HOIMask: Towards Generative Masked Modeling for Human Object Interaction Generation

Yihong Ji[1,2⋆], Jinsong Zhang[3], He Hu[1,2], and Hongbo Xu[2]

[1] College of Computer Science and Software Engineering, Shenzhen University
[2] Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ)
[3] Fuzhou University

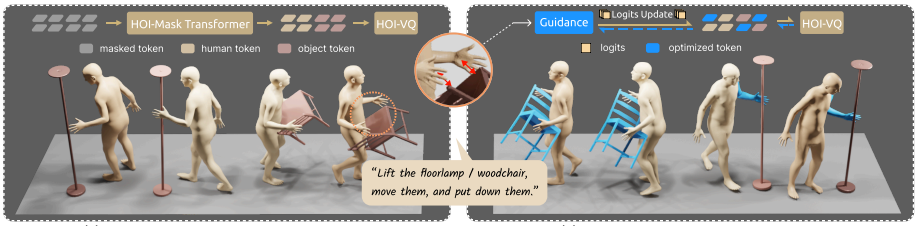

Fig. 1: HOIMask is a generative masked model to synthesis high fidelity Human-Object Interaction (HOI) motion from text description and object geometry. (a) Achieving accurate spatial and temporal coordination with our proposed HOI VQ-VAE and HOI-Mask Transformer, and (b) guiding the model to generate more rational interaction through Contact-aware Reconstruction Guidance.

Abstract. Diffusion-based methods have dominated the HOI generation, as they enable critical contact fusions or signals to guide the diffusion process. However, they often result in high artifacts and unstable interaction quality due to error accumulation during iterative denoising. In this work, we propose HOIMask, the first generative masked framework for modeling HOI motion in discrete space. HOIMask first encodes both motion sequences and contact-aware signals into discrete 2D human and object token maps via HOI Vector Quantization (VQ), preserving fine-grained spatial-temporal structure beyond conventional 1D representations. On this basis, a generative masked modeling framework is employed to jointly capture human-object interaction dynamics, leveraging a transformer architecture designed to model complex spatial-temporal and interaction dependencies. To generate more coherent and physically plausible motions, we further introduce a novel contact-aware reconstruction guidance in discrete space during inference, which fuses contact signals to optimize HOI tokens that forces the generated motion with higher spatio-temporal consistency. With craftily designed motion interaction tokens, dedicated architecture and guidance strategy, HOIMask outperforms state-of-the-art diffusion-based methods, generating more realistic and semantically aligned HOI motions. Please refer to our project page for more results.

⋆ Corresponding author.

Keywords: Human-Object Interaction Generation · Generative Masked Model · VQ-VAE

1 Introduction

HOI generation has attracted significant attention and plays an important role in a broad range of applications in VR, animation and spatial-intelligence. Text-driven HOI generation [2, 5, 10, 37, 60] synthesizes vivid and coherent human-object motions guided by textual descriptions. Achieving realistic HOI requires synthesizing natural human and object motions while deeply understanding the underlying spatio-temporal dynamics between them.

Recently many methods [2,5,10,14,23,27,37,55,58,60] generate HOI motions based on diffusion models [9, 15, 36, 44, 45]. HOI-Diff [37] introduces a two-stage diffusion framework to generate the contact and motion. CHOIS [27] incorporates additional control signals like object waypoints to generate more stable and long-term HOI motions. Recent studies further focus on modeling joint-level interactions [17,60] or designing new interaction representations [14, 38, 58] to improve the quality of generated motions. However, these methods apply diffusion processes to raw HOI motion sequences which involves strong kinematic constraints and tight human–object coupling, where small denoising errors can accumulate across iterative sampling steps, leading to unstable trajectories and physically implausible interactions.

Meanwhile, autoregressive models [35,49,51] like generative masked model [3] further enhance motion generation fidelity by capturing temporal coherence within motion sequences [11, 34, 39–41] compared with diffusion models. Motion autoregressive methods model sequential dependencies by predicting each motion token conditioned on the text token and previously generated tokens. This autoregressive decoding mechanism inherently captures long-range temporal dependencies, resulting in enhanced coherence and generation fidelity, showing great potential for HOI generation. However, due to the inherent complexity of HOI motions, it is essential not only to model the individual temporal dynamics of the human and the object, but also to capture their intricate spatial interaction dependencies.

In this work, we present HOIMask, a novel framework for HOI motion generation based on generative masked modeling in discrete space. First, we propose a 2D HOI VQ-VAE to model HOI motions into discrete 2D human and object token maps using two learned codebooks. In contrast to 1D VQ encodings [18] that disregard spatial interactions, our method captures the intricate interaction motion inherent in each HOI sequence. Using 2D convolutions, we generate 2D HOI token map, which captures both human and object motions along with contact information, enabling better spatial awareness. Second, we design a HOI-Mask Transformer to collaboratively model human and object tokens. The HOI-Mask Transformer incorporates spatial-temporal attention block to model detailed motion dependencies, along with human/object-centric attention blocks to capture HOI motions relationships. This architecture allows the

model to reason jointly about spatial and temporal dynamics within and across individuals, thereby enhancing its ability to generate interactive HOI motion sequences. Third, to generate more physically plausible interactions, we introduce a novel contact-aware reconstruction guidance during discrete sampling process. We define a discrete probability gradient based on predicted logits, guided by a reconstruction objective at inference stage. The reconstruction objective contains contact loss and penetration loss, and we propose a differentiable Logits Update strategy to optimize the HOI tokens. This guidance enables the HOI-Mask Transformer to optimize logits effectively, resulting in higher-quality and more coherent interactions.

Our main contributions are as follows:

- We propose HOIMask, the first Vector Quantization (VQ)-based generative masked modeling framework for full-body HOI motion.
- We introduce two key components of HOIMask: (1) a 2D HOI VQ-VAE for capturing complex spatial and temporal interaction to reconstruct high-fidelity HOI motions. (2) an interaction-aware HOI-Mask Transformer for modeling human-object interaction dependencies to generate accurate HOI tokens.
- We introduce a novel contact-aware reconstruction guidance in discrete sampling space. Specifically, we design a logits update method with differentiable strategy during inference to optimize the HOI tokens which represent more rational HOI motions.
- We achieve state-of-the-art performance on FullBodyManipulation [28] and BEHAVE [1] datasets compared to the diffusion-based methods.

2 Related work

2.1 Human Object Interaction Generation

Synthesizing HOI motions has recently gained increasing attention. Extensive datasets [1, 19, 28, 31, 65] capture HOI motions that includes complex interactions and multiple object categories. GRAB [47] is one of the earliest datasets for full-body human–object interaction. BEHAVE [1] and OMOMO [28] provide more complex interaction scenarios, while IMHD [65] contains highly dynamic interactions such as sports activities. Recent works like Text2HOI [2] focuses on hand-object interaction generation, while InterDiff [57] predicts full-body and object motions. HOI-Diff [37] and CG-HOI [10] incorporate contact guidance, and CHOIS [27] introduces waypoint-based controllability. ChainHOI [60] enhances interaction accuracy through joint-level representations, ROG [58] introduces an Interactive Distance Field (IDF) for interaction modeling, and SyncDiff [14] presents a frequency-domain modeling paradigm. However, these diffusion-based methods often suffer from error accumulation and inconsistent interactions. In contrast, we present the first VQ-based generative masked model that generates more rational and coherent HOI motions from text.

