Long-Term Composition of Human-Object Interactions

Dayoung Gong[1] **Minsu Cho**[1,2] **Gül Varol**[3]

[1]Pohang University of Science and Technology (POSTECH) [2]RLWRLD
[3]LIGM, École des Ponts, IP Paris, Univ Gustave Eiffel, CNRS
https://gongda0e.github.io/LTHOI/

Abstract

Real-world human activities unfold as sequences of temporally dependent human-object interactions (HOIs), yet existing methods address either atomic HOIs in isolation or text-driven motion composition without object or scene context. To bridge this gap, we introduce the task of long-term action composition of HOIs, where a human navigates between and interacts with multiple objects within a 3D scene. The primary challenges lie in producing smooth transitions, scene-aware locomotion, and plausible hand-object interaction. To this end, we propose LT-HOI, a diffusion-based model that jointly conditions on past and future contexts—the past providing kinematic continuity across transitions, the future enabling anticipatory navigation toward upcoming interacting objects. We further integrate Diffusion Noise Optimization for collision avoidance and introduce a hand-object displacement loss to improve contact quality. On the ParaHome dataset, we establish the first benchmark for this task and demonstrate that LT-HOI composes long, temporally coherent HOIs.

1 Introduction

Human motion is inherently contextual, emerging from continuous interactions with the surrounding world. Real-world activities unfold not as isolated actions, but as sequences of temporally dependent human-object interactions (HOIs). Each interaction builds upon the previous one—sharing object states, spatial configurations, and intentions—while transitioning smoothly to the next. Capturing such temporal dependencies is essential for generating long-term and goal-oriented motion.

Despite recent progress in temporal action composition [1, 3, 46, 49] and HOI generation [7, 16, 19–21, 28, 29, 41, 43, 47], a significant gap remains. Temporal composition methods generate temporally aligned human-only motions from text without modeling objects or scenes, while HOI models focus on single short interactions. This is partly due to the scarcity of long HOI sequences in existing datasets [4, 12, 14, 19, 37], leaving long-term HOI composition largely unexplored. The recently released ParaHome dataset [18] provides long-duration HOI sequences in 3D home environments, opening up the possibility of studying this problem.

In this paper, we introduce the task of long-term action composition of HOIs: synthesizing motion sequences where a 3D human smoothly navigates between and interacts with multiple objects within a 3D scene (Fig. 1). We assume object motions are given as input and use them as spatio-temporal guidance for human motion synthesis. In practice, object motions can be obtained either from planning [2, 45] or from object pose estimation methods that recover object trajectories from human demonstrations or videos [39, 40]. The primary challenges then lie in three aspects: (1) smooth transitions between consecutive interactions, (2) scene-aware locomotion toward upcoming interactions, and (3) plausible hand-object contact.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

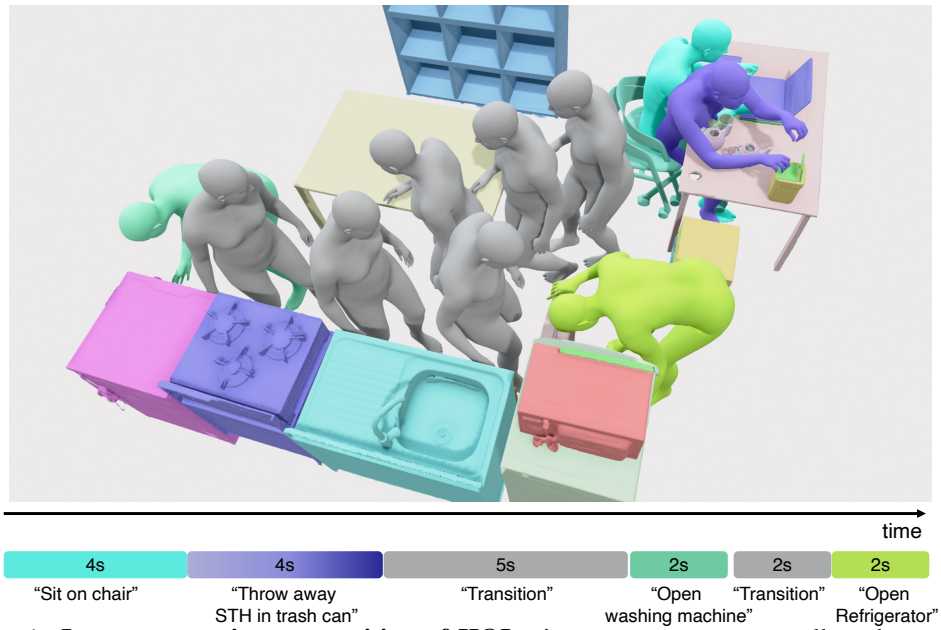


Figure 1: **Long-term action composition of HOIs** aims to generate temporally coherent HOI sequences within a 3D scene. The figure illustrates motions generated by our model. Different body colors denote different actions.

To address these challenges, we propose LT-HOI, a diffusion-based model that composes long-term HOI sequences. At its core, LT-HOI jointly conditions on past and future contexts that play complementary roles: the past comprises previously generated human motion along with text, object motion, and object geometry, providing kinematic continuity for smooth transitions; the future comprises the upcoming text, object motion, and object geometry, providing spatial awareness for goal-directed locomotion. In practice, we adopt an autoregressive-across-subsequences, non-autoregressive-within-subsequence generation scheme [1]. At each step, the model produces a motion segment covering the tail of the previous action, the current action, and the beginning of the next—naturally aligning with our past, current, and future context design. While future context guides the model to the right destination, navigating around scene obstacles requires explicit geometric reasoning. To this end, we further leverage Diffusion Noise Optimization (DNO) [17] for collision-aware locomotion and introduce a hand-object displacement loss for plausible interaction.

Our contributions are as follows: **1)** We introduce the task of long-term action composition of HOIs and establish the first benchmark on the ParaHome dataset. **2)** We propose LT-HOI, a diffusion-based model that jointly conditions on past and future contexts to compose long HOI sequences, where past context ensures smooth transitions and future context enables goal-directed locomotion. **3)** We further integrate DNO to enable collision-aware navigation and introduce a hand-object displacement loss to improve contact quality during interactions. **4)** Comprehensive experiments validate the effectiveness of our approach, with ablations confirming the contribution of each component.

2 Related Work

Temporal action composition for human motions. Given a series of text descriptions and their corresponding durations, temporal action composition aims to generate human motions that follow the instructions and exhibit coherent temporal structure. TEACH [1] first introduces this task through a hierarchical formulation that generates motions non-autoregressively within segments and autoregressively across them, building upon a text-conditioned 3D motion synthesis model [30]. Subsequent works [3, 34, 46] improve transition quality by refining overlapping segments—either via post-hoc inpainting [34], joint diffusion over temporal windows [46], or blended positional encodings [3]. More recent works further explore streaming motion generation [22, 42]. However, these approaches focus on text-driven human-only motion composition and do not model interactions with objects or the surrounding environment. In this paper, we address temporal composition of HOIs within a 3D scene.

Human-object interactions. Modeling 3D human interactions with objects [5, 7, 19, 20, 28, 36, 43] and scenes [11, 23, 44, 48] is a crucial task for embodied AI. In this paper, we mainly focus on interacting with dynamic objects within a 3D scene. The goal of human-object interaction motion generation is to synthesize human motions that interact with objects. Early methods such as OMOMO [19] take a sequence of object states as input and generate full-body human motions that interact with the given object. InterDiff [43] introduces the task of anticipating future 3D human-object interactions, given past human and object motions. To achieve this, it employs an interaction diffusion model together with an interaction correction module. HOI-Diff [27] proposes an affordance-guided diffusion model for HOI generation, while CG-HOI [7] and TriDi [28] model human, object, and contact interactions within a unified network. Recently, CHOIS [20] presented a diffusion-based framework that predicts both human and object motions conditioned on a text prompt, the initial human pose, the target object pose, object waypoints, and object geometries. DecHOI [16] decouples trajectory generation from interaction modeling for HOI synthesis, and recent works further extend HOI generation to bimanual interactions with articulated objects [47], human-level instructions [41], zero-shot generation with part affordance guidance [21], egocentric modeling [29], and large-scale mocap data [24]. However, existing approaches primarily focus on single, short interactions, rather than multi-step, temporally coherent sequences involving multiple objects. In this paper, we address long-term action composition of HOIs, aiming to generate temporally coherent human motions that sequentially interact with multiple objects within a 3D scene.

3 Approach

In this section, we present LT-HOI, a diffusion-based model for long-term action composition of HOIs. We formulate the generation as autoregressive across subsequences and non-autoregressive within each subsequence. At each step, the subsequence is generated conditioned on both past and future contexts—past context provides kinematic continuity for smooth transitions, and future context guides goal-directed locomotion toward the next interaction. We first describe the problem formulation (Sec. 3.1), data representation (Sec. 3.2), and the model architecture (Sec. 3.3), followed by training (Sec. 3.4) and inference (Sec. 3.5) strategies.

3.1 Problem Formulation

Given the text descriptions, object geometries, and object motions of the interactions, together with a global scene map, the goal is to synthesize human motions of N consecutive interactions within a shared 3D scene. Specifically, the full motion sequence $\boldsymbol{X}$ consists of N subsequences $\boldsymbol{X} = [\boldsymbol{X}_1, \cdots, \boldsymbol{X}_N]$, where each $\boldsymbol{X}_n = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_L] \in \mathbb{R}^{L \times D}$ represents a motion segment, $\boldsymbol{x}_l \in \mathbb{R}^D$ is the human pose at frame l, and D is the pose dimension. Each subsequence $\boldsymbol{X}_n$ is generated by conditioning on three contexts, each consisting of text, object geometry, and object motion: the current context from its own interaction, and the past and future contexts from the previous and next interactions. The past context also includes the previously generated human motion.

3.2 Data Representation

Human. We represent a human pose as $\boldsymbol{x} = [\boldsymbol{J}, \boldsymbol{r}, \boldsymbol{\theta}]$, where $\boldsymbol{J} \in \mathbb{R}^{N^{\text{joint}} \times 3}$ denotes the global joint positions, $\boldsymbol{r} \in \mathbb{R}^6$ is the 6D rotation [50] of the global body orientation, and $\boldsymbol{\theta} \in \mathbb{R}^{(N^{\text{joint}}-1) \times 6}$ represents the 6D rotations of the SMPL-X [26] pose parameters. N^{joint} includes both body and hand joints, and all components are flattened and concatenated, yielding $D = 3N^{\text{joint}} + 6 + 6(N^{\text{joint}} - 1)$.
Object. The object pose $\boldsymbol{o} \in \mathbb{R}^9$ consists of its translation and 6D rotation. The object geometry $\boldsymbol{g} \in \mathbb{R}^K$ is encoded using Basis Point Sets (BPS) [31], following [19, 20], where K is the number of basis points. BPS provides a compact shape descriptor by measuring distances from a fixed set of basis points to the object surface.
Scene. We represent the scene through three complementary maps projected onto the ground plane. First, a signed distance field $\boldsymbol{s}^{\text{sdf}} \in \mathbb{R}^{H \times W \times 1}$ encodes the walkable space around static obstacles, with values normalized from 0 (occupied) to 1 (free). Second, the initial positions of the current and future interacting objects are encoded as Gaussian heatmaps, $\boldsymbol{s}^{\text{curr}}, \boldsymbol{s}^{\text{future}} \in \mathbb{R}^{H \times W \times 1}$, respectively. The final scene representation $\boldsymbol{s} = [\boldsymbol{s}^{\text{sdf}}, \boldsymbol{s}^{\text{curr}}, \boldsymbol{s}^{\text{future}}] \in \mathbb{R}^{H \times W \times 3}$ captures both navigable geometry and the spatial layout of interactions.

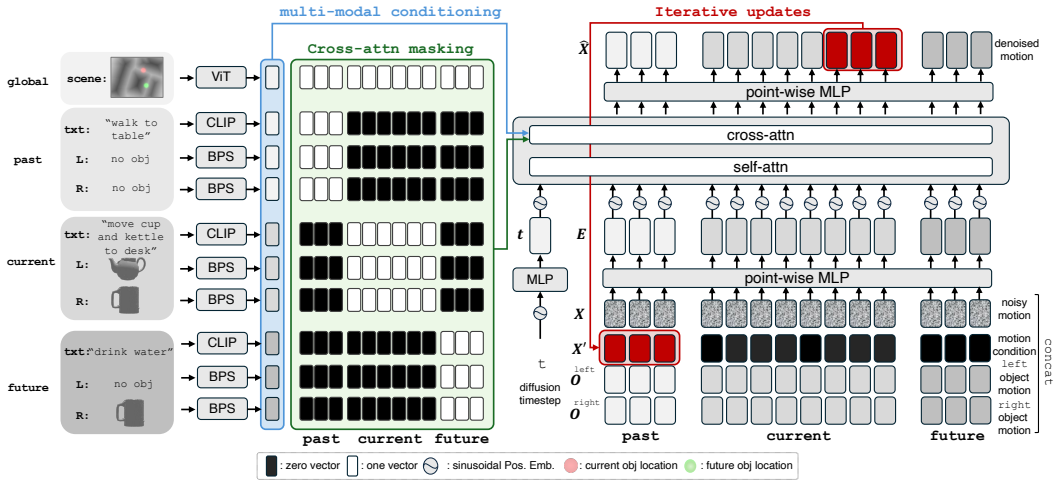


Figure 2: **Overall framework of LT-HOI.** LT-HOI is a diffusion-based generative model for long-term HOI composition. The model generates HOIs autoregressively across subsequences and non-autoregressively within each subsequence, conditioning on past and future context while iteratively feeding previous predictions into subsequent generation.

