: human motion as a foreign language. In: Proceedings of Advances in Neural Information Processing Systems, 2023, 36. 20067–20079
27 Zhang Y, Huang D, Liu B, et al. MotionGPT: finetuned LLMs are general-purpose motion generators. AAAI, 2024, 38: 7368–7376

28 Pinyoanuntapong E, Saleem M U, Wang P, et al. Bamm: bidirectional autoregressive motion model. In: Proceedings of European Conference on Computer Vision, 2025. 172–190
29 Pinyoanuntapong E, Wang P, Lee M, et al. Mmm: generative masked motion model. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024. 1546–1555
30 Guo C, Mu Y, Javed M G, et al. Momask: generative masked modeling of 3d human motions. 2023. ArXiv:2312.00063
31 Zhang M, Cai Z, Pan L, et al. MotionDiffuse: text-driven human motion generation with diffusion model. IEEE Trans Pattern Anal Mach Intell, 2024, 46: 4115–4128
32 Tevet G, Raab S, Gordon B, et al. Human motion diffusion model. In: Proceedings of International Conference on Learning Representations, 2023
33 Kim J, Kim J, Choi S. FLAME: free-form language-based motion synthesis & editing. AAAI, 2023, 37: 8255–8263
34 Chen X, Jiang B, Liu W, et al. Executing your commands via motion diffusion in latent space. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023. 18000–18010
35 Wang Y, Leng Z, Li F W B, et al. Fg-t2m: fine-grained text-driven human motion generation via diffusion model. In: Proceedings of IEEE/CVF International Conference on Computer Vision, 2023. 22035–22044
36 Zhang M, Guo X, Pan L, et al. Remodiffuse: retrieval-augmented motion diffusion model. 2023. ArXiv:2304.01116
37 Zhang M, Jin D, Gu C, et al. Large motion model for unified multi-modal motion generation. In: Proceedings of European Conference on Computer Vision, 2025. 397–421
38 Zhang M, Li H, Cai Z, et al. Finemogen: fine-grained spatio-temporal motion generation and editing. In: Proceedings of Advances in Neural Information Processing Systems, 2023. 36: 13981–13992
39 Wang Y, Li M, Liu J, et al. Fg-t2m++: llms-augmented fine-grained text driven human motion generation. In: Proceedings of International Journal of Computer Vision, 2025. 1–17
40 Wang Y, Li M, Leng Z, et al. MOST: motion diffusion model for rare text via temporal clip banzhaf interaction. IEEE Trans Visual Comput Graphics, 2025, 31: 8994–9007
41 Xu S, Wang Y-X, Gui L, et al. Interdreamer: zero-shot text to 3d dynamic human-object interaction. In: Proceedings of Advances in Neural Information Processing Systems, 2024. 37: 52858–52890
42 Taheri O, Ghorbani N, Black M J, et al. Grab: a dataset of whole-body human grasping of objects. In: Proceedings of European Conference on Computer Vision, 2020. 581–600
43 Xu S, Li Z, Wang Y-X, et al. Interdiff: generating 3d human-object interactions with physics-informed diffusion. In: Proceedings of IEEE/CVF International Conference on Computer Vision, 2023. 14928–14940
44 Diller C, Dai A. Cg-hoi: contact-guided 3d human-object interaction generation. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024. 19888–19901
45 Ghosh A, Dabral R, Golyanik V, et al. IMoS: intent-driven full-body motion synthesis for human-object interactions. Comput Graphics Forum, 2023, 42: 1–12
46 Devlin J, Chang M-W, Lee K, et al. Bert: pre-training of deep bidirectional transformers for language understanding. 2018. ArXiv:1810.04805
47 Achiam J, Adler S, Agarwal S, et al. GPT-4 Technical Report. 2023. ArXiv:2303.08774
48 Raffel C, Shazeer N, Roberts A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. J Mach Learn Res, 2020, 21: 5485–5551
49 Kalakonda S S, Maheshwari S, Sarvadevabhatla R K. Action-GPT: leveraging large-scale language models for improved and generalized action generation. In: Proceedings of IEEE International Conference on Multimedia and Expo, 2023. 31–36
50 Athanasiou N, Petrovich M, Black M J, et al. Sinc: spatial composition of 3d human motions for simultaneous action generation. 2023. ArXiv:2304.10417
51 Jacobs R A, Jordan M I, Nowlan S J, et al. Adaptive mixtures of local experts. Neural Computation, 1991, 3: 79–87
52 Eigen D, Ranzato M, Sutskever I. Learning factored representations in a deep mixture of experts. 2013. ArXiv:1312.4314
53 Zoph B, Bello I, Kumar S, et al. St-moe: designing stable and transferable sparse expert models. 2022. ArXiv:2202.08906
54 Lepikhin D, Lee H, Xu Y, et al. Gshard: scaling giant models with conditional computation and automatic sharding. 2020. ArXiv:2006.16668
55 Pavlakos G, Choutas V, Ghorbani N, et al. Expressive body capture: 3d hands, face, and body from a single image. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019. 10975–10985
56 Zhou Y, Barnes C, Lu J, et al. On the continuity of rotation representations in neural networks. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019. 5745–5753
57 Liu A, Feng B, Xue B, et al. Deepseek-v3 technical report. 2024. ArXiv:2412.19437
58 Betker J, Goh G, Jing L, et al. Improving image generation with better captions. Comput Sci, 2023, 2: 8
59 Liu Y, Zhang K, Li Y, et al. Sora: a review on background, technology, limitations, and opportunities of large vision models. 2024. ArXiv:2402.17177
60 Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024
61 Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision. In: Proceedings of International Conference on Machine Learning, 2021. 8748–8763
62 Qi C R, Su H, Mo K, et al. Pointnet: deep learning on point sets for 3d classification and segmentation. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2017. 652–660
63 Fan Z, Sarkar R, Jiang Z, et al. M^3vit: mixture-of-experts vision transformer for efficient multi-task learning with model-accelerator co-design. In: Proceedings of Advances in Neural Information Processing Systems, 2022. 35: 28441–28457
64 Chen T, Chen X, Du X, et al. Adamv-moe: adaptive multi-task vision mixture-of-experts. In: Proceedings of IEEE/CVF International Conference on Computer Vision, 2023. 17346–17357
65 Du N, Huang Y, Dai A M, et al. Glam: efficient scaling of language models with mixture-of-experts. In: Proceedings of International Conference on Machine Learning, 2022. 5547–5569
66 Fedus W, Zoph B, Shazeer N. Switch transformers: scaling to trillion parameter models with simple and efficient sparsity. J Mach Learn Res, 2022, 23: 1–39
67 Perez E, Strub F, De Vries H, et al. Film: visual reasoning with a general conditioning layer. In: Proceedings of AAAI Conference on Artificial Intelligence, 2018. 32
68 Shafir Y, Tevet G, Kapon R, et al. Human motion diffusion as a generative prior. In: Proceedings of International Conference on Learning Representations, 2023