2.2 Motion Quantization

Encoding human motions into discrete tokens has been extensively studied in single-human motion generation [13, 24, 32, 33, 62, 67, 68]. TM2T [13], T2M-GPT [62], and MotionGPT [24] establish mappings between motions and discrete tokens. For HOI representation, HOIGPT [18] introduces a dual VQ-VAE [50] for encoding hand-object motions, but its 1D quantization amplifies interaction errors and simply quantizing HOI motions can't capture the inherent interaction. In this work, we propose a contact-aware HOI quantization that transforms HOI motions and corresponding contact information into discrete 2D token maps, capturing fine-grained spatio-temporal interactions.

2.3 Generative Masked Modeling

Recent studies show that generative masked modeling methods [11, 22, 34, 39–41] outperform diffusion-based approaches [4, 7, 8, 26, 29, 43, 48, 52, 61, 63, 64, 66] in generating realistic single-human motions. VQ-based models such as MoMask [11] and MMM [41] employ masked transformers to predict all tokens jointly, while InterMask [21] extends this to human-human interactions. However, in the HOI generation, diffusion-based methods still dominate. Motivated by these advances, we develop a HOI VQ-VAE and HOI-Mask Transformer to generate high-fidelity, contact-aware human-object interactions.

3 Methodology

Overview Given object geometry and text description, our objective is generating HOI motion. As shown in Figure 2 (a), we start by learning discrete latent representations via HVQ and OVQ (3.1) to encode human-object motions and their relative contact information into 2D tokens. Then the HOI-Mask Transformer (3.2) generates 2D HOI tokens conditioned on text and object geometry, which are decoded by the HOI-Decoder to reconstruct HOI motions. During inference, the contact-aware reconstruction guidance (3.3) is applied to optimize the HOI tokens for more physically plausible interactions.

3.1 HOI VQ-VAE

HVQ The human motion $M_h \in \mathbb{R}^{T\times J}$ is accompanied by a contact state $C_h \in \mathbb{R}^{T\times J}$, indicating whether each joint is in contact with the object at every time step. Given M_h and C_h, a 2D spatio-temporal token map $m_h \in \mathbb{R}^{T\times J\times D}$ is constructed (T is the frames of motion sequence, J is the number of joints, and D is the joint feature dimension), where spatial and temporal dependencies are jointly captured through 2D convolutions. The resulting token maps are then encoded into 2D latent features $\tilde{l}_h \in \mathbb{R}^{t\times j\times d}$ (with downsampling ratios $T/t, J/j$), and each d-dimensional feature vector is discretized by mapping it to the nearest entry in a learnable human codebook, producing a quantized sequence $l_h = Q(\tilde{l}_h)$.

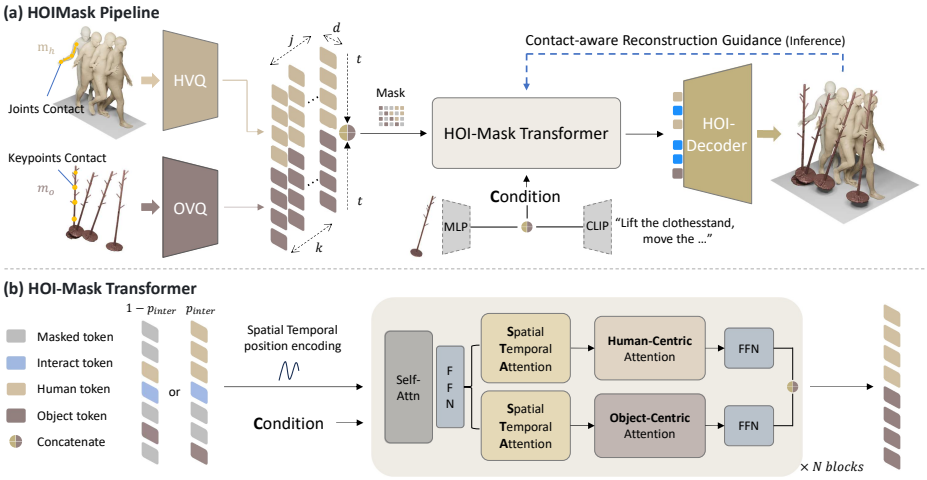

Fig. 2: Overview of **HOIMask**. (a) During training, HOI motions first are quantized via HVQ and OVQ into 2D HOI tokens, which are partially masked and predicted by the HOI-Mask Transformer. During inference, the HOI tokens are optimized with contact-aware reconstruction guidance to generate more physically plausible interactions through HOI-Decoder. (b) The HOI-Mask Transformer incorporates spatial-temporal attention and human/object-centric attention to model fine-grained spatio-temporal dependencies.

OVQ The object motion $M_o \in \mathbb{R}^T$ includes the object's centroid rotation and global 3D position. We uniformly sample K surface points per frame via Poisson Disk Sampling [59] and pre-calculate their contact states. Similar to the human branch, the motion and contact signals are concatenated into a 2D object token map $m_o \in \mathbb{R}^{T\times K\times D}$ and encoded into latent features $\tilde{l}_o \in \mathbb{R}^{t\times k\times d}$.

HOI-Decoder The quantized HOI sequence is decoded into motion using a dual-branch 2D HOI-Decoder for humans and objects. The object decoder first generates the object motion conditioned on the object geometry, followed by the human decoder, which generates the human motion conditioned on the object motion and geometry. By conditioning the human motion on the predicted object trajectory and contact states, the generated human movements are constrained to align with the object's spatial and geometric characteristics, ensuring physically plausible interactions.

Loss Following [18], we first incorporate a reconstruction loss and the commitment loss [50]:

$$\mathcal{L}_{vq} = \|m_h - \hat{m}_h\|_1 + \|m_o - \hat{m}_o\|_1 + \beta\|\tilde{l}_h - \text{sg}(l_h)\|_2^2 + \beta\|\tilde{l}_o - \text{sg}(l_o)\|_2^2, \quad (1)$$

where $\text{sg}(\cdot)$ indicates the stop-gradient operation, and β serves as a weighting coefficient. To reconstruct both high-fidelity motions and high-quality interactions, we define a training objective composed of a motion loss and an interaction loss.