3.3 LT-HOI

The overall pipeline is illustrated in Fig. 2. At each autoregressive step, the model generates a single subsequence $\boldsymbol{X}_n$ of total length $L^{\texttt{all}}$, composed of three temporally distinct segments: the tail of the previous motion $\boldsymbol{X}_n^{\texttt{past}}$ (length $L^{\texttt{past}}$), the current target motion $\boldsymbol{X}_n^{\texttt{curr}}$ (length $L^{\texttt{curr}}$), and the beginning of the next motion $\boldsymbol{X}_n^{\texttt{future}}$ (length $L^{\texttt{future}}$). The full subsequence is $\boldsymbol{X}_n = [\boldsymbol{X}_n^{\texttt{past}}, \boldsymbol{X}_n^{\texttt{curr}}, \boldsymbol{X}_n^{\texttt{future}}] \in \mathbb{R}^{L^{\texttt{all}} \times D}$, where $L^{\texttt{all}} = L^{\texttt{past}} + L^{\texttt{curr}} + L^{\texttt{future}}$.

Input embedding. Since each temporal segment involves a distinct action with potentially different objects, we encode the text description and object geometry per segment $p \in \{\texttt{past}, \texttt{curr}, \texttt{future}\}$. Text descriptions are encoded with CLIP [33] and projected through MLPs to obtain text embeddings $\boldsymbol{w}^p \in \mathbb{R}^C$. BPS representations of the left and right object geometries, $\boldsymbol{g}^{\texttt{left},p}, \boldsymbol{g}^{\texttt{right},p} \in \mathbb{R}^K$, are similarly projected to $\tilde{\boldsymbol{g}}^{\texttt{left},p}, \tilde{\boldsymbol{g}}^{\texttt{right},p} \in \mathbb{R}^C$, with learnable `[left]` and `[right]` indicator tokens added via summation. The scene map $\boldsymbol{s}$ is patchified and processed by a Vision Transformer (ViT) [8], with the resulting `[CLS]` token projected to obtain the scene feature $\tilde{\boldsymbol{s}} \in \mathbb{R}^C$. To facilitate smooth transitions, we construct a motion condition $\boldsymbol{X}'_n \in \mathbb{R}^{L^{\texttt{all}} \times D}$, where the past segment is filled with the clean previous motion $\boldsymbol{X}_n^{\texttt{past}}$ and the remaining positions are zero-padded. The diffusion timestep t is encoded via a sinusoidal positional embedding followed by an MLP, yielding $\boldsymbol{t} \in \mathbb{R}^C$.

Multimodal conditioning. We integrate conditions based on their temporal granularity. Frame-wise conditions—the per-frame object motions $\boldsymbol{O}^{\{\texttt{left},\texttt{right}\}}$ and the past motion condition $\boldsymbol{X}'_n$—are concatenated channel-wise with the noisy motion $\boldsymbol{X}_n$ and projected through an MLP to produce the motion embedding:

$$\boldsymbol{E} = \texttt{MLP}([\boldsymbol{X}_n, \boldsymbol{X}'_n, \boldsymbol{O}^{\texttt{left}}, \boldsymbol{O}^{\texttt{right}}]). \tag{1}$$

This channel-wise integration provides dense spatial correspondence between the human and objects at every frame. In contrast, subsequence-wise conditions—text descriptions $\boldsymbol{w}^p$ and object geometries $\tilde{\boldsymbol{g}}^p$ for each segment p—and the global scene feature $\tilde{\boldsymbol{s}}$ are incorporated via cross-attention. These are assembled into condition tokens $\boldsymbol{M}$.

Denoising transformer. The motion embedding $\boldsymbol{E}$, together with the timestep embedding $\boldsymbol{t}$, is fed into a denoising transformer f consisting of $N^{\texttt{dec}}$ decoder layers. Each layer applies self-attention over the motion sequence followed by cross-attention to $\boldsymbol{M}$. During cross-attention, a subsequence-aware mask restricts each motion frame to attend only to condition tokens of its corresponding segment, while the scene token is globally accessible across all segments. The transformer output $\boldsymbol{F} \in \mathbb{R}^{L^{\texttt{all}} \times C}$ is:

$$\boldsymbol{F} = f(\boldsymbol{E}, \boldsymbol{t};\, \boldsymbol{M}). \tag{2}$$

The final predicted motion $\hat{\boldsymbol{X}}_n \in \mathbb{R}^{L^{\text{all}} \times D}$ is obtained by a point-wise MLP:

$$\hat{\boldsymbol{X}}_n = \texttt{MLP}(\boldsymbol{F}). \tag{3}$$

For brevity, we denote the entire denoising network—comprising input embedding (Eq. 1), denoising transformer (Eq. 2), and output projection (Eq. 3)—as f_θ, so that $\hat{\boldsymbol{X}}_n = f_\theta(\boldsymbol{X}_n, t; \boldsymbol{M})$.

3.4 Training

The model is trained via the standard diffusion denoising objective. During the forward process, noise is added to the clean motion $\boldsymbol{X}^0$ at a randomly sampled timestep t:

$$\boldsymbol{X}^t = \sqrt{\gamma(t)}\boldsymbol{X}^0 + \sqrt{1-\gamma(t)}\epsilon, \quad \epsilon \sim \mathcal{N}(0, \boldsymbol{I}), \tag{4}$$

where $\gamma(t)$ is a pre-defined variance schedule. The denoising network f_θ is trained to recover $\boldsymbol{X}^0$ from $\boldsymbol{X}^t$ using an ℓ_1 reconstruction loss:

$$\mathcal{L}^{\text{recon}} = \|f_\theta(\boldsymbol{X}^t, t; \boldsymbol{M}) - \boldsymbol{X}^0\|_1. \tag{5}$$

Hand-object displacement loss. To improve the quality of hand-object interactions, we introduce an auxiliary loss that supervises the spatial relationship between hand joints and object surfaces. For each interacting object, we first obtain a set of P sparse representative points $\boldsymbol{Q} \in \mathbb{R}^{P\times 3}$ by applying Farthest Point Sampling (FPS) [32] to the object mesh vertices, which are then transformed per frame using the ground-truth object pose. Given N^{hand} hand joint positions $\boldsymbol{J}^{\text{hand}} \in \mathbb{R}^{N^{\text{hand}}\times 3}$, we define the BPS feature function $\phi(\boldsymbol{J}^{\text{hand}}, \boldsymbol{Q}) \in \mathbb{R}^{N^{\text{hand}}\times 3}$ as the displacement vector from each hand joint to its nearest point in $\boldsymbol{Q}$. The displacement loss is then defined as the ℓ_1 distance between the BPS features computed from the predicted hand joints $\hat{\boldsymbol{J}}^{\text{hand}}$ and those from the ground-truth hand joints $\boldsymbol{J}^{\text{hand}}$:

$$\mathcal{L}^{\text{disp}} = \|\phi(\hat{\boldsymbol{J}}^{\text{hand}}, \boldsymbol{Q}) - \phi(\boldsymbol{J}^{\text{hand}}, \boldsymbol{Q})\|_1. \tag{6}$$

This loss is computed independently for the left and right hands with their respective interacting objects, and is applied only at frames where an object interaction is present. By supervising the relative spatial relationship between hands and objects rather than absolute hand positions, the loss encourages more plausible hand-object contacts.

The final training objective is $\mathcal{L} = \lambda^{\text{recon}}\mathcal{L}^{\text{recon}} + \lambda^{\text{disp}}\mathcal{L}^{\text{disp}}$, where λ^{recon} and λ^{disp} control the weights of the respective losses.

3.5 Inference

The model generates the full long-term sequence by autoregressively synthesizing N subsequences during inference. The process begins with no past context ($L^{\text{past}} = 0$) for the first subsequence. After generating $\boldsymbol{X}_n$, the last L^{past} frames of $\boldsymbol{X}_n^{\text{curr}}$ serve as the past context for the next step, forming $\boldsymbol{X}_{n+1}^{\text{past}}$. The canonical frame for the first subsequence is set using the ground-truth yaw orientation; for subsequent steps, it is derived from the predicted motion.

Within each subsequence, motion is synthesized using the reverse diffusion process. Starting from pure Gaussian noise $\hat{\boldsymbol{X}}^S$, the denoising network f_θ iteratively refines the sequence, $\hat{\boldsymbol{X}}^S \to \hat{\boldsymbol{X}}^{S-\Delta} \to \cdots \to \hat{\boldsymbol{X}}^0$, following the deterministic update rule of DDIM [35], where Δ denotes the step size that skips intermediate diffusion steps to accelerate sampling.

Obstacle-avoidance navigation using DNO. To enable obstacle-avoidance navigation without compromising the learned motion prior, we adopt Diffusion Noise Optimization (DNO) [17], a test-time optimization framework that steers a pre-trained diffusion model toward a target criterion by updating its initial latent noise rather than the model parameters or the motion itself. We instantiate this criterion as a scene-collision loss defined on the pelvis trajectory following DNO [17]. Specifically, at each optimization iteration, we first run a full reverse-diffusion pass starting from the initial noise $\hat{\boldsymbol{X}}^S$ to obtain the predicted clean motion $\hat{\boldsymbol{X}}^0$. We then extract the pelvis trajectory from $\hat{\boldsymbol{X}}^0$, project it onto the 2D ground plane, and query the scene SDF to compute the collision loss:

$$\mathcal{L}^{\text{col}} = \frac{1}{L^{\text{curr}} + L^{\text{future}}} \sum_{l=L^{\text{past}}+1}^{L^{\text{all}}} c_l \; + \; \max_l c_l, \tag{7}$$

where $c_l = \max\Big(0,\ \delta - \boldsymbol{s}^{\mathtt{sdf}}(\hat{\boldsymbol{J}}_l^{\mathtt{pel}})\Big)$ is the per-frame collision penalty, $\hat{\boldsymbol{J}}_l^{\mathtt{pel}}$ is the projected pelvis position at frame l, and δ is a safety margin. The first term penalizes the average collision over the current and future segments, and the second penalizes the worst-case collision. The gradient of $\mathcal{L}^{\mathtt{col}}$ is then backpropagated through the entire reverse-diffusion process to update $\hat{\boldsymbol{X}}^S$. Since the update modifies only the latent noise while keeping the diffusion model fixed, the resulting locomotion stays close to the learned motion prior while reducing scene collisions.

4 Experiments

We first describe the dataset (Sec. 4.1), evaluation metrics (Sec. 4.2), and implementation details (Sec. 4.3). We then compare LT-HOI with state-of-the-art methods (Sec. 4.4), present analyses on the generation strategy, individual components, and the role of past and future contexts (Sec. 4.5), and provide qualitative results (Sec. 4.6).

4.1 ParaHome Dataset

We use the ParaHome dataset [18] to evaluate long-term composition of HOI motion generation. The dataset contains 207 scenes collected from 38 participants, where each participant interacts with 22 objects, resulting in a total of 486 minutes of human motions performing daily activities in real home environments. Each scene includes, on average, 26 sequences of HOI actions and 39 sequences of transition actions between HOI interactions. Transition actions refer to the in-between motions connecting consecutive HOI actions—including locomotion where the human moves across the 3D scene to reach the next interaction—which are crucial for evaluating temporal smoothness. We additionally annotate these motions as 'Transition'.

Splits. Since our primary goal is to evaluate the temporal composition of motions, and each scene in the dataset contains consecutive motion sequences, we split the dataset by scene to prevent data leakage across splits. The training, validation, and test sets consist of 165, 21, and 21 scenes, respectively, maintaining an approximate 8:1:1 ratio. In total, these splits contain 5,970, 1,005, and 989 subsequences, respectively. All sequences are processed at 30 FPS. Training subsequences average 113.5 frames (3.8 s), and each test scene is generated as a single continuous sequence by composing approximately 47 consecutive subsequences on average (4,556 frames, about 2.5 minutes of continuous HOIs). Notably, 13.7% of all subsequences are locomotion without any object interaction, where the model must navigate toward the next interaction using only the text, scene, and future context.