Motion Loss The motions of human and object can be decomposed into global

positions m_p and relative rotations m_r. The motion loss is defined as the combination of the L1 loss on the predicted positions and the rotation loss relative to the ground truth.

$$\mathcal{L}_{motion} = \|m_p - \hat{m}_p\|_1 + \|m_r - \hat{m}_r\|_1. \tag{2}$$

Interaction Loss We compute the interaction loss by measuring the spatial relationship between the Poisson Disk Sampled objects K_o and the human pose. Specifically, following [56], we compute the object points' positions in the human frame, formulate as:

$$\hat{N}_h = {R_K}^{-1}(K_o - T_h), \tag{3}$$

where R_K and T_h represent the rotation of the sampled object and the position of the human, respectively. While [56] focuses solely on hand-level and object-centric interactions, our formulation extends this to full-body and bi-centric modeling, enabling the model to jointly reason about both human and object motions. Furthermore, the human/object-centric contact maps C_h and C_o are used to mask out when the human and object are not in contact. The interaction loss is formulated as:

$$\mathcal{L}_{interact} = \sum_{t=1}^{T} C_h \left\| \hat{N}_{h,t} - N_{h,t} \right\|_1 + C_o \left\| \hat{N}_{o,t} - N_{o,t} \right\|_1. \tag{4}$$

The overall loss function is a weighted (λ_{motion}, $\lambda_{interact}$) sum of motion and interaction loss. The human and object codebooks are jointly updated using Exponential Moving Average (EMA) with periodic resets to ensure stable and consistent quantization, following [62].

$$\mathcal{L}_{\text{vqvae}} = \mathcal{L}_{vq} + \lambda_{\text{motion}}\mathcal{L}_{motion} + \lambda_{interact}\mathcal{L}_{interact}. \tag{5}$$

3.2 HOI-Mask Transformer

Subsequently, the motion tokens of both the human and the object are jointly processed by our HOI-Mask Transformer, which captures rich spatial-temporal dependencies within each entity and across their interactions. An additional interaction token is introduced to explicitly separate human and object tokens, facilitating inter-entity communication. The total input tokens sequence $t' \in \mathbb{Z}^{t(j+k)+1}$ is formed by concatenating the flattened 1D human and object tokens with the interaction token. We use the 2D positional encodings [53] to impose spatial-temporal structure on the embeddings.

HOI-Mask Mechanism As shown in Figure 2 (b), we employ two masking strategies. The first is a random masking strategy, controlled by a cosine scheduling function [3] $\gamma(\tau) = \cos\left(\frac{\pi\tau}{2}\right) \in [0, 1]$, where $\tau \in [0, 1]$ that $\tau = 0$ means the sequence is completely corrupted. During random masking, the HOI tokens are randomly masked with ratio $\gamma(\tau)$. To further enhance interaction modeling, we introduce a human-object masking strategy with probability p_{inter}. Specifically, the human-object masking randomly masks either the object or the human, and

the other one is set to be fully unmasked. This mechanism improves the human and object tokens' dependencies.

STA In the Transformer module, the 2D HOI tokens are first processed by a self-attention layer [51] to compute global attention scores. We then decouple the tokens and introduce a Spatial-Temporal Attention (STA) to respectively model the human and object tokens. The STA is composed of two distinct attention mechanisms: spatial attention and temporal attention. In the spatial attention branch, tokens are restricted to attend to other spatial tokens within the same time step. In the temporal attention branch, the same operations will be executed at the same spatial position. The outputs of both branches are finally aggregated through element-wise addition to hybrid spatial and temporal information for more coherent motion representation:

$$STA = Attn_{(Q_j,K_j,V_j)} + Attn_{(Q_t,K_t,V_t)}, \tag{6}$$

where Q_j, K_j, V_j and Q_t, K_t, V_t are the spatial and temporal queries, keys and values of human and object.

HOCA We further introduce a Human/Object-Centric Attention (HOCA) to enhance cross-entity interaction. The *STA* features from the human and object branches are fed into the HOCA to exchange information and strengthen the human and object tokens dependencies. In this module, each token of STA_{human} or STA_{object} attends to the entire set of another *STA* feature, enabling comprehensive modeling of cross-entity dependencies.

Training The objective of HOI-Mask Transformer is to predict masked tokens conditioned on text and object geometry. Text features are extracted using a frozen CLIP [42], while object features are encoded via an MLP. These are concatenated as the final condition $[c = c_t|c_o]$. Given the masked sequence $\tilde{t}$, the model is trained to minimize the negative log-likelihood of the masked tokens, i.e., the cross-entropy loss between predicted probability distribution and one-hot ground truth:

$$\mathcal{L}_{\text{mask}} = \sum_{\tilde{t}_k \in [\text{MASK}]} -\log p_\theta(t_k \mid \tilde{t}, c). \tag{7}$$

Inference At inference, the HOI-Mask Transformer begins with all positions masked and iteratively generates tokens over N steps. At each step n, it predicts token distributions, samples candidates, and re-masks low-confidence tokens $\lceil \gamma(\frac{n}{N} \cdot (t(j+k)+1)) \rceil$ for refinement. This process continues until $n = N$, after which the HOI-Decoder reconstructs motion from the generated 2D HOI tokens. Classifier-free guidance [16] is further employed to improve generation diversity and controllability under conditional and unconditional settings.

3.3 Contact-aware Reconstruction Guidance

The HOI-Mask Transformer aims to generate tokens semantically aligned with the given conditions. However, semantic relevance alone does not guarantee

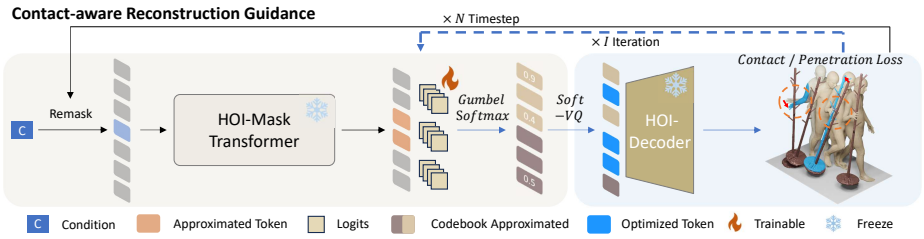

Fig. 3: At Inference stage, Contact-aware Reconstruction Guidance is applied to optimize HOI tokens. The predicted logits are updated and pass through Soft Vector Quantization (Soft-VQ) to obtain optimized HOI tokens in differentiable paradigm during every iteration.

physically plausible interactions. Irrational motions may happen because the human-object tokens lack contact-guided coupling between HOI motions. The same problem also occurs in diffusion-based generation models, and some methods [10, 27, 54] design contact guidance during the denoising process to improve interaction quality. Unlike diffusion models that directly perform guidance in continuous space, applying guidance in discrete token space is considerably more challenging due to the non-differentiability of token representations. Therefore, an effective guidance strategy should enable diffusion-like differentiable optimization while effectively refining the human-object tokens. To this end, we introduce a contact-aware reconstruction guidance that incorporates contact signals to fully exploit the model's potential in HOI generation during inference.