Interacting objects. Since the dataset does not provide annotations for interacting objects, we automatically detect them from the data based on hand-object proximity. Specifically, we obtain human vertices using the SMPL-X model [26] and compute distances between hand vertices and object vertices. If the minimum distance falls below a predefined threshold of 0.02 m, the object is regarded as interacting. Detection is performed independently for the left and right hands. Since ParaHome represents articulated objects with multiple parts—for example, a gas stove is composed of a main body and two knobs, a laptop consists of top and bottom parts, and a washing machine consists of a main body and a door—we adopt part-level object representations, where each part has its own translation and rotation. Non-articulated objects are represented as a single part. Each object part is encoded with object geometry $\boldsymbol{g}$ and object motions $\boldsymbol{O}$.

4.2 Evaluation Metrics

We evaluate the performance of our model from three perspectives: subsequence, transition, and sequence, following previous work [3, 13, 20].

Subsequence. For subsequence evaluation, we assess human motion, goal reaching, and contact quality. Following prior single-HOI generation work [20], we use MPJPE, foot height, foot sliding (FS), R@3, and FID for human motion, and contact precision, recall, F1, and hand penetration for contact quality. We further use translation error (T_{goal}) and orientation error (O_{goal}) to evaluate how accurately the human reaches the target object at the moment of interaction.

Transition. To evaluate the motion smoothness when transitioning between consecutive HOI actions, we define and evaluate the transition motions. Specifically, the transition motion $\boldsymbol{X}^{\mathtt{tr}} \in \mathbb{R}^{L^{\mathtt{tr}} \times D}$ is defined as the motion segment of length $L^{\mathtt{tr}}$ that bridges the termination of subsequence $\boldsymbol{X}_n$ and the start of $\boldsymbol{X}_{n+1}$. As evaluation metrics for transition motions, we adopt FID, PeakJerk (PJ), and Area

Table 1: **Comparison with the State of the Art**

Method	Subsequence												Transition			Sequence	
	Motion							Contact					Motion	Jerk		Scene	
	MPJPE↓	H_{feet} →	FS↓	R@3 ↑	FID↓	T_{goal} ↓	O_{goal} ↓	C_{prec} ↑	C_{rec} ↑	C_{F1} ↑	$C_{\%}$ →	P_{hand} ↓	FID↓	PJ↓	AUJ↓	$Coll_{\%}$ ↓	P_{max} ↓
GT	-	-3.30	0.29	-	-	-	-	-	-	-	0.49	1.7e-6	-	0.03	2.22	0.35	0.04
MDM	28.76	-6.39	0.33	0.43	0.70	54.14	1.04	0.34	0.18	0.21	0.18	6.2e-5	1.59	0.16	**2.05**	0.67	0.07
CHOIS	28.47	-6.90	0.34	0.44	1.22	54.81	1.07	0.33	0.18	0.20	0.17	3.4e-5	1.21	0.17	2.09	0.60	0.06
DecHOI	27.26	**-3.43**	0.39	0.44	0.53	50.33	0.98	0.40	0.23	0.26	0.24	8.7e-5	2.44	0.21	2.23	0.58	0.06
InterDiff	23.17	-5.17	**0.28**	0.36	1.09	39.85	0.65	0.38	0.20	0.23	0.19	**2.5e-6**	4.60	0.80	3.44	0.51	0.06
FlowMDM	24.46	-4.44	1.03	0.46	0.64	53.25	0.71	0.41	0.24	0.27	0.24	3.2e-6	6.16	**0.12**	2.12	0.53	0.06
LT-HOI (ours)	**19.16**	-4.28	0.33	**0.54**	**0.45**	**24.41**	**0.48**	**0.49**	**0.36**	**0.38**	**0.35**	8.2e-5	**1.12**	0.13	2.16	**0.32**	**0.05**

Under the Jerk (AUJ) [3].

Sequence. To evaluate scene-aware locomotion, we further report the scene collision ratio ($Coll_{\%}$) and the maximum scene penetration depth (P_{max}) following [13], which respectively measure the fraction of frames in which the body interpenetrates the 3D scene and the worst-case penetration depth into scene geometry. Detailed explanations of metrics are provided in Sec. A.

4.3 Implementation Details

We use $N^{\texttt{joint}} = 52$ joints in total, including 22 body joints and 30 hand joints. Global joint positions are extracted using the SMPL-X joint regressor [26]. All human, object, and scene data are canonicalized with respect to the yaw orientation of the human pose at the first frame of the current subsequence. Joint positions and object translations are then normalized to the range $[-1, 1]$. The denoising transformer consists of $N^{\texttt{dec}} = 8$ decoder layers with 4 attention heads and a channel dimension of $C = 256$. We train the model with a batch size of 32 and an initial learning rate of 2.5×10^{-4}, using a linear warm-up over the first 2,000 iterations followed by cosine decay. The total number of training iterations is 200,000. During inference, we use 75 DDIM steps and 150 DNO optimization steps. We set the boundary margin δ to 0.15 m. The transition length $L^{\texttt{tr}}$ is set to 30 frames. We set $\lambda^{\texttt{recon}} = 1$ and $\lambda^{\texttt{disp}} = 0.3$.

4.4 Comparison with the State of the Art

We compare LT-HOI with MDM [38], CHOIS [20], DecHOI [16], InterDiff [43], and FlowMDM [3], which target text-conditioned motion generation, HOI synthesis (CHOIS and DecHOI), HOI prediction, and temporal action composition, respectively. For a fair comparison, all baselines are retrained from scratch on ParaHome with the same input modalities and training recipe as LT-HOI: MDM and FlowMDM additionally take object motion, object geometry, and scene features; CHOIS and DecHOI take scene features; and InterDiff takes text and scene features. Other design choices, including the network architecture, follow the original works: MDM, CHOIS, and DecHOI generate each subsequence conditioned on the last frame of the previous one, following their original setting that takes the initial human pose; InterDiff predicts a subsequence from past observations; and FlowMDM jointly generates the full sequence. Implementation details are provided in Sec. B.

Table 1 presents the results. LT-HOI outperforms the baselines on most metrics, and the gap reflects how each baseline handles temporal context. Observing only a single past frame, MDM, CHOIS, and DecHOI cannot anticipate where the next interaction occurs, resulting in large goal-reaching errors (over 50 cm vs. 24.41 cm for LT-HOI). DecHOI improves over MDM and CHOIS in MPJPE and contact quality, but its MPJPE remains higher than those of InterDiff, FlowMDM, and LT-HOI, suggesting that strong atomic HOI synthesis alone does not address long-term composition. InterDiff conditions on past motion but, without future context, has higher goal-reaching error and lower contact quality than LT-HOI (T_{goal} 39.85 vs. 24.41, C_{F1} 0.23 vs. 0.38). FlowMDM avoids subsequence boundaries by generating the entire sequence jointly and thus achieves low transition jerk, yet its transition FID is several times higher than LT-HOI (6.16 vs. 1.12), indicating that low jerk alone does not yield realistic transitions. By leveraging both past and future contexts, LT-HOI achieves smooth transitions, accurate goal reaching, and better contact quality.

4.5 Analysis

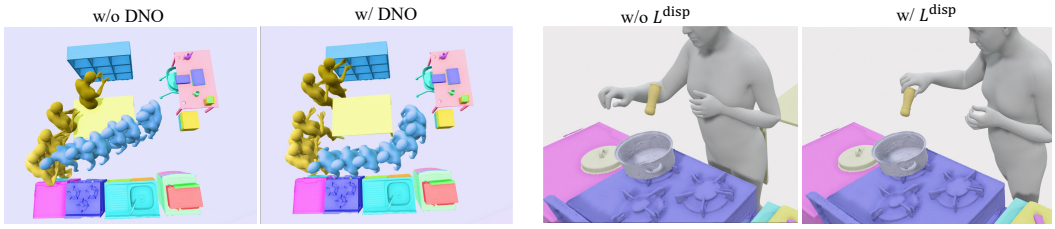
Effect of the generation strategy. To validate our generation scheme, which produces HOI motions in parallel within each subsequence and autoregressively across subsequences, we compare against

Table 2: **Effect of the generation strategy.** Generating each subsequence in parallel and composing subsequences autoregressively outperforms both variants on most metrics.

Method	Subsequence												Transition			Sequence	
	Motion							Contact					Motion	Jerk		Scene	
	MPJPE↓	H_{feet} →	FS↓	R@3↑	FID↓	T_{goal} ↓	O_{goal} ↓	C_{prec} ↑	C_{rec} ↑	C_{F1} ↑	$C_{\%}$ →	P_{hand} ↓	FID↓	PJ↓	AUJ↓	$Coll_{\%}$ ↓	P_{max} ↓
Joint	22.61	**-4.24**	0.48	0.41	3.35	37.19	0.56	0.35	0.23	0.23	0.23	**4.0e-6**	12.84	0.82	3.35	0.57	0.06
Autoregressive	35.23	0.37	0.42	0.15	55.91	121.49	1.30	0.03	0.01	0.01	0.01	4.4e-6	25.30	0.64	2.95	0.68	0.06
LT-HOI	**19.16**	-4.28	**0.33**	**0.54**	**0.45**	**24.41**	**0.48**	**0.49**	**0.36**	**0.38**	**0.35**	8.2e-5	**1.12**	**0.13**	**2.16**	**0.32**	**0.05**

Table 3: **Ablation study.** DNO reduces scene collisions and improves goal reaching, while $\mathcal{L}^{\texttt{disp}}$ improves contact quality.

Method	DNO	$\mathcal{L}^{\text{disp}}$	Subsequence												Transition			Sequence	
			Motion							Contact					Motion	Jerk		Scene	
			MPJPE↓	H_{feet} →	FS↓	R@3↑	FID↓	T_{goal} ↓	O_{goal} ↓	C_{prec} ↑	C_{rec} ↑	C_{F1} ↑	$C_{\%}$ →	P_{hand} ↓	FID↓	PJ↓	AUJ↓	$Coll_{\%}$ ↓	P_{max} ↓
LT-HOI	✓	✓	**19.16**	**-4.28**	0.33	**0.54**	0.45	**24.41**	**0.48**	**0.49**	**0.36**	**0.38**	**0.35**	8.2e-5	1.12	0.13	2.16	**0.32**	**0.05**
	-	✓	19.85	-4.47	0.35	**0.54**	**0.41**	25.70	0.52	0.46	0.35	0.36	0.34	8.6e-5	0.81	0.13	**2.15**	0.45	**0.05**
	-	-	20.65	-5.87	**0.30**	0.52	0.54	26.58	0.56	0.43	0.28	0.30	0.27	**6.5e-5**	**0.43**	**0.09**	**2.15**	0.48	**0.05**



(a) Effect of DNO (b) Effect of hand-object displacement loss

Figure 3: **Ablation study on DNO and $\mathcal{L}^{\texttt{disp}}$**

two variants: a *joint* variant that generates the entire sequence in parallel, and an *autoregressive* (AR) variant that generates motions sequentially in fixed-length chunks. For the AR variant, we use a chunk size of 32 frames. Table 2 presents the results.

LT-HOI achieves the best results on most metrics among the three schemes. The AR variant suffers from long-horizon error accumulation, showing high errors on goal-reaching metrics (T_{goal} 121.49, O_{goal} 1.30), which in turn limits contact quality (C_{F1} 0.01). The joint variant avoids this issue by modeling the entire sequence at once, but produces less precise interactions (C_{F1} 0.23, FID 3.35), likely because jointly modeling all subsequences makes it harder to capture fine-grained per-segment details. LT-HOI combines the strengths of both: parallel generation within each subsequence for local coherence, and autoregressive composition across subsequences for long-horizon consistency.

Ablation study. Table 3 progressively ablates DNO and the hand-object displacement loss $\mathcal{L}^{\texttt{disp}}$. Without DNO, the collision ratio ($\text{Coll}_{\%}$) increases from 0.32 to 0.45, and the goal-reaching error increases (T_{goal} 24.41 → 25.70). As illustrated in Fig. 3(a), DNO enables collision-avoiding navigation. Further analyses of DNO, including per-body-part collisions, the objective design, and runtime, are provided in Sec. C.4. Further removing $\mathcal{L}^{\texttt{disp}}$ degrades contact quality. Figure 3(b) shows that without $\mathcal{L}^{\texttt{disp}}$ the generated hands hover near the object without making proper contact, whereas applying it guides the hands toward the object surface for more plausible interactions.