We first formulate two key objectives: a contact loss and a penetration loss. As described in Section 3.1, the 2D HOI tokens encode both motion dynamics and contact-aware signals. At every timestep n, the HOI-Mask Transformer predicts HOI tokens, which are decoded into motion sequences and corresponding contact maps $\tilde{C}_h$ and $\tilde{C}_o$. These reconstructed maps are used to compute the contact-aware guidance loss, serving as a crucial signal for improving interaction quality.
Contact Loss The predicted contact representation includes both human joints and object keypoints contact maps. We apply a thresholding of 0.95 to $\tilde{C}_h$ and $\tilde{C}_o$ to determine whether a contact occurs. Given the SMPL [30] parameters and the generated HOI motions, we compute the 3D positions of the human joints J_h and determine which joints are in contact with the object mesh $\tilde{K}_o$, as well as which object regions are contacted by the human throughout the sequence. The contact loss is then formulated as:

$$\mathcal{L}_c = \tilde{C}_h \|J_h - \tilde{K}_o\|_1 + \tilde{C}_o \|\tilde{K}_o - J_h\|_1. \tag{8}$$

Penetration Loss We design a penetration loss to constrain interpenetration between the human and the object, as well as penetration into the floor. Following [18, 25, 27], for each human joint $J \in \mathcal{J}_{in}^{o}$ that penetrates the object mesh, we find the closest object vertex and minimize the squared distance. Similar to the human-object penetration, we set a floor height of $(z = 0.02)$ meters to evaluate whether human feet or object keypoints penetrate the floor. The penetration

loss is defined as follows:

$$\mathcal{L}_{h_p} = \frac{1}{\mathcal{J}_{in}^o} \sum_{J \in \mathcal{J}_{in}^o} min \|J - \tilde{K_o}\|_2^2, \tag{9}$$

$$\mathcal{L}_p = \lambda_1 \mathcal{L}_{h_p} + \lambda_2 min \|J_h^z - z\|_2 + \lambda_3 min \|\tilde{K_o^z} - z\|_1, \tag{10}$$

where λ_1, λ_2, λ_3 denote the loss weights, J_h^z and $\tilde{K_o^z}$ represent the z-coordinates of the human feet and object vertices.

Logits Update After calculating the contact and penetration loss, we back-propagate their gradients to the logits generated by the classifier-free guided HOI-Mask Transformer. VQ-based model natively decode the final indices generated from the Transformer process, which is non-differentiable and it is needed for gradient in reconstruction guidance. Recently, [39] tackle the similar problem using a Differentiable Expectation Sampling (DES). To address this, we introduce a similar differentiable strategy that estimates the probability distribution over codebook indices and applies Gumbel-Softmax [20] sampling to approximate categorical sampling in a differentiable paradigm. Further, we propose a Soft-VQ in HOI-Decoder to transform these updated logits into the HOI tokens.

As shown in Figure 3, we set an iteration steps I in every timestep n. In timestep n, the HOI-Mask Transformer generates the logits, and the optimization process starts in iteration i. During training or optimization, motion tokens are sampled from a categorical distribution, a process that is inherently non-differentiable. To make this process differentiable for backpropagation, we apply the Gumbel-Softmax to sample from a categorical distribution.

$$p_\theta\left(x_k \mid X_{\overline{\mathbf{M}}}, W, S\right) = \frac{\exp\left((\ell_k + g_k)/\tau\right)}{\sum_{j=1}^{K} \exp\left((\ell_j + g_j)/\tau\right)}, \tag{11}$$

where ℓ, τ represent the logits and temperature, g denotes the Gumbel noise with $g_1, ..., g_k$ are independent and identically distributed (i.i.d.) samples from a Gumbel (0, 1) distribution. The logits are passed through the Gumbel-Softmax to obtain a differentiable approximation of discrete token probabilities. The predicted probability p is utilized to reconstruct HOI motion through HOI-Decoder. Contact and penetration loss are applied to provide gradient signals with respect to the logits. Subsequently, the logits at iteration i+1 are updated using the gradient-based guidance as follows:

$$\mathcal{L}_{i+1} = L_i - \lambda_g \nabla_{L_t} L_r(l_i, r), \tag{12}$$

where λ_g is the guidance scale, $L_r(l_i, r)$ represents the gradient of the reconstruction loss L_r. The HOI-Decoder uses the predicted probability p to reconstruct HOI motion by querying the discrete token embedding using the index with the highest probability, $argmax_s p_\theta(x_s)$, this kind of hard quantization which is also non-differentiable. To obtain continuous and differentiable representations from discrete codebooks, we employ a Soft-VQ module. Unlike standard vector quantization that selects the nearest codeword via a hard assignment, Soft-VQ

computes a probabilistic mixture of all codewords based on their probabilities. Formally, given the probability $p_{n,s}$ for the s-th codeword in the quantizer, and the corresponding codebook entries $\mathbf{e}_s \in \mathbb{R}^d$, the soft code embedding is obtained as:

$$\mathbf{z}_n = \sum_{S=1}^{S} p_{n,s}\, \mathbf{e}_s. \tag{13}$$

4 Experiments

4.1 Dataset

FullBodyManipulation Dataset We conduct our experiments on the FullBodyManipulation dataset [28], a comprehensive, large-scale collection that contains 10 hours human-object interaction motion data, accompanied by corresponding videos of 17 subjects interacting with 15 different objects. We follow the train-test split of CHOIS for both training and evaluation.
BEHAVE Dataset This dataset [1] contains 20 diverse objects and 8 subjects. We follow the official train-test split and use the version processed by the HOI-Diff [37].

4.2 Baseline methods

In our experiment, we compare a series of diffusion-based methods with our VQ-based generative masked model.

- **MDM*** We extend the MDM [48] by incorporating object motion to enable HOI generation.
- **InterDiff*** We adapt InterDiff [57], a general motion diffusion model, by introducing text conditioning for HOI generation.
- **HOI-Diff** [37] employs a dual diffusion model for modeling human and object motions. We follow its official configuration for comparison.
- **CHOIS*** [27] introduces additional control signals such as object waypoints. We remove them to align it with our setting.
- **SemGeoMo*** [6] takes the object trajectory as additional condition, we remove this conditioning and treat it as the objective to maintain consistency.
- **ROG** [58] constructs an Interactive Distance Field (IDF) to generate HOI motions, we keep the official setting to comparison.
- **LIGHT** [54] builds upon diffusion forcing and introduces a data-driven guidance strategy.

4.3 Evaluation Metric

Motion Quality Metric We first evaluate the motion quality metrics following [12,46]. Fréchet Inception Distance (FID) evaluates how closely the generated motions align with the ground truth motions. R-Precision (Top3) (R-Pre$^{(Top3)}$),

Table 1: Quantitative comparisons on the **FullBodyManipulation** and **BEHAVE** datasets. Bold indicates the best result.

FullBodyManipulation Dataset						
Method	FID$^{\downarrow}$	Diversity$^{\rightarrow}$	R-Pre$^{(Top3)\uparrow}$	PS$^{\downarrow}$	$C_{prec}{}^{\uparrow}$	$C_{\%}$
GT	$0.01^{\pm 0.000}$	$1.756^{\pm 0.082}$	$0.750^{\pm 0.000}$	$0.060^{\pm 0.000}$	–	$70.68^{\pm 0.000}$
Diffusion Models						
MDM* [48]	$4.273^{\pm 0.001}$	$2.391^{\pm 0.082}$	$0.405^{\pm 0.001}$	$0.088^{\pm 0.001}$	$0.39^{\pm 0.007}$	$22.56^{\pm 0.095}$
InterDiff* [57]	$6.890^{\pm 0.002}$	$2.290^{\pm 0.053}$	$0.605^{\pm 0.000}$	$0.079^{\pm 0.001}$	$0.42^{\pm 0.004}$	$25.63^{\pm 0.100}$
HOI-Diff [37]	$1.875^{\pm 0.000}$	$2.463^{\pm 0.079}$	$0.792^{\pm 0.000}$	$0.068^{\pm 0.001}$	$0.57^{\pm 0.005}$	$29.20^{\pm 0.090}$
CHOIS* [27]	$0.810^{\pm 0.001}$	$2.130^{\pm 0.050}$	$0.800^{\pm 0.001}$	$0.066^{\pm 0.002}$	$0.71^{\pm 0.005}$	$45.46^{\pm 0.107}$
SemGeoMo* [6]	$1.325^{\pm 0.000}$	$1.453^{\pm 0.057}$	$0.778^{\pm 0.002}$	$0.070^{\pm 0.002}$	$0.55^{\pm 0.007}$	$40.25^{\pm 0.090}$
ROG [58]	$0.199^{\pm 0.002}$	$1.646^{\pm 0.066}$	$0.817^{\pm 0.000}$	$\mathbf{0.054}^{\pm 0.001}$	$0.47^{\pm 0.008}$	$27.56^{\pm 0.120}$
LIGHT [54]	$\mathbf{0.105}^{\pm 0.002}$	$\mathbf{1.739}^{\pm 0.070}$	$0.779^{\pm 0.000}$	$0.088^{\pm 0.002}$	$0.71^{\pm 0.007}$	$49.57^{\pm 0.091}$
Generative Masked Model						
HOIMask	$0.109^{\pm 0.000}$	$1.845^{\pm 0.083}$	$\mathbf{0.826}^{\pm 0.000}$	$0.065^{\pm 0.000}$	$\mathbf{0.74}^{\pm 0.005}$	$\mathbf{52.96}^{\pm 0.080}$
BEHAVE Dataset						
Method	FID$^{\downarrow}$	Diversity$^{\rightarrow}$	R-Pre$^{(Top3)\uparrow}$	PS$^{\downarrow}$	$C_{\%}$	
GT	$0.01^{\pm 0.000}$	$1.896^{\pm 0.047}$	$0.713^{\pm 0.000}$	$0.088^{\pm 0.000}$	$74.20^{\pm 0.000}$	
Diffusion Models						
InterDiff* [57]	$1.060^{\pm 0.000}$	$1.697^{\pm 0.097}$	$0.554^{\pm 0.002}$	$0.145^{\pm 0.002}$	$22.59^{\pm 0.113}$	
HOI-Diff [37]	$0.725^{\pm 0.000}$	$1.743^{\pm 0.089}$	$0.517^{\pm 0.001}$	$0.178^{\pm 0.001}$	$36.31^{\pm 0.104}$	
CHOIS* [27]	$0.571^{\pm 0.002}$	$1.710^{\pm 0.128}$	$0.606^{\pm 0.002}$	$0.156^{\pm 0.001}$	$41.70^{\pm 0.087}$	
SemGeoMo* [6]	$0.985^{\pm 0.001}$	$1.529^{\pm 0.104}$	$0.452^{\pm 0.000}$	$0.165^{\pm 0.002}$	$43.06^{\pm 0.121}$	
ROG [58]	$0.490^{\pm 0.000}$	$\mathbf{1.887}^{\pm 0.119}$	$0.595^{\pm 0.002}$	$0.149^{\pm 0.004}$	$25.71^{\pm 0.078}$	
LIGHT [54]	$0.467^{\pm 0.005}$	$1.861^{\pm 0.087}$	$0.561^{\pm 0.000}$	$\mathbf{0.135}^{\pm 0.002}$	$65.89^{\pm 0.074}$	
Generative Masked Model						
HOIMask	$\mathbf{0.424}^{\pm 0.001}$	$1.873^{\pm 0.098}$	$\mathbf{0.606}^{\pm 0.001}$	$0.141^{\pm 0.007}$	$\mathbf{66.90}^{\pm 0.100}$	

which evaluate semantic alignment between the input text and generated motions. Diversity quantifies the variation within the generated motions.

Interaction Quality Metric We validate the interaction quality on Penetration Score (PS), Contact Precision (C_{prec}) and Contact Percentage ($C_{\%}$) [27]. PS calculates the Signed Distance Function (SDF) values between the human and object meshes in each sequence. C_{prec} measures the accuracy of predicted contact regions. $C_{\%}$ measures the ratio of frames in which human-object interactions are identified as being in contact.

4.4 Results

Quantitative Comparison Table 1 present quantitative comparisons between our **HOIMask** and various diffusion-based baselines. The results demonstrate that HOIMask achieves state-of-the-art performance, with substantial improvements in FID, R-Pre$^{(Top3)}$, C_{prec} and $C_{\%}$. A lower FID indicates superior realism and visual fidelity of the generated interactions, while the highest R-Precision reflects stronger semantic alignment between motion and textual descriptions. The highest $C_{\%}$ and lowest PS confirm that our model produces more physically

plausible and coherent human object interactions.

Qualitative Comparison Figure 4 presents qualitative comparisons between HOIMask, CHOIS [27], ROG [58], and SemGeoMo [6] across three different objects. HOIMask generates motions that are semantically aligned with text inputs and maintain realistic interactions. For the prompt "lift the trashcan, rotate it, and set it back down," only HOIMask successfully performs the rotation. Overall, HOIMask produces physically plausible and contact-aware interactions, while other methods often show semantic misalignment or penetration artifacts.