Effect of past context. We analyze the effect of past context length $L^{\texttt{past}}$ on temporal composition quality by varying it from 0 to 256 while fixing $L^{\texttt{future}} = 64$, as shown in Fig. 4. Without any past context ($L^{\texttt{past}} = 0$), the peak jerk reaches 0.58, indicating abrupt transitions between consecutive subsequences. This is expected, as the model has no information about how the previous motion ended, forcing it to generate the beginning of each segment independently. Introducing even a single frame of past context ($L^{\texttt{past}} = 1$) reduces PJ from 0.58 to 0.19, indicating that even minimal past information substantially improves transition smoothness. As $L^{\texttt{past}}$ further increases, PJ continues to decrease, reaching 0.10 at $L^{\texttt{past}} = 256$, suggesting that longer past context provides richer kinematic history for the model to maintain temporal continuity. Beyond transition quality, longer past context also improves motion quality (MPJPE from 19.69 to 18.80), contact quality (F1 from 0.23 to 0.38), and goal-reaching accuracy (T_{goal} from 34.95 to 23.46). This indicates that richer temporal history

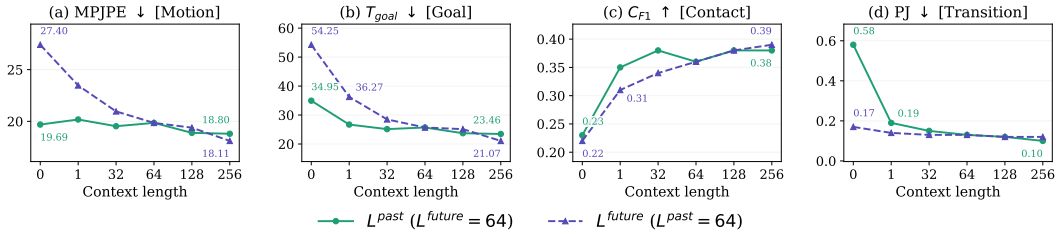


Figure 4: **Effect of past and future context length.** Past context drives transition smoothness (d), future context drives motion accuracy (a) and goal reaching (b), while contact F1 (c) benefits from both contexts.

Table 4: **Robustness to object motion perturbations.** Metrics remain close to the clean setting ($\sigma = 0$) up to $\sigma = 5$ cm, and contact recall decreases at $\sigma = 10$ cm.

σ (cm)	p (%)	T_{obj} (cm)	MPJPE↓	FS↓	R@3 ↑	FID↓	T_{goal} ↓	C_{prec} ↑	C_{rec} ↑	C_{F1} ↑	FID_{tr} ↓	PJ↓	$Coll_{\%}$ ↓
0	0	0.00	19.85	0.35	0.54	0.41	25.70	0.46	0.35	0.36	0.81	0.13	0.45
(A) Temporal coverage													
5	10	0.55	19.80	0.34	0.50	0.44	25.70	0.46	0.36	0.37	0.72	0.12	0.46
5	25	1.33	19.93	0.34	0.51	0.43	25.84	0.46	0.34	0.36	0.77	0.13	0.46
5	50	2.68	19.98	0.34	0.52	0.43	26.09	0.46	0.34	0.36	0.66	0.13	0.47
5	75	4.19	19.91	0.33	0.53	0.42	26.42	0.46	0.33	0.36	1.17	0.13	0.46
5	100	5.33	19.80	0.34	0.53	0.43	25.23	0.45	0.34	0.35	1.00	0.13	0.47
(B) Noise magnitude													
0.5	100	0.53	19.71	0.34	0.53	0.41	25.64	0.46	0.35	0.36	0.82	0.13	0.46
1	100	1.07	19.74	0.35	0.52	0.39	25.20	0.46	0.35	0.37	0.87	0.13	0.47
2	100	2.13	19.62	0.33	0.53	0.39	25.35	0.46	0.35	0.36	1.18	0.13	0.47
5	100	5.33	19.80	0.34	0.53	0.43	25.23	0.45	0.34	0.35	1.00	0.13	0.47
10	100	10.67	20.28	0.34	0.51	0.48	26.59	0.46	0.29	0.33	0.98	0.13	0.48

helps the model better coordinate not only smooth transitions, but also hand-object interactions and spatial navigation. Past context thus improves the coherence of compositions.

Effect of future context. We examine the effect of future context length $L^{\texttt{future}}$ by varying it from 0 to 256 while fixing $L^{\texttt{past}} = 64$, as shown in Fig. 4. The results reveal that future context plays a different role from past context. Without future context ($L^{\texttt{future}} = 0$), the largest degradation appears in goal-reaching accuracy: T_{goal} rises to 54.25, indicating that the model struggles to navigate toward the next interacting object. Contact quality also drops (F1 of 0.22), as the generated motion lacks awareness of where the upcoming interaction will occur and thus cannot position the body appropriately. Providing a single frame of future context ($L^{\texttt{future}} = 1$) already reduces T_{goal} from 54.25 to 36.27 and improves contact F1 from 0.22 to 0.31, indicating that even a minimal spatial cue about the next interaction target is informative for navigation. As $L^{\texttt{future}}$ increases further, MPJPE, contact F1, and goal-reaching accuracy continue to improve: MPJPE decreases from 27.40 to 18.11, contact F1 increases from 0.22 to 0.39, and T_{goal} drops to 21.07 at $L^{\texttt{future}} = 256$.
These results suggest that past and future contexts serve complementary roles. Past context mainly provides kinematic continuity by stabilizing transitions at segment boundaries, while future context provides spatial awareness of the upcoming interaction, guiding navigation toward the correct location for the next interaction. Combining both leads to improved overall performance, as each addresses a different aspect of long-term motion coherence.

Robustness to object motion perturbations. Since LT-HOI takes object motions as input, we evaluate its robustness to imperfect object trajectories. At inference time, we perturb object translations with temporally smooth random 3D offsets (low-pass filtered noise) of magnitude σ over a portion p of the trajectory, with DNO disabled; T_{obj} denotes the resulting mean object translation error. As shown in Table 4, the metrics remain comparable to the clean setting even when the entire trajectory is perturbed with $\sigma = 5$ cm (contact F1 0.35 vs. 0.36). Under $\sigma = 10$ cm, contact recall decreases (0.35 → 0.29) while precision is unchanged. Full results are provided in Sec. C.3.

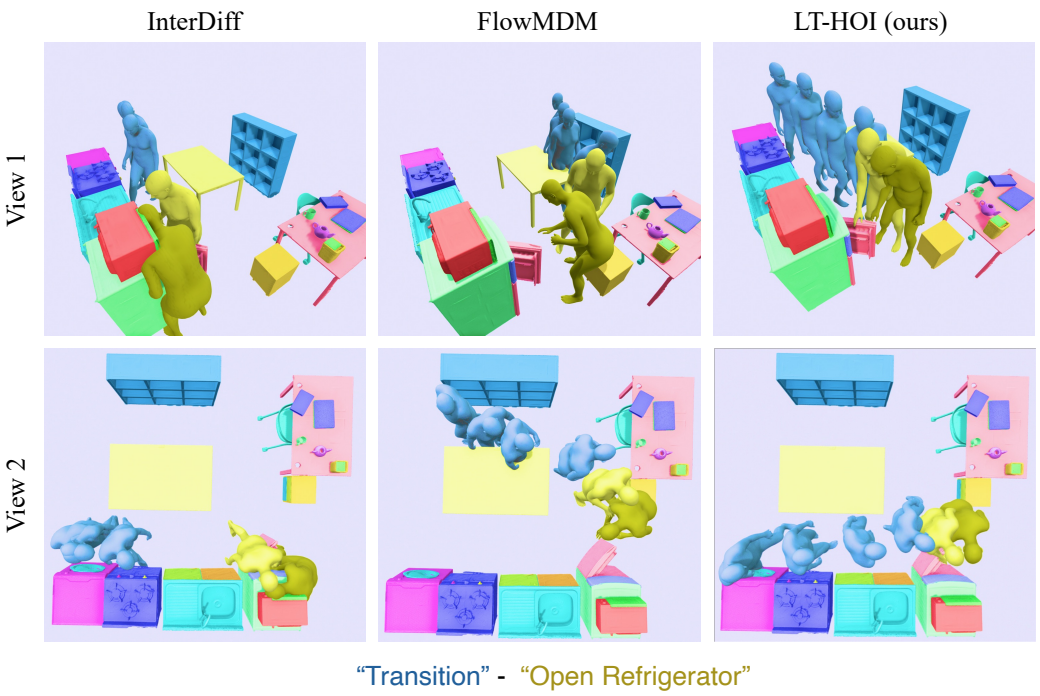


Figure 5: **Qualitative results** on a sequence of 'Transition' followed by 'Open Refrigerator', shown from two viewpoints with 20 frames uniformly sampled across the sequence.

4.6 Qualitative Results

Figure 5 presents qualitative comparisons between LT-HOI and the baselines. The example sequence consists of a 'Transition' motion followed by an 'Open Refrigerator' motion, with 20 frames uniformly sampled across the sequence. InterDiff reaches the next interaction location but skips the walking transition in between, resulting in an abrupt jump rather than continuous motion. FlowMDM produces reasonable locomotion but penetrates the desk. In contrast, LT-HOI generates smooth transitions between interaction locations while avoiding collisions with surrounding objects, producing plausible interaction motions throughout the sequence.

Perceptual study. We further conducted a perceptual study with 25 participants on 20 randomly sampled test sequences, where participants compared LT-HOI against InterDiff or FlowMDM in terms of semantic correctness, visual realism, and physical plausibility. LT-HOI is preferred over InterDiff in 95.2–96.0% and over FlowMDM in 70.4–80.0% of the comparisons across all three criteria, and the preference over FlowMDM rises to 89.5% when participants perceive a clear difference (Sec. D). This suggests that the quantitative improvements of LT-HOI translate into motions that humans perceive as more faithful to the task, more natural, and more physically plausible.

5 Conclusion

We have introduced the task of long-term action composition of HOIs, aiming to generate human motions that coherently connect multiple object manipulations within a 3D scene. We have proposed a diffusion-based generative model dubbed LT-HOI, which synthesizes temporally aligned motions conditioned on past and future contexts—each comprising text, object motion, and object geometry. Past context enables smooth transitions between interactions, while future context guides goal-directed, scene-aware locomotion toward upcoming objects. Experiments on the ParaHome dataset demonstrate that our framework synthesizes long, temporally coherent human motions. We hope this work provides a useful step toward goal-oriented, long-horizon motion generation for embodied AI.

Acknowledgements

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grants funded by the Korean government (MSIT) (RS-2024-00457882: National AI Research Lab Project (50%), RS-2022-II220264: Comprehensive Video Understanding and Generation with Knowledge-based Deep Logic Neural Network (45%), RS-2019-II191906: AI Graduate School Program at POSTECH (5%)). This work also benefited from Hi! PARIS and State funding managed by the ANR under the France 2030 program (reference: ANR-23-IACL-0005), as well as the ANR project CorVis (ANR-21-CE23-0003-01). We thank Jeonghwan Kim, Jisoo Kim, and Hanbyul Joo for their assistance in reproducing the baseline implementations on the ParaHome dataset.

References

[1] N. Athanasiou, M. Petrovich, M. J. Black, and G. Varol. TEACH: Temporal action composition for 3D humans. In *3DV*, pages 414–423, 2022.

[2] S. Baik, G. Kim, M. Choi, and H. Joo. Text-guided 6D object pose rearrangement via closed-loop VLM agents. In *ECCV*, 2026.

[3] G. Barquero, S. Escalera, and C. Palmero. Seamless human motion composition with blended positional encodings. In *CVPR*, pages 457–469, 2024.

[4] B. L. Bhatnagar, X. Xie, I. A. Petrov, C. Sminchisescu, C. Theobalt, and G. Pons-Moll. BEHAVE: Dataset and method for tracking human object interactions. In *CVPR*, pages 15935–15946, 2022.

[5] Y. Cao, L. Pan, K. Han, K.-Y. K. Wong, and Z. Liu. AvatarGO: Zero-shot 4D human-object interaction generation and animation. In *ICLR*, 2025.

[6] K. Clark, P. Vicol, K. Swersky, and D. J. Fleet. Directly fine-tuning diffusion models on differentiable rewards. In *ICLR*, 2024.

[7] C. Diller and A. Dai. CG-HOI: Contact-guided 3D human-object interaction generation. In *CVPR*, pages 19888–19901, 2024.

[8] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

[9] C. Guo, S. Zou, X. Zuo, S. Wang, W. Ji, X. Li, and L. Cheng. Generating diverse and natural 3D human motions from text. In *CVPR*, pages 5152–5161, 2022.

[10] C. He, J. Saito, J. Zachary, H. Rushmeier, and Y. Zhou. NeMF: Neural motion fields for kinematic animation. *NeurIPS*, 35:4244–4256, 2022.

[11] S. Huang, Z. Wang, P. Li, B. Jia, T. Liu, Y. Zhu, W. Liang, and S.-C. Zhu. Diffusion-based generation, optimization, and planning in 3D scenes. In *CVPR*, 2023.

[12] Y. Huang, O. Taheri, M. J. Black, and D. Tzionas. InterCap: Joint markerless 3D tracking of humans and objects in interaction. In *GCPR*, volume 13485, pages 281–299, 2022.

[13] I. Hwang, B. Zhou, Y. M. Kim, J. Wang, and C. Guo. SceneMI: Motion in-betweening for modeling human-scene interactions. In *ICCV*, 2025.

[14] N. Jiang, T. Liu, Z. Cao, J. Cui, Z. Zhang, Y. Chen, H. Wang, Y. Zhu, and S. Huang. Full-body articulated human-object interaction. In *ICCV*, pages 9365–9376, 2023.

[15] N. Jiang, Z. Zhang, H. Li, X. Ma, Z. Wang, Y. Chen, T. Liu, Y. Zhu, and S. Huang. Scaling up dynamic human-scene interaction modeling. In *CVPR*, pages 1737–1747, 2024.

[16] H. Jung, S. Lee, J. Yoon, S. Kim, G. Nam, Q. Huang, and S. Kim. Decoupled generative modeling for human-object interaction synthesis. In *CVPR*, 2026.

[17] K. Karunratanakul, K. Preechakul, E. Aksan, T. Beeler, S. Suwajanakorn, and S. Tang. Optimizing diffusion noise can serve as universal motion priors. In *CVPR*, 2024.

[18] J. Kim, J. Kim, J. Na, and H. Joo. ParaHome: Parameterizing everyday home activities towards 3D generative modeling of human-object interactions. In *CVPR*, pages 1816–1828, 2025.

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

[20] J. Li, A. Clegg, R. Mottaghi, J. Wu, X. Puig, and C. K. Liu. Controllable human-object interaction synthesis. In *ECCV*, pages 54–72, 2024.

[21] L. Li and A. Dai. HOI-PAGE: Zero-shot human-object interaction generation with part affordance guidance. In *ICML*, 2026.

[22] Z. Ling, Q. Shuai, T. Zhang, S. Li, B. Han, and C. Zou. PRISM: Streaming human motion generation with per-joint latent decomposition. *arXiv preprint arXiv:2603.08590*, 2026.

[23] Z. Lou, Q. Cui, T. Wang, Z. Song, L. Zhang, C. Cheng, H. Wang, X. Tang, H. Li, and H. Zhou. Harmonizing stochasticity and determinism: Scene-responsive diverse human motion prediction. *NeurIPS*, 37:39784–39811, 2024.

[24] J. Lu, C.-H. P. Huang, U. Bhattacharya, Q. Huang, and Y. Zhou. HUMOTO: A 4D dataset of mocap human object interactions. In *ICCV*, pages 10886–10897, 2025.

[25] X. Lv, L. Xu, Y. Yan, X. Jin, C. Xu, S. Wu, Y. Liu, L. Li, M. Bi, W. Zeng, and X. Yang. HIMO: A new benchmark for full-body human interacting with multiple objects. In *ECCV*, pages 300–318, 2024.

[26] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. Osman, D. Tzionas, and M. J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *CVPR*, pages 10975–10985, 2019.

[27] X. Peng, Y. Xie, Z. Wu, V. Jampani, D. Sun, and H. Jiang. HOI-Diff: Text-driven synthesis of 3D human-object interactions using diffusion models. In *CVPR*, pages 2878–2888, 2025.

[28] I. A. Petrov, R. Marin, J. Chibane, and G. Pons-Moll. TriDi: Trilateral diffusion of 3D humans, objects, and interactions. In *ICCV*, pages 5523–5535, 2025.

[29] I. A. Petrov, V. Guzov, R. Marin, E. Aksan, X. Chen, D. Cremers, T. Beeler, and G. Pons-Moll. ECHO: Ego-centric modeling of human-object interactions. In *ECCV*, 2026.

[30] M. Petrovich, M. J. Black, and G. Varol. TEMOS: Generating diverse human motions from textual descriptions. In *ECCV*, pages 480–497, 2022.

[31] S. Prokudin, C. Lassner, and J. Romero. Efficient learning on point clouds with basis point sets. In *ICCV*, pages 4332–4341, 2019.

[32] C. R. Qi, L. Yi, H. Su, and L. J. Guibas. PointNet++: Deep hierarchical feature learning on point sets in a metric space. *NeurIPS*, 2017.

[33] A. Radford, J. W. Kim, C. Chen, A. v. d. Oord, and I. Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021.

[34] Y. Shafir, G. Tevet, R. Kapon, and A. H. Bermano. Human motion diffusion as a generative prior. In *ICLR*, 2024.

[35] J. Song, C. Meng, and S. Ermon. Denoising diffusion implicit models. In *ICLR*, 2021.

[36] W. Song, X. Zhang, S. Li, Y. Gao, A. Hao, X. Hou, C. Chen, N. Li, and H. Qin. HOIAnimator: Generating text-prompt human-object animations using novel perceptive diffusion models. In *CVPR*, pages 811–820, 2024.

[37] O. Taheri, N. Ghorbani, M. J. Black, and D. Tzionas. GRAB: A dataset of whole-body human grasping of objects. In *ECCV*, pages 581–600, 2020.

[38] G. Tevet, S. Raab, B. Gordon, Y. Shafir, D. Cohen-or, and A. H. Bermano. Human motion diffusion model. In *ICLR*, 2023.

[39] B. Wen, J. Tremblay, V. Blukis, S. Tyree, T. Müller, A. Evans, D. Fox, J. Kautz, and S. Birchfield. BundleSDF: Neural 6-DoF tracking and 3D reconstruction of unknown objects. In *CVPR*, pages 606–617, 2023.

[40] B. Wen, W. Yang, J. Kautz, and S. Birchfield. FoundationPose: Unified 6D pose estimation and tracking of novel objects. In *CVPR*, pages 17868–17879, 2024.

[41] Z. Wu, J. Li, P. Xu, and C. K. Liu. Human-object interaction from human-level instructions. In *ICCV*, pages 11176–11186, 2025.

[42] L. Xiao, S. Lu, H. Pi, K. Fan, L. Pan, Y. Zhou, Z. Feng, X. Zhou, S. Peng, and J. Wang. MotionStreamer: Streaming motion generation via diffusion-based autoregressive model in causal latent space. In *ICCV*, pages 10086–10096, 2025.

[43] S. Xu, Z. Li, Y.-X. Wang, and L.-Y. Gui. InterDiff: Generating 3D human-object interactions with physics-informed diffusion. In *ICCV*, pages 14928–14940, 2023.

[44] H. Yi, J. Thies, M. J. Black, X. B. Peng, and D. Rempe. Generating human interaction motions in scenes with text control. In *ECCV*, pages 246–263, 2024.

[45] T. Yoshida, S. Kurita, T. Nishimura, and S. Mori. Generating 6DoF object manipulation trajectories from action description in egocentric vision. In *CVPR*, pages 17370–17382, 2025.

[46] Q. Zhang, J. Song, X. Huang, Y. Chen, and M.-Y. Liu. DiffCollage: Parallel generation of large content with diffusion models. In *CVPR*, pages 10188–10198, 2023.

[47] W. Zhang, R. Dabral, V. Golyanik, V. Choutas, E. Alvarado, T. Beeler, M. Habermann, and C. Theobalt. BimArt: A unified approach for the synthesis of 3D bimanual interaction with articulated objects. In *CVPR*, pages 27694–27705, 2025.

[48] K. Zhao, Y. Zhang, S. Wang, T. Beeler, and S. Tang. Synthesizing diverse human motions in 3D indoor scenes. In *ICCV*, 2023.

[49] K. Zhao, G. Li, and S. Tang. DartControl: A diffusion-based autoregressive motion model for real-time text-driven motion control. In *ICLR*, 2025.

[50] Y. Zhou, C. Barnes, J. Lu, J. Yang, and H. Li. On the continuity of rotation representations in neural networks. In *CVPR*, pages 5745–5753, 2019.

Appendix

In this supplementary material, we provide evaluation metrics in Sec. A, details of the baselines, their past and future conditioned variants, and their short-term HOI synthesis results in Sec. B, additional experimental results in Sec. C (including additional ablation studies, scene representation, robustness to object motion perturbations, DNO analysis, and results on the HIMO dataset), a perceptual study in Sec. D, and discussions in Sec. E.

A Evaluation Metrics

In this section, we provide detailed explanations of evaluation metrics used in our experiments (Sec. 4.2). Following the previous work on atomic HOI [20] and temporal action composition of human motions [3], we evaluate the generated motions in three parts—subsequence, transition, and sequence—using the following metrics. The symbols $\uparrow$ and $\downarrow$ indicate that higher or lower values correspond to better performance, respectively. The symbol $\rightarrow$ indicates that the closer the value is to GT, the better.

Mean per joint position error (MPJPE) $\downarrow$: MPJPE measures the average Euclidean distance between the predicted and GT 3D joint positions, reported in centimeters.

Foot height $\rightarrow$: Foot height measures the foot height from the ground, reported in centimeters.

Foot sliding $\downarrow$: The foot sliding is computed as the weighted accumulated foot translation on the ground plane (xy-plane), following prior work [10, 20], reported in centimeters.

R@3 $\uparrow$: R@3 evaluates motion-to-text retrieval accuracy within the top-3 candidates, following Guo et al. [9]. It measures the capability of a motion representation to be semantically aligned with its corresponding natural language description. A higher R@3 score indicates better motion-text correspondence and representation quality. For evaluation, we adhere to the standard protocol by training separate motion and text encoders on the training split of the dataset. This training procedure follows the official implementation of Guo et al. [9].

Fréchet Inception Distance (FID) $\downarrow$: FID measures the distributional discrepancy between the predicted motions and the ground-truth motions in the learned feature space. We leverage the same motion encoder trained on the ParaHome training split for the $\mathbf{R@3}$ evaluation, ensuring consistency in the feature space used for distribution assessment.

Goal reaching translation error (T_{goal}) $\downarrow$: T_{goal} measures how accurately the human reaches the target object at the moment of interaction. It is computed as the Euclidean distance between the human pelvis position and the target object position at the first frame of contact, reported in centimeters.

Goal reaching orientation error (O_{goal}) $\downarrow$: O_{goal} measures the angular alignment between the human body and the target object at the moment of interaction. It is computed as the angular difference (in radians) between the body's forward direction and the direction toward the target object position, evaluated at the first frame of contact.

Contact: We evaluate hand–object interaction quality using contact precision $\mathbf{C_{prec}} \uparrow$, recall $\mathbf{C_{rec}} \uparrow$, F1 score $\mathbf{C_{F1}} \uparrow$, and contact percentage $\mathbf{C}_{\%} \rightarrow$ [19, 20].

Contact is defined per frame by computing the minimum distance between each wrist joint (left and right) and all object vertices; if the minimum distance is below a fixed threshold, the frame is considered to be in contact. The same procedure is applied to both ground-truth and predicted motions, and we compute precision, recall, and F1 based on the resulting binary contact labels. Contact percentage measures the proportion of frames in which contact occurs, providing an overall estimate of how frequently the hand interacts with the object throughout the sequence.

$\mathbf{P_{hand}}$ $\downarrow$: Hand penetration distance measures the average depth of hand–object interpenetration using a signed distance field (SDF). Specifically, the score is computed as the mean absolute value of all negative signed distances. Note that the penetration score is computed for the objects that interact with the hands.

Peak jerk (PJ) $\downarrow$: PJ measures abrupt motion changes during transitions. The metric is defined as the maximum jerk value observed across all joints during the transition motion, where jerk is the third derivative of joint position with respect to time [3].

Area under jerk (AUJ) $\downarrow$: AUJ evaluates the overall motion smoothness of the transition motion. It is computed as the difference between the maximum jerk across all joints and the average dataset jerk value, and this difference is further averaged over all transition motions [3].

Scene collision ratio ($\text{Coll}_{\%}$) $\downarrow$: $\text{Coll}_{\%}$ quantifies how often the generated human body interpenetrates the static 3D scene. We voxelize each scene into a 3D signed distance field (SDF) and, for every

frame, query the SDF at the body vertices obtained from the SMPL-X forward pass. To exclude trivial floor contact, only vertices above a fixed floor height of 10 cm are considered. A frame is counted as colliding when at least one queried vertex has a signed distance below -3 cm, i.e., a penetration depth larger than 3 cm. $\text{Coll}_{\%}$ is then computed as the fraction of colliding frames over the total sequence length and averaged across all test sequences [13].

Scene penetration max ($\mathbf{P}_{\max}$) ↓: $\text{P}_{\max}$ measures the worst-case penetration depth between the body and the scene. Using the same 3D SDF query as $\text{Coll}_{\%}$, the per-frame penetration depth is defined as the maximum of $\max(-d, 0)$ over all body vertices, where d denotes the signed distance to the scene. $\text{P}_{\max}$ takes the maximum of these per-frame values across the entire sequence and is averaged across test sequences, reported in meters. While $\text{Coll}_{\%}$ captures how frequently collisions occur, $\text{P}_{\max}$ reflects how severe the most extreme collision is [13].

B Comparison with Baselines

We compare LT-HOI against five diffusion baselines: MDM [38], CHOIS [20], DecHOI [16], InterDiff [43], and FlowMDM [3].