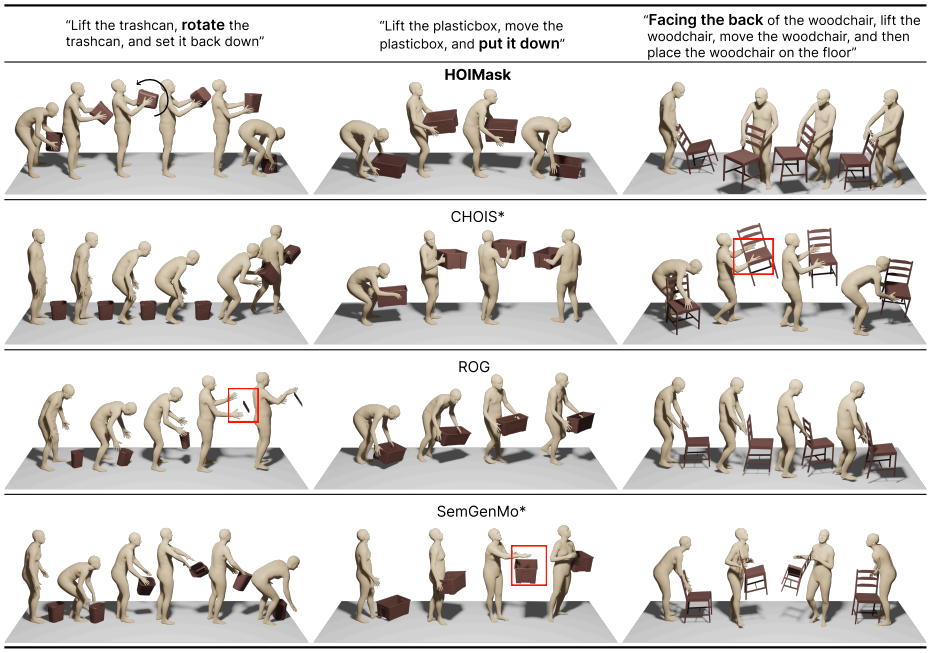

Fig. 4: Qualitative comparison on the FullBodyManipulation dataset. We use red boxes to highlight the incorrect interaction. The rotating arrow demonstrates the successful action "rotate". In contrast, our method generates more realistic and physically consistent human-object interactions than the compared approaches.

User Study We further conduct a user study to validate the generation quality. We randomly generate 20 HOI sequences from our method and a series of diffusion-based methods, and 20 participants were asked to choose the method that better matched the given text description or demonstrated superior interaction quality. As shown in Figure 5, HOIMask outperforms other methods in interaction quality and text adherence.

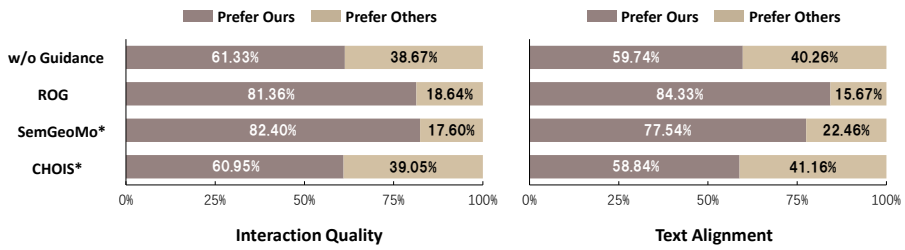

Fig. 5: User Study. The color bar and numbers represent the preference rate.

5 Ablation Studies

We conduct a series of ablation studies to validate the key components of our HOIMask framework. We first examine the effectiveness of the 2D HOI VQ-VAE design, followed by an analysis of the core modules within the HOI-Mask Transformer. Finally, we evaluate the contribution of the Contact-aware Reconstruction Guidance in enhancing the interaction quality.

Table 2: Ablation Study results on the FullBodyManipulation test set demonstrating the effectiveness of each key component in the proposed **HOI-Mask Transformer**. **Bold** face indicates the best result.

Method	FID ↓	Diversity →	R-Prec ↑	C_{prec} ↑	$C_{\%}$
w/o STA	$0.505^{\pm 0.001}$	$1.126^{\pm 0.044}$	$0.781^{\pm 0.001}$	$0.69^{\pm 0.008}$	$43.92^{\pm 0.056}$
w/o HOCA	$0.242^{\pm 0.000}$	$1.496^{\pm 0.098}$	$0.794^{\pm 0.000}$	$0.73^{\pm 0.005}$	$52.50^{\pm 0.055}$
w/o Guidance	$0.134^{\pm 0.000}$	$1.864^{\pm 0.084}$	$0.800^{\pm 0.001}$	$0.64^{\pm 0.006}$	$37.54^{\pm 0.067}$
HOIMask	$\mathbf{0.109}^{\pm 0.000}$	$\mathbf{1.845}^{\pm 0.083}$	$\mathbf{0.826}^{\pm 0.000}$	$\mathbf{0.74}^{\pm 0.005}$	$\mathbf{52.96}^{\pm 0.080}$

5.1 2D or 1D Tokens

In Table 3, we compare our 2D HOI-VQ-VAE with a 1D variant that encodes HOI motion using a 1D convolutional encoder. The results show that whether in motion or interaction quality, the 2D quantization process significantly reduce errors. Moreover, removing either the Interaction Loss or Motion Loss leads to clear performance degradation, highlighting their critical roles in VQ training.

5.2 HOI-Mask Transformer

Table 2 presents an ablation study to evaluate the key components of HOI-Mask Transformer. The results show that removing the STA, Human/Object Centric

Table 3: Ablation study results on the FullBodyManipulation test set demonstrating the effectiveness of each key component in the proposed **HOI VQ-VAE**. **Bold** face indicates the best result.

Method	MPJPE ↓	R_o ↓	$C_{\%}$	PS ↓
GT	0.00	0.00	70.68	0.060
1D-VQVAE	65.43	0.26	43.15	0.063
w/o Interaction Loss	**27.56**	0.10	60.73	0.062
w/o Motion Loss	35.87	0.12	55.41	0.061
2D-VQVAE	28.95	**0.08**	**65.33**	**0.061**

Table 4: Ablation study results on **Contact-aware Reconstruction Guidance**.

Method	FID ↓	Diversity →	R-Prec ↑	$C_{\%}$	PS ↓
w/o Contact Loss	$0.121^{\pm 0.001}$	$1.880^{\pm 0.041}$	$0.779^{\pm 0.002}$	$39.01^{\pm 0.082}$	$\mathbf{0.061}^{\pm 0.000}$
w/o Penetration Loss	$0.114^{\pm 0.000}$	$1.866^{\pm 0.075}$	$0.814^{\pm 0.002}$	$50.88^{\pm 0.055}$	$0.068^{\pm 0.001}$
w GT Contact	$0.112^{\pm 0.001}$	$\mathbf{1.823}^{\pm 0.088}$	$0.821^{\pm 0.001}$	$\mathbf{54.60}^{\pm 0.049}$	$0.067^{\pm 0.001}$
Ours	$\mathbf{0.109}^{\pm 0.000}$	$1.845^{\pm 0.083}$	$\mathbf{0.826}^{\pm 0.000}$	$52.96^{\pm 0.080}$	$0.065^{\pm 0.000}$

Attention or Self Attention degraded the performance.The components of HOI-Mask Transformer can effectively modeling interaction and only connect them together can leverage the model's maximum potential.

5.3 Contact-aware Reconstruction Guidance

The Reconstruction Guidance can significantly improve the interaction quality through logits update process to optimize the HOI tokens. In Table 2, removing this component causes a severe drop in C_{prec} and $C_{\%}$, highlights this novel technique in generative masked model for improving the interaction quality.

$$\mathcal{L}_g = \lambda_c \mathcal{L}_c + \lambda_p \mathcal{L}_p, \tag{14}$$

where λ_c and λ_p denote the weights of the contact loss and penetration loss, respectively. We provide additional ablation results on the guidance loss in Table 4. Removing the contact loss reduces the contact percentage $C_{\%}$, while excluding the penetration loss increases the Penetration Score (PS), demonstrating that both terms are essential for Guidance to ensure accurate contact modeling and mitigate physically implausible interpenetration. We further compare our predicted contact guidance with Ground Truth (GT) contact. Although GT contact increases interaction probability, it leads to worse penetration score performance. The GT contact signal may biases the model towards spatial overlap to ensure contact occurrence. Directly applying GT contact as guidance without modeling its coupling with motion dynamic may cause the guidance to drift toward physically implausible configurations. In contrast, the predicted highly correlate contact achieves a better balance between motion fidelity and interaction quality.