B.1 Implementation Details

For a fair comparison, all baselines share the motion representation, diffusion schedule (cosine, 1,000 steps, x_0-prediction), CLIP ViT-B/32 text encoder, BPS encoder, ViT scene-BEV encoder, and transformer hyperparameters (d_{model}=256, 4 heads, 8 layers, FFN size 1,024) of LT-HOI. The shared conditioning inputs are a per-segment text embedding $\boldsymbol{w}_s \in \mathbb{R}^{512}$, per-segment left/right object BPS $\boldsymbol{g}_s^L, \boldsymbol{g}_s^R \in \mathbb{R}^{1024\times 3}$, a per-segment scene BEV image $\boldsymbol{s}_s \in \mathbb{R}^{3\times 224\times 224}$, and per-frame object motion $\boldsymbol{o}_t^L, \boldsymbol{o}_t^R \in \mathbb{R}^9$ (3D translation + 6D rotation). The five baselines differ in how these conditions are injected and whether past/future contexts are used. Object motions are concatenated with human motions channel-wise.

MDM [38]. A bidirectional transformer that denoises the current subsequence in one pass, conditioned only on the last frame of the previous subsequence ($L^{\texttt{past}}$=1, $L^{\texttt{future}}$=0), analogous to the initial-pose condition used in short-term HOI synthesis. Each frame token is the linear projection of $[\boldsymbol{x}_t, \boldsymbol{o}_t^L, \boldsymbol{o}_t^R]$. All segment-level conditions are summed into a single per-segment token $\boldsymbol{c}_s = \tau(t) + \phi_{\text{text}}(\boldsymbol{w}_s) + \phi_{\text{scene}}(\boldsymbol{s}_s) + \phi_{\text{bps}}(\boldsymbol{g}_s^L) + \phi_{\text{bps}}(\boldsymbol{g}_s^R)$, where $\tau(t)$ is the diffusion-timestep embedding and each ϕ is an MLP into $\mathbb{R}^{d_{\text{model}}}$. A segment-restricted attention mask lets each frame attend only to its own segment's $\boldsymbol{c}_s$.

CHOIS [20]. Identical to MDM except that BPS is moved from the segment token to per-frame tokens, providing dense geometric guidance. Frame tokens become $[\boldsymbol{x}_t, \boldsymbol{o}_t^L, \boldsymbol{o}_t^R, \boldsymbol{g}_{s(t)}^L, \boldsymbol{g}_{s(t)}^R]$ where $s(t)$ is the segment containing frame t, and the segment token reduces to $\boldsymbol{c}_s = \tau(t) + \phi_{\text{text}}(\boldsymbol{w}_s) + \phi_{\text{scene}}(\boldsymbol{s}_s)$.

DecHOI [16]. Similar to CHOIS, DecHOI is designed for atomic HOI generation. Following its formulation, which takes object motion, BPS features, and a text embedding as input, we additionally provide the scene embedding alongside the text embedding for global conditioning. We also adopt the decoupled path denoiser and HOI adversarial discriminator from the original method for trajectory generation and interaction modeling, respectively, based on their official code repository. As DecHOI mainly tackles atomic motion generation, we follow the same setting as MDM and CHOIS ($L^{\texttt{past}}$=1, $L^{\texttt{future}}$=0).

InterDiff [43]. An encoder–decoder transformer that uses $L^{\texttt{past}}$=64 past frames and $L^{\texttt{future}}$=0, matching its original prediction setting. The encoder ingests the past frames, with each token aggregating a body-motion projection, per-slot object-motion projections gated by a hand-index embedding, the hand-specific BPS, and the past-segment text embedding. The decoder denoises the body target over the past and current windows; each decoder token additionally incorporates per-frame BPS, per-frame text, the broadcast scene feature, and the timestep token, and attends to the encoder via standard cross-attention.

Table S1: **Past and future conditioning for baselines**

Method	L^{past}	L^{future}	Subsequence													Transition			Sequence	
			Motion							Contact						Motion	Jerk		Scene	
			MPJPE↓	H_{feet} →	FS↓	R@3 ↑	FID↓	T_{goal} ↓	O_{goal} ↓	C_{prec} ↑	C_{rec} ↑	C_{F1} ↑	$C_{\%}$ →	P_{hand} ↓		FID↓	PJ↓	AUJ↓	$\text{Coll}_{\%}$ ↓	$P_{\max}$ ↓
GT	-	-	-	-3.30	0.29	-	-	-	-	-	-	-	0.49	1.7e-6		-	0.03	2.22	0.35	0.04
MDM	1	0	28.76	-6.39	0.33	0.43	0.70	54.14	1.04	0.34	0.18	0.21	0.18	6.2e-5		1.59	0.16	2.05	0.67	0.07
MDM†	64	64	22.50	-4.82	0.30	0.49	0.45	31.47	0.65	0.45	0.29	0.32	0.29	6.4e-5		0.70	0.10	2.06	0.52	0.06
CHOIS	1	0	28.47	-6.90	0.34	0.44	1.22	54.81	1.07	0.33	0.18	0.20	0.17	3.4e-5		1.21	0.17	2.09	0.60	0.06
CHOIS†	64	64	20.83	-3.50	0.29	0.50	0.32	29.17	0.58	0.44	0.30	0.32	0.29	8.7e-5		0.49	0.09	2.06	0.48	0.05
DecHOI	1	0	27.26	-3.43	0.39	0.44	0.53	50.33	0.98	0.40	0.23	0.26	0.24	8.7e-5		2.44	0.21	2.23	0.58	0.06
DecHOI†	64	64	20.21	-1.09	0.40	0.49	0.62	24.91	0.51	0.48	0.37	0.38	0.37	1.2e-4		1.81	0.13	2.11	0.49	0.06
InterDiff	64	0	23.17	-5.17	0.28	0.36	1.09	39.85	0.65	0.38	0.20	0.23	0.19	2.5e-6		4.60	0.80	3.44	0.51	0.06
InterDiff†	64	64	19.84	-3.86	0.31	0.49	0.18	28.03	0.43	0.40	0.25	0.27	0.24	4.5e-6		5.24	0.58	2.98	0.48	0.05

Table S2: **Short-term HOI synthesis of baselines**

Method	MPJPE↓	H_{feet} →	FS↓	R@3 ↑	FID↓	T_{goal} ↓	O_{goal} ↓	C_{prec} ↑	C_{rec} ↑	C_{F1} ↑	$C_{\%}$ →	P_{hand} ↓
GT	–	-3.66	0.28	0.49	–	–	–	–	–	–	0.70	–
MDM	12.63	-4.40	0.24	0.50	0.26	17.59	0.33	0.70	0.50	0.53	0.47	1.4e-4
CHOIS	12.37	-4.30	0.23	0.49	0.22	16.94	0.35	0.73	0.53	0.56	0.49	1.3e-4
DecHOI	12.04	-4.26	0.27	0.50	0.22	15.86	0.32	0.72	0.57	0.60	0.54	1.3e-4

FlowMDM [3]. A bidirectional denoiser that generates the full sequence in one pass. FlowMDM replaces MDM's self-attention with the Pose-Centric Cross-Attention (PCCAT) block and its positional encoding with Blended Positional Encoding (BPE). Conditioning enters as a per-frame vector $\boldsymbol{c}_t = \boldsymbol{w}_{s(t)} + \boldsymbol{g}^L_{s(t)} + \boldsymbol{g}^R_{s(t)} + \boldsymbol{s}_{s(t)} + \tau(t)$. Inside each PCCAT layer, queries are formed from the concatenation of the motion feature and $\boldsymbol{c}_t$, while keys and values come from the motion feature alone, so that the condition steers attention without entering the value space.

B.2 Past and Future Conditioning for Baselines

In Table 1, each baseline follows its original temporal setting: MDM, CHOIS, and DecHOI are short-horizon models conditioned only on the last frame of the previous subsequence (L^{past}=1, L^{future}=0), and InterDiff forecasts future motion from past observations (L^{past}=64, L^{future}=0). To examine whether the gap stems from temporal context rather than architecture, we extend each baseline to take both past and future contexts with L^{past}=L^{future}=64, as in LT-HOI, and denote the extended variants with †. All models are retrained from scratch on ParaHome for the same number of iterations, without using released weights; Table S1 lists the temporal context of each model. FlowMDM has no † variant as it generates the whole sequence jointly.

For MDM†, we construct segment-wise conditioning tokens for the past, current, and future subsequences from CLIP text features [33], BPS, and ViT scene features [8], fuse them with the diffusion timestep embedding, and apply segment-wise masking in the self-attention layers so that each frame attends to the token of its own subsequence. CHOIS† and DecHOI† follow the same scheme, with tokens built from text and scene features only, since BPS is already provided per frame in these models. For InterDiff†, the encoder ingests the past motion as in the original model, while the decoder jointly predicts the past, current, and future motions.

Results. As shown in Table S1, all four † variants improve over their original counterparts on most subsequence motion, contact, and transition metrics. Compared with a single past frame, longer past and future contexts reduce the goal-reaching error (*e.g.*, 54.81 → 29.17 for CHOIS and 54.14 → 31.47 for MDM) and the transition FID (1.21 → 0.49 for CHOIS and 1.59 → 0.70 for MDM), indicating that the contexts help bridge adjacent subsequences. DecHOI† likewise reduces MPJPE from 27.26 to 20.21, and InterDiff†, which already observes the past, further improves subsequence motion and contact quality with the added future context. These gains across architecturally distinct baselines suggest that past and future conditioning is a key factor for long-term HOI composition, beyond architectural differences.

Table S3: **Effect of $\lambda^{\texttt{disp}}$**

$\lambda^{\texttt{disp}}$	Subsequence												Transition			Sequence	
	Motion							Contact					Motion	Jerk		Scene	
	MPJPE↓	H_{feet} →	FS↓	R@3 ↑	FID↓	T_{goal} ↓	O_{goal} ↓	C_{prec} ↑	C_{rec} ↑	C_{F1} ↑	$C_{\%}$ →	P_{hand} ↓	FID↓	PJ↓	AUJ↓	$Coll_{\%}$ ↓	P_{max} ↓
GT	-	-3.30	0.29	-	-	-	-	-	-	-	0.49	1.7e-6	-	0.03	2.22	0.35	0.04
0.1	20.20	-4.52	0.30	0.52	0.55	26.27	0.52	0.46	0.35	0.37	0.34	8.5e-5	0.26	0.10	2.16	0.44	0.05
0.2	19.77	-4.08	0.32	0.53	0.34	25.07	0.52	0.47	0.35	0.37	0.34	6.9e-5	0.23	0.11	2.16	0.45	0.05
0.3	19.85	-4.47	0.35	0.54	0.41	25.70	0.52	0.46	0.35	0.36	0.34	8.6e-5	0.81	0.13	2.15	0.46	0.05
0.4	19.57	-4.84	0.34	0.54	0.27	25.02	0.51	0.47	0.37	0.38	0.35	8.0e-5	0.46	0.14	2.17	0.46	0.05
0.5	19.92	-4.90	0.35	0.50	0.39	25.75	0.53	0.47	0.37	0.38	0.36	7.9e-5	0.75	0.15	2.19	0.46	0.05

Table S4: **Self-attention vs. cross-attention**

Attn type	Subsequence												Transition			Sequence	
	Motion							Contact					Motion	Jerk		Scene	
	MPJPE↓	H_{feet} →	FS↓	R@3 ↑	FID↓	T_{goal} ↓	O_{goal} ↓	C_{prec} ↑	C_{rec} ↑	C_{F1} ↑	$C_{\%}$ →	P_{hand} ↓	FID↓	PJ↓	AUJ↓	$Coll_{\%}$ ↓	P_{max} ↓
GT	-	-3.30	0.29	-	-	-	-	-	-	-	0.49	1.7e-6	-	0.03	2.22	0.35	0.04
Self-attn	19.90	-5.49	0.39	0.52	0.49	25.73	0.51	0.47	0.36	0.37	0.35	8.8e-5	1.44	0.20	2.20	0.46	0.05
Cross-attn	19.85	-4.47	0.35	0.54	0.41	25.70	0.52	0.46	0.35	0.36	0.34	8.6e-5	0.81	0.13	2.15	0.46	0.05

B.3 Short-term HOI Synthesis of Baselines

We evaluate the short-term HOI synthesis models (MDM, CHOIS, and DecHOI) on atomic HOI generation on ParaHome. Following the setup of short-term HOI synthesis [16, 20], we retrain these models with the ground-truth initial human pose as input; all other training inputs and model configurations are identical to Table 1. Among the 989 test subsequences in ParaHome, 688 subsequences without transition motions are used in the evaluation. As shown in Table S2, all baselines achieve results close to real motion on most metrics in the atomic HOI setting. Semantic correctness remains similar (R@3: 0.49–0.50 vs. 0.49), and low-level artifacts are also close to ground truth (foot height: -4.26 to -4.40 cm vs. -3.66 cm; foot sliding: 0.23–0.27 vs. 0.28). Although contact frequency is lower (0.47–0.54 vs. 0.70), the gap is modest compared with the degradation in long-term composition (MPJPE: 27.26–28.76 cm, contact F1: 0.20–0.26 in Table 1). These results suggest that their degradation reflects the difficulty of composition rather than an inability to synthesize atomic HOI on ParaHome or a domain gap with the dataset. Long-term composition requires maintaining transition continuity, anticipating future object interactions, and adapting to changing scene contexts—challenges that LT-HOI addresses through past and future conditioning.