6 Conclusion

In this paper, we propose HOIMask, the first generative masked modeling framework for HOI generation. HOIMask employs a HOI VQ-VAE to encode HOI motions into 2D discrete token space, preserving fine-grained spatial-temporal structures. Then, a specialized HOI-Mask Transformer is designed to model HOI tokens and capture complex interaction dependencies. Furthermore, we propose a contact-aware reconstruction guidance to optimize HOI tokens during inference, enabling contact signals to guide the generative masked modeling process for producing physically coherent interactions. Extensive experiments demonstrate that HOIMask generates more plausible and semantically consistent HOI motions compared to diffusion-based methods.

Acknowledgments

We would like to thank Professor Joshua Zhexue Huang and Professor Yulin He for their help and support in investigation, writing - review & editing, and funding acquisition. This work was supported by the Science and Technology Major Project of Shenzhen (KJZD20230923114809020), Basic Research Foundation of Shenzhen (JCYJ20250604175602004), and Research Task Assignment Project from Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ) (No. GML-26420011).

References

1. Bhatnagar, B.L., Xie, X., Petrov, I.A., Sminchisescu, C., Theobalt, C., Pons-Moll, G.: Behave: Dataset and method for tracking human object interactions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15935–15946 (2022)
2. Cha, J., Kim, J., Yoon, J.S., Baek, S.: Text2hoi: Text-guided 3d motion generation for hand-object interaction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1577–1585 (2024)
3. Chang, H., Zhang, H., Jiang, L., Liu, C., Freeman, W.T.: Maskgit: Masked generative image transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11315–11325 (2022)
4. Chen, X., Jiang, B., Liu, W., Huang, Z., Fu, B., Chen, T., Yu, G.: Executing your commands via motion diffusion in latent space. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18000–18010 (2023)
5. Christen, S., Hampali, S., Sener, F., Remelli, E., Hodan, T., Sauser, E., Ma, S., Tekin, B.: Diffh2o: Diffusion-based synthesis of hand-object interactions from textual descriptions. In: SIGGRAPH Asia 2024 Conference Papers. pp. 1–11 (2024)
6. Cong, P., Wang, Z., Ma, Y., Yue, X.: Semgeomo: Dynamic contextual human motion generation with semantic and geometric guidance. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17561–17570 (2025)

7. Dabral, R., Mughal, M.H., Golyanik, V., Theobalt, C.: Mofusion: A framework for denoising-diffusion-based motion synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9760–9770 (2023)
8. Dai, W., Chen, L.H., Wang, J., Liu, J., Dai, B., Tang, Y.: Motionlcm: Real-time controllable motion generation via latent consistency model. In: European Conference on Computer Vision. pp. 390–408 (2024)
9. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021)
10. Diller, C., Dai, A.: Cg-hoi: Contact-guided 3d human-object interaction generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19888–19901 (2024)
11. Guo, C., Mu, Y., Javed, M.G., Wang, S., Cheng, L.: Momask: Generative masked modeling of 3d human motions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1900–1910 (2024)
12. Guo, C., Zou, S., Zuo, X., Wang, S., Ji, W., Li, X., Cheng, L.: Generating diverse and natural 3d human motions from text. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5152–5161 (2022)
13. Guo, C., Zuo, X., Wang, S., Cheng, L.: Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts. In: European Conference on Computer Vision. pp. 580–597 (2022)
14. He, W., Liu, Y., Liu, R., Yi, L.: Syncdiff: Synchronized motion diffusion for multibody human-object interaction synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11731–11743 (2025)
15. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
16. Ho, J., Salimans, T.: Classifier-free diffusion guidance. In: NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications (2021)
17. Huang, G., Zeng, L.A., Zheng, Z., Gu, S., Zheng, W.S.: Efficient explicit joint-level interaction modeling with mamba for text-guided hoi generation. arXiv preprint arXiv:2503.23121 (2025)
18. Huang, M., Chu, F.J., Tekin, B., Liang, K.J., Ma, H., Wang, W., Chen, X., Gleize, P., Xue, H., Lyu, S., et al.: Hoigpt: Learning long-sequence hand-object interaction with language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7136–7146 (2025)
19. Huang, Y., Taheri, O., Black, M.J., Tzionas, D.: Intercap: Joint markerless 3d tracking of humans and objects in interaction. In: DAGM German Conference on Pattern Recognition. pp. 281–299. Springer (2022)
20. Jang, E., Gu, S., Poole, B.: Categorical reparameterization with gumbel-softmax. In: The fifth International Conference on Learning Representations (2017)
21. Javed, M.G., Li, X., et al.: Intermask: 3d human interaction generation via collaborative masked modeling. In: The Thirteenth International Conference on Learning Representations (2025)
22. Jeong, M., Hwang, Y., Lee, J., Jung, S., Kim, W.H.: Hgm^3: Hierarchical generative masked motion modeling with hard token mining. In: The Thirteenth International Conference on Learning Representations (2025)
23. Ji, Y., Liu, Y., Zhuo, Y., Yu, W., Ma, F., Huang, J.Z., Yu, F.: Onlinehoi: Towards online human-object interaction generation and perception. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 9395–9403 (2025)
24. Jiang, B., Chen, X., Liu, W., Yu, J., Yu, G., Chen, T.: Motiongpt: Human motion as a foreign language. Advances in Neural Information Processing Systems **36**, 20067–20079 (2023)

25. Jiang, H., Liu, S., Wang, J., Wang, X.: Hand-object contact consistency reasoning for human grasps generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11107–11116 (2021)
26. Jin, P., Li, H., Cheng, Z., Li, K., Yu, R., Liu, C., Ji, X., Yuan, L., Chen, J.: Local action-guided motion diffusion model for text-to-motion generation. In: European Conference on Computer Vision. pp. 392–409 (2024)
27. Li, J., Clegg, A., Mottaghi, R., Wu, J., Puig, X., Liu, C.K.: Controllable human-object interaction synthesis. In: European Conference on Computer Vision. pp. 54–72 (2024)
28. Li, J., Wu, J., Liu, C.K.: Object motion guided human motion synthesis. ACM Transactions on Graphics (TOG) **42**(6), 1–11 (2023)
29. Liu, X., Li, Y.L., Zeng, A., Zhou, Z., You, Y., Lu, C.: Bridging the gap between human motion and action semantics via kinematic phrases. In: European Conference on Computer Vision. pp. 223–240 (2024)
30. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: Smpl: a skinned multi-person linear model. ACM Transactions on Graphics (TOG) **34**(6), 1–16 (2015)
31. Lu, J., Huang, C.H.P., Bhattacharya, U., Huang, Q., Zhou, Y.: Humoto: A 4d dataset of mocap human object interactions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10886–10897 (2025)
32. Lu, S., Chen, L.H., Zeng, A., Lin, J., Zhang, R., Zhang, L., Shum, H.Y.: Humantomato: text-aligned whole-body motion generation. In: Proceedings of the 41st International Conference on Machine Learning. pp. 32939–32977 (2024)
33. Lucas, T., Baradel, F., Weinzaepfel, P., Rogez, G.: Posegpt: Quantization-based 3d human motion generation and forecasting. In: European Conference on Computer Vision. pp. 417–435 (2022)
34. Meng, Z., Xie, Y., Peng, X., Han, Z., Jiang, H.: Rethinking diffusion for text-driven human motion generation: Redundant representations, evaluation, and masked autoregression. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 27859–27871 (2025)
35. Parmar, N., Vaswani, A., Uszkoreit, J., Kaiser, L., Shazeer, N., Ku, A., Tran, D.: Image transformer. In: International conference on machine learning. pp. 4055–4064. PMLR (2018)
36. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4195–4205 (2023)
37. Peng, X., Xie, Y., Wu, Z., Jampani, V., Sun, D., Jiang, H.: Hoi-diff: Text-driven synthesis of 3d human-object interactions using diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 2878–2888 (2025)
38. Petrov, I.A., Marin, R., Chibane, J., Pons-Moll, G.: Tridi: Trilateral diffusion of 3d humans, objects, and interactions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5523–5535 (2025)
39. Pinyoanuntapong, E., Saleem, M., Karunratanakul, K., Wang, P., Xue, H., Chen, C., Guo, C., Cao, J., Ren, J., Tulyakov, S.: Maskcontrol: Spatio-temporal control for masked motion synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9955–9965 (2025)
40. Pinyoanuntapong, E., Saleem, M.U., Wang, P., Lee, M., Das, S., Chen, C.: Bamm: Bidirectional autoregressive motion model. In: European Conference on Computer Vision. pp. 172–190 (2024)

41. Pinyoanuntapong, E., Wang, P., Lee, M., Chen, C.: Mmm: Generative masked motion model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1546–1555 (2024)
42. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: Proceedings of the 38st International conference on machine learning. pp. 8748–8763. PmLR (2021)
43. Ren, Z., Huang, S., Li, X.: Realistic human motion generation with cross-diffusion models. In: European Conference on Computer Vision. pp. 345–362 (2024)
44. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10684–10695 (2022)
45. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: The Ninth International Conference on Learning Representations (2021)
46. Song, W., Zhang, X., Li, S., Gao, Y., Hao, A., Hou, X., Chen, C., Li, N., Qin, H.: Hoianimator: Generating text-prompt human-object animations using novel perceptive diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 811–820 (2024)
47. Taheri, O., Ghorbani, N., Black, M.J., Tzionas, D.: Grab: A dataset of whole-body human grasping of objects. In: European conference on computer vision. pp. 581–600. Springer (2020)
48. Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-or, D., Bermano, A.H.: Human motion diffusion model. In: The Eleventh International Conference on Learning Representations (2023)
49. Tian, K., Jiang, Y., Yuan, Z., Peng, B., Wang, L.: Visual autoregressive modeling: Scalable image generation via next-scale prediction. Advances in neural information processing systems **37**, 84839–84865 (2024)
50. Van Den Oord, A., Vinyals, O., et al.: Neural discrete representation learning. Advances in neural information processing systems **30** (2017)
51. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
52. Wang, Y., Leng, Z., Li, F.W., Wu, S.C., Liang, X.: Fg-t2m: Fine-grained text-driven human motion generation via diffusion model. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22035–22044 (2023)
53. Wang, Z., Liu, J.C.: Translating math formula images to latex sequences using deep neural networks with sequence-level training. International Journal on Document Analysis and Recognition (IJDAR) **24**(1), 63–75 (2021)
54. Wang, Z., Xu, S., Guo, C., Zhou, B., Gong, J., Wang, J., Wang, Y.X., Gui, L.Y.: Unleashing guidance without classifiers for human-object interaction animation. In: The Fourteenth International Conference on Learning Representations (2026)
55. Wu, L., Chen, Z., Lan, J.: Hoi-dyn: Learning interaction dynamics for human-object motion diffusion. arXiv preprint arXiv:2507.01737 (2025)
56. Wu, Z., Li, J., Xu, P., Liu, C.K.: Human-object interaction from human-level instructions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11176–11186 (2025)
57. Xu, S., Li, Z., Wang, Y.X., Gui, L.Y.: Interdiff: Generating 3d human-object interactions with physics-informed diffusion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14928–14940 (2023)

58. Xue, M., Liu, Y., Guo, L., Huang, S., Ding, C.: Guiding human-object interactions with rich geometry and relations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22714–22723 (2025)
59. Yuksel, C.: Sample elimination for generating poisson disk sample sets. In: Computer Graphics Forum. vol. 34, pp. 25–32. Wiley Online Library (2015)
60. Zeng, L.A., Huang, G., Wei, Y.L., Gu, S., Tang, Y.M., Meng, J., Zheng, W.S.: Chainhoi: Joint-based kinematic chain modeling for human-object interaction generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12358–12369 (2025)
61. Zeng, L.A., Huang, G., Wu, G., Zheng, W.S.: Light-t2m: A lightweight and fast model for text-to-motion generation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 9797–9805 (2025)
62. Zhang, J., Zhang, Y., Cun, X., Zhang, Y., Zhao, H., Lu, H., Shen, X., Shan, Y.: Generating human motion from textual descriptions with discrete representations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14730–14740 (2023)
63. Zhang, M., Cai, Z., Pan, L., Hong, F., Guo, X., Yang, L., Liu, Z.: Motiondiffuse: Text-driven human motion generation with diffusion model. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(6), 4115–4128 (2024)
64. Zhang, M., Guo, X., Pan, L., Cai, Z., Hong, F., Li, H., Yang, L., Liu, Z.: Remodiffuse: Retrieval-augmented motion diffusion model. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 364–373 (2023)
65. Zhao, C., Zhang, J., Du, J., Shan, Z., Wang, J., Yu, J., Wang, J., Xu, L.: I'm hoi: Inertia-aware monocular capture of 3d human-object interactions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 729–741 (2024)
66. Zhong, C., Hu, L., Zhang, Z., Xia, S.: Attt2m: Text-driven human motion generation with multi-perspective attention mechanism. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 509–519 (2023)
67. Zhou, Z., Wan, Y., Wang, B.: Avatargpt: All-in-one framework for motion understanding planning generation and beyond. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1357–1366 (2024)
68. Zou, Q., Yuan, S., Du, S., Wang, Y., Liu, C., Xu, Y., Chen, J., Ji, X.: Parco: Part-coordinating text-to-motion synthesis. In: European Conference on Computer Vision. pp. 126–143 (2024)