C Additional Experiments

In this section, we provide additional experimental results. Unless otherwise specified, we set the past context length $L^{\texttt{past}}$ and future context length $L^{\texttt{future}}$ to 64, and set the number of decoder layers $N^{\texttt{dec}}$ to 8.

C.1 Ablation Studies

Effect of $\lambda^{\texttt{disp}}$. To analyze the effect of the hand–object displacement loss weight $\lambda^{\texttt{disp}}$, we vary its value from 0.1 to 0.5. As shown in Table S3, increasing $\lambda^{\texttt{disp}}$ generally improves contact metrics, indicating that the displacement loss encourages more accurate hand–object interactions. However, larger values slightly degrade motion quality. Balancing contact performance and motion quality, we set $\lambda^{\texttt{disp}} = 0.3$ and fix this value for all subsequent experiments.

Self-attention vs. cross-attention. To analyze the effect of the attention strategy for multimodal conditioning, we compare our cross-attention design with a variant that replaces cross-attention layers with self-attention. In this variant, all multimodal conditions—scene, object geometry, and text features—are concatenated to the input sequence of the denoising transformer f, analogous to the diffusion timestep embedding shown in Fig. 2. As shown in Table S4, cross-attention improves motion quality and transition smoothness (*e.g.*, transition FID 1.44 → 0.81, PJ 0.20 → 0.13), while contact metrics remain comparable. We hypothesize that cross-attention decouples the motion and conditioning representation spaces: motion tokens attend to conditioning signals without the

Table S5: **Effect of the number of decoder layers $N^{\texttt{dec}}$**

$N^{\texttt{dec}}$	Subsequence												Transition			Sequence	
	Motion							Contact					Motion	Jerk		Scene	
	MPJPE↓	H_{feet} →	FS↓	R@3↑	FID↓	T_{goal} ↓	O_{goal} ↓	C_{prec} ↑	C_{rec} ↑	C_{F1} ↑	$C_{\%}$ →	P_{hand} ↓	FID↓	PJ↓	AUJ↓	$Coll_{\%}$ ↓	P_{max} ↓
GT	-	-3.30	0.29	-	-	-	-	-	-	-	0.49	1.7e-6	-	0.03	2.22	0.35	0.04
6	20.11	-6.04	0.36	0.54	0.43	25.64	0.53	0.48	0.37	0.38	0.36	1.1e-4	0.54	0.11	2.16	0.45	0.05
8	19.85	-4.47	0.35	0.54	0.41	25.70	0.52	0.46	0.35	0.36	0.34	8.6e-5	0.81	0.13	2.15	0.46	0.05
10	19.90	-4.73	0.34	0.55	0.43	25.84	0.55	0.46	0.35	0.36	0.34	1.0e-4	0.54	0.15	2.15	0.47	0.05

Table S6: **Numerical results of Fig. 4**

method	L^{past}	L^{future}	Subsequence													Transition			Sequence	
			Motion							Contact						Motion	Jerk		Scene	
			MPJPE↓	H_{feet} →	FS↓	R@3↑	FID↓	T_{goal} ↓	O_{goal} ↓	C_{prec} ↑	C_{rec} ↑	C_{F1} ↑	$C_{\%}$ →	P_{hand} ↓		FID↓	PJ↓	AUJ↓	$Coll_{\%}$ ↓	P_{max} ↓
GT	-	-	-	-3.30	0.29	-	-	-	-	-	-	-	0.49	1.7e-6		-	0.03	2.22	0.35	0.04
Effect of past context length ($L^{future} = 64$)																				
LT-HOI	0	64	19.69	-0.14	0.32	0.51	0.46	34.95	0.51	0.32	0.22	0.23	0.21	6.2e-5		20.72	0.58	2.92	0.42	0.05
	1	64	20.18	-4.83	0.31	0.52	0.36	26.71	0.55	0.46	0.33	0.35	0.32	8.5e-5		1.15	0.19	2.13	0.47	0.05
	32	64	19.53	-3.98	0.34	0.53	0.31	25.14	0.51	0.47	0.36	0.38	0.35	7.0e-5		0.63	0.15	2.16	0.43	0.05
	64	64	19.85	-4.47	0.35	0.54	0.41	25.70	0.52	0.46	0.35	0.36	0.34	8.6e-5		0.81	0.13	2.15	0.45	0.05
	128	64	18.89	-4.44	0.33	0.52	0.42	23.72	0.48	0.48	0.37	0.38	0.35	7.9e-5		0.25	0.12	2.15	0.44	0.05
	256	64	18.80	-4.54	0.31	0.53	0.53	23.46	0.47	0.47	0.36	0.38	0.36	7.2e-5		0.22	0.10	2.15	0.44	0.05
Effect of future context length ($L^{past} = 64$)																				
LT-HOI	64	0	27.40	-7.04	0.33	0.43	0.88	54.25	0.99	0.34	0.20	0.22	0.20	3.5e-5		1.42	0.17	2.19	0.56	0.06
	64	1	23.47	-7.30	0.33	0.47	0.70	36.27	0.70	0.42	0.29	0.31	0.29	7.1e-5		1.76	0.14	2.17	0.54	0.06
	64	32	20.96	-3.03	0.34	0.52	0.22	28.49	0.59	0.45	0.33	0.34	0.32	7.5e-5		0.86	0.13	2.16	0.47	0.05
	64	64	19.85	-4.47	0.35	0.54	0.41	25.70	0.52	0.46	0.35	0.36	0.34	8.6e-5		0.81	0.13	2.15	0.45	0.05
	64	128	19.39	-6.75	0.33	0.53	0.58	25.13	0.50	0.47	0.37	0.38	0.35	1.1e-4		0.24	0.12	2.16	0.47	0.05
	64	256	18.11	-5.25	0.34	0.53	0.53	21.07	0.41	0.48	0.37	0.39	0.36	7.9e-5		0.22	0.12	2.14	0.43	0.05

Table S7: **Effect of scene encoder**

method	Subsequence												Transition			Sequence	
	Motion							Contact					Motion	Jerk		Scene	
	MPJPE↓	H_{feet} →	FS↓	R@3↑	FID↓	T_{goal} ↓	O_{goal} ↓	C_{prec} ↑	C_{rec} ↑	C_{F1} ↑	$C_{\%}$ →	P_{hand} ↓	FID↓	PJ↓	AUJ↓	$Coll_{\%}$ ↓	P_{max} ↓
GT	-	-3.30	0.29	-	-	-	-	-	-	-	0.49	1.7e-6	-	0.03	2.22	0.35	0.04
CNN	19.65	-5.94	0.34	0.52	0.74	24.95	0.50	0.47	0.36	0.37	0.35	8.9e-5	0.61	0.13	2.15	0.45	0.05
ViT	19.85	-4.47	0.35	0.54	0.41	25.70	0.52	0.46	0.35	0.36	0.34	8.6e-5	0.81	0.13	2.15	0.46	0.05

conditioning features being modified in return, allowing the model to preserve task-relevant motion structure while incorporating external guidance.

Effect of the number of decoder layers $N^{\texttt{dec}}$. We conduct an ablation study by varying the number of decoder layers $N^{\texttt{dec}}$ from 6 to 10, as shown in Table S5. Performance varies modestly across depths, and the improvements from 8 to 10 layers are marginal compared to the increased computational cost. Therefore, we adopt $N^{\texttt{dec}} = 8$ as the default configuration to balance performance and efficiency.

Numerical results of Fig. 4. Full results of Fig. 4, *i.e.*, the effect of past and future context lengths, are presented in Table S6.

C.2 Scene Representation

Effect of scene encoder. To verify the effect of the scene encoder architecture, we compare CNN and ViT as the scene encoder backbone in Table S7. CNN captures local spatial features through hierarchical convolutions, while ViT models global spatial context via self-attention over image patches. In this experiment, we use a 4-layer CNN. Although CNN yields marginally better motion quality—likely due to its locality bias being well-suited for processing the structured 2D scene layout—both variants achieve comparable performance across most metrics. The CNN encoder achieves slightly lower MPJPE (19.65 vs. 19.85) and better T_{goal} (24.95 vs. 25.70), while ViT achieves better subsequence FID (0.41 vs. 0.74). These differences are minor, suggesting that our framework is robust to the choice of scene encoder architecture. We adopt ViT as the default due to its flexibility and scalability.

2D BEV vs. 3D voxel scene representation. To evaluate whether a 3D scene representation benefits full-body HOI generation, we replace our BEV scene representation with a 3D voxel. Specifically,

Table S8: **2D BEV vs. 3D voxel scene representation**

Scene	MPJPE↓	H_{feet} →	FS↓	R@3 ↑	FID↓	T_{goal} ↓	O_{goal} ↓	C_{prec} ↑	C_{rec} ↑	C_{F1} ↑	$C_{\%}$ →	P_{hand} ↓	FID_{tr} ↓	PJ↓	AUJ↓	$Coll_{\%}$ ↓	P_{max} ↓
GT	–	-3.30	0.29	0.49	–	–	–	–	–	–	0.49	1.7e-6	–	0.03	2.22	0.35	0.04
2D BEV	19.85	-4.47	0.35	0.54	0.41	25.70	0.52	0.46	0.35	0.36	0.34	8.6e-5	0.81	0.13	2.15	0.45	0.05
3D voxel	19.16	-4.34	0.33	0.52	0.48	24.95	0.51	0.47	0.37	0.38	0.36	7.4e-5	0.34	0.13	2.16	0.44	0.05

Table S9: **Robustness to object motion perturbations**

σ (cm)	p (%)	T_{obj} (cm)	MPJPE↓	H_{feet} →	FS↓	R@3 ↑	FID↓	T_{goal} ↓	O_{goal} ↓	C_{prec} ↑	C_{rec} ↑	C_{F1} ↑	$C_{\%}$ →	P_{hand} ↓	FID_{tr} ↓	PJ↓	AUJ↓	$Coll_{\%}$ ↓	P_{max} ↓
GT	–	–	-	-3.30	0.29	0.49	–	–	–	–	–	–	0.49	1.7e-6	–	0.03	2.22	0.35	0.04
0	0	0.00	19.85	-4.47	0.35	0.54	0.41	25.70	0.52	0.46	0.35	0.36	0.34	8.6e-5	0.81	0.13	2.15	0.45	0.05
(A) Temporal coverage																			
5	10	0.55	19.80	-4.26	0.34	0.50	0.44	25.70	0.53	0.46	0.36	0.37	0.35	9.5e-5	0.72	0.12	2.16	0.46	0.05
5	25	1.33	19.93	-3.77	0.34	0.51	0.43	25.84	0.53	0.46	0.34	0.36	0.34	1.0e-4	0.77	0.13	2.16	0.46	0.05
5	50	2.68	19.98	-4.42	0.34	0.52	0.43	26.09	0.53	0.46	0.34	0.36	0.34	9.3e-5	0.66	0.13	2.16	0.47	0.05
5	75	4.19	19.91	-4.19	0.33	0.53	0.42	26.42	0.53	0.46	0.33	0.36	0.32	9.2e-5	1.17	0.13	2.16	0.46	0.05
5	100	5.33	19.80	-4.72	0.34	0.53	0.43	25.23	0.51	0.45	0.34	0.35	0.34	8.8e-5	1.00	0.13	2.16	0.47	0.05
(B) Noise magnitude																			
0.5	100	0.53	19.71	-4.03	0.34	0.53	0.41	25.64	0.52	0.46	0.35	0.36	0.35	9.6e-5	0.82	0.13	2.16	0.46	0.05
1	100	1.07	19.74	-4.22	0.35	0.52	0.39	25.20	0.52	0.46	0.35	0.37	0.35	1.1e-4	0.87	0.13	2.16	0.47	0.05
2	100	2.13	19.62	-4.17	0.33	0.53	0.39	25.35	0.52	0.46	0.35	0.36	0.34	9.5e-5	1.18	0.13	2.16	0.47	0.05
5	100	5.33	19.80	-4.72	0.34	0.53	0.43	25.23	0.51	0.45	0.34	0.35	0.34	8.8e-5	1.00	0.13	2.16	0.47	0.05
10	100	10.67	20.28	-4.32	0.34	0.51	0.48	26.59	0.54	0.46	0.29	0.33	0.29	7.7e-5	0.98	0.13	2.16	0.48	0.06

we voxelize the full scene at 2 cm resolution, extract a human-centric 32^3 occupancy volume aligned with the human heading direction, and encode it using a ViT following TRUMANS [15]. For a clear comparison, DNO optimization is disabled in both settings. As shown in Table S8, adopting a 3D voxel representation improves MPJPE (19.85 → 19.16 cm) and contact F1 (0.36 → 0.38), with a slight reduction in collision ratio ($Coll_{\%}$: 0.45 → 0.44). However, these gains come with a substantial computational overhead: the voxel variant requires approximately $4\times$ longer training due to the full-scene 3D I/O pipeline. We therefore find that the 2D BEV representation offers a more practical balance between geometric awareness and efficiency.

C.3 Robustness to Object Motion Perturbations

To evaluate robustness to imperfect object motion inputs, we perturb the object trajectories at inference time along two axes: (A) temporal coverage and (B) noise magnitude. Specifically, we perturb object translations using temporally smooth random 3D offsets generated by low-pass filtering of random noise, which better simulates realistic trajectory errors than frame-level jitter. To isolate the effect of noisy object motion conditioning, DNO optimization is disabled in this experiment. We additionally report the mean object translation error T_{obj} to quantify the actual deviation from the original object motion. In Table S9, σ denotes the noise magnitude and p the perturbed portion of the trajectory.

Temporal coverage: We gradually increase the portion of the trajectory affected by perturbation while fixing the noise magnitude to $\sigma = 5$ cm. LT-HOI demonstrates stable performance against temporal corruption. Even as we increase the corrupted portion of the trajectory up to 100%, the overall metrics remain comparable to the clean setting.

Noise magnitude: We vary the perturbation magnitude from $\sigma = 0.5$ cm to $\sigma = 10$ cm, applying noise to the entire trajectory. LT-HOI remains stable up to $\sigma = 5$ cm, maintaining a contact F1 of 0.35. Even under the perturbation of $\sigma = 10$ cm, while contact recall decreases, the precision remains unaffected. These results suggest that LT-HOI remains effective under moderate object motion errors, beyond the ideal ground-truth object motion setting.

C.4 DNO Analysis

What DNO optimizes and why collisions remain. The collision objective in DNO and the reported collision metric measure different aspects of collision. The DNO objective is designed to correct coarse navigation-level conflicts by optimizing the pelvis trajectory on the ground plane. In contrast, $Coll_{\%}$ reports the fraction of frames where any body vertex penetrates the scene geometry by more than 3 cm, over all body parts. Under this full-body metric, DNO reduces $Coll_{\%}$ to a level comparable to the ground-truth motion collision ratio of 0.35. In the GT motion, approximately 90% of the intersections belong to the hands (65.4%) and head (24.5%). Since the evaluation metric uses a strict 3 cm penetration threshold, it also captures genuine close-range contacts (*e.g.*, hand-object

Table S10: **Per-body-part collision analysis**

Part	(A) All frames			(B) Pelvis cleared by DNO		
	GT	w/o DNO	w/ DNO	GT	w/o DNO	w/ DNO
hand	0.271	0.278	0.250	0.373	0.450	0.348
arm	0.091	0.150	0.104	0.254	0.422	0.233
leg	0.056	0.157	0.031	0.178	0.627	0.102
torso	0.051	0.102	0.015	0.229	0.470	0.092
pelvis	0.044	0.106	0.024	0.167	0.429	0.084
head	0.027	0.032	0.028	0.025	0.069	0.034
foot	0.004	0.001	0.001	–	–	–
any (= $\text{Coll}_{\%}$)	0.357	0.454	0.314	0.623	0.997	0.519

Table S11: **Extending the DNO objective to all joints**

DNO objective	Subsequence												Transition			Sequence	
	Motion							Contact					Motion	Jerk		Scene	
	MPJPE↓	H_{feet} →	FS↓	R@3 ↑	FID↓	T_{goal} ↓	O_{goal} ↓	C_{prec} ↑	C_{rec} ↑	C_{F1} ↑	$C_{\%}$ →	P_{hand} ↓	FID↓	PJ↓	AUJ↓	$\text{Coll}_{\%}$ ↓	$P_{\max}$ ↓
pelvis only (ours)	19.10	-4.52	0.33	0.52	0.28	24.64	0.47	0.47	0.37	0.38	0.35	6.7e-5	1.07	0.13	2.12	0.33	0.05
per-joint	19.21	-4.04	0.34	0.52	0.19	25.46	0.50	0.46	0.32	0.34	0.31	6.0e-5	1.07	0.13	2.13	0.31	0.04

interactions), and the non-zero collision ratio of GT motion suggests that this metric has a non-negligible floor on contact-rich sequences. On a pelvis-level collision metric that directly evaluates DNO on the axis it optimizes, *i.e.*, the fraction of frames in which the predicted pelvis projected onto the ground plane lies in an occupied region of the BEV scene map, DNO reduces collisions from 0.09 to 0.01, an 89% reduction.

Per-body-part collision analysis. Table S10(A) reports the per-part penetration ratio over all test frames (989 subsequences, 95,703 frames), *i.e.*, the fraction of frames containing at least one vertex of each body part penetrating the scene by more than 3 cm, using the same 3D SDF query as $\text{Coll}_{\%}$; the union over all parts corresponds to $\text{Coll}_{\%}$. Before DNO, leg penetration is the main source of intersection (15.7% vs. 5.6% in GT), mainly near dining tables. Hand penetration remains close to GT, reflecting that the metric evaluates rigid meshes against a dilated voxel SDF, where even captured motion has non-zero collisions. Although DNO optimizes only the pelvis trajectory, it reduces collisions in other body parts through the kinematic chain: relative to w/o DNO, penetration decreases by 0.125 for legs, 0.046 for arms, and 0.028 for hands. After DNO, five of seven body parts have lower penetration than GT, while the arm (+0.013) and head (+0.001) are slightly higher. We further isolate frames where the pelvis initially intersected the 2D obstacle region but became collision-free after DNO (12,670 frames, 13.2% of the test set); DNO moved the pelvis outside the obstacle region in 95.7% of these frames. As shown in Table S10(B), these frames correspond to close-proximity interactions, where captured motion already has a high collision rate (62.3%). Without DNO, 99.7% of frames contain penetration, while DNO reduces penetration across all body parts toward GT levels, and the residual collisions remain comparable to captured motion. Contact F1 is unchanged, and hand joint error improves by 16.0 mm.

Extending the DNO objective to all joints. We implement a per-joint collision objective and compare it with our pelvis-level objective on the full test set (Table S11); both use 25 DDIM steps and 100 optimization steps. The per-joint objective further reduces collisions (0.33 → 0.31) and penetration depth (0.05 → 0.04), but this comes at the cost of interaction quality: contact F1 decreases from 0.38 to 0.34, mainly due to reduced recall (0.37 → 0.32), and T_{goal} increases from 24.64 to 25.46. Therefore, although per-joint optimization achieves stronger collision reduction, we retain the pelvis-level objective as it provides a better balance between collision avoidance and functional interaction quality.

DNO configuration and runtime. Table S12 reports the inference cost of DNO, averaged per subsequence on a single RTX A6000. Our default setting (75 DDIM and 150 optimization steps) takes 240 s. Reducing the DDIM steps to 25 and the optimization steps to 100 lowers this to 85.3 s and 59.3 s, respectively, and further skipping DNO for subsequences that are already collision-free yields 34.0 s, a 7× speedup. These reductions largely preserve quality: the fastest variant achieves MPJPE

Table S12: **DNO configuration and runtime**

Model	DDIM steps	Opt. steps	Subsequence													Transition			Sequence		Time/ subseq. (s)
			Motion							Contact						Motion	Jerk		Scene		
			MPJPE↓	H_{feet} →	FS↓	R@3 ↑	FID↓	T_{goal} ↓	O_{goal} ↓	C_{prec} ↑	C_{rec} ↑	C_{F1} ↑	$C_{\%}$ →	P_{hand} ↓		FID↓	PJ↓	AUJ↓	$Coll_{\%}$ ↓	P_{max} ↓	
GT	–	–	-	-3.30	0.29	-	-	-	-	-	-	-	0.49	1.7e-6		-	0.03	2.22	0.35	0.04	–
w/o DNO	75	–	19.85	-4.47	0.35	0.54	0.41	25.70	0.52	0.46	0.35	0.36	0.34	8.6e-5		0.81	0.13	2.15	0.45	0.05	∼1
DNO (ours)	75	150	19.16	-4.28	0.33	0.54	0.45	24.41	0.48	0.49	0.36	0.38	0.35	8.2e-5		1.12	0.13	2.16	0.32	0.05	240.0
DNO	25	150	19.12	-4.89	0.34	0.54	0.32	23.48	0.48	0.48	0.37	0.38	0.36	7.5e-5		0.53	0.14	2.12	0.32	0.05	85.3
DNO	25	100	19.10	-4.52	0.33	0.52	0.28	24.64	0.47	0.47	0.37	0.38	0.35	6.7e-5		1.07	0.13	2.12	0.33	0.05	59.3
DNO + skip	25	100	18.87	-5.13	0.35	0.53	0.39	23.45	0.43	0.47	0.35	0.37	0.35	8.2e-5		0.54	0.13	2.13	0.32	0.04	34.0

Table S13: **Results on the HIMO dataset [25]**

Method	MPJPE↓	H_{feet} →	FS↓	R@3 ↑	FID↓	C_{prec} ↑	C_{rec} ↑	C_{F1} ↑	$C_{\%}$ →	P_{hand} ↓	FID_{tr} ↓	PJ↓	AUJ↓
GT	–	1.71	0.02	–	–	–	–	–	0.65	6.5e-4	–	0.02	2.25
MDM	9.59	1.23	0.05	0.80	0.31	0.71	0.63	0.62	0.56	7.5e-4	0.46	0.35	2.55
CHOIS	9.53	1.27	0.05	0.77	0.36	0.71	0.60	0.60	0.54	6.7e-4	0.58	0.34	2.50
InterDiff	10.93	0.37	0.05	0.80	0.33	0.68	0.50	0.52	0.45	9.5e-3	0.48	0.18	2.28
DecHOI	10.80	0.70	0.07	0.75	0.57	0.70	0.56	0.58	0.52	6.4e-4	2.32	0.61	2.95
FlowMDM	8.95	1.43	0.04	0.76	0.60	0.73	0.63	0.62	0.55	9.5e-3	0.87	0.01	2.30
LT-HOI (ours)	9.05	0.56	0.05	0.80	0.26	0.74	0.68	0.67	0.61	6.1e-4	0.47	0.09	2.19

18.87 and T_{goal} 23.45 (vs. 19.16 and 24.41) with the same $Coll_{\%}$ of 0.32. Truncated backpropagation through only a subset of denoising steps [6] could further improve efficiency.

C.5 Results on the HIMO Dataset

We additionally evaluate all methods on the HIMO dataset [25], a large-scale dataset for full-body human-object interaction with multiple objects. All methods are trained on HIMO and evaluated on its held-out split. Since HIMO sequences consist of multiple interaction segments, we adapt the dataset into a long-term HOI composition benchmark by treating 120 frames as a single subsequence, as the official release does not provide timeline annotations. Sequences are processed at 30 FPS without temporal downsampling, yielding 389 continuous test sequences with an average duration of 9.8 s (up to 22.8 s). Since HIMO does not provide 3D scene geometry, we disable the scene embedding, do not apply DNO during inference, and omit navigation-related metrics ($Coll_{\%}$, P_{max}, T_{goal}, O_{goal}). For R@3 and FID, we use the official evaluator released with HIMO. As shown in Table S13, LT-HOI achieves the tied-best performance on R@3 and the best performance on FID, contact quality (precision, recall, and F1), and transition smoothness (AUJ), while FlowMDM yields the lowest MPJPE and PJ. We attribute this performance variation to the distinct characteristics of the two datasets. HIMO primarily focuses on short-term (9.8 s on average), object-centric interactions without scene boundaries, a setting where full-sequence non-autoregressive models may be effective. We thus view HIMO as a complementary benchmark: it shows that LT-HOI generalizes to shorter, unconstrained interactions, while ParaHome serves to evaluate the core challenges addressed by our formulation: long-horizon composition, transitions between multiple objects, and scene-aware navigation.

D Perceptual Study

To evaluate semantic correctness, visual realism, and physical plausibility beyond quantitative metrics, we conducted a perceptual study comparing LT-HOI with InterDiff and FlowMDM.

D.1 Setup

Participants. 25 participants who were not involved in this work took part as unpaid volunteers. All participants voluntarily agreed to participate in the study.

Stimuli. We randomly sampled 20 test sequences (fixed seed, no manual selection), each containing 3–4 consecutive interactions. All methods are rendered identically, so that only the motion differs.

Procedure. Each trial presents two videos side by side (LT-HOI vs. InterDiff or FlowMDM) in random left/right order, together with the task list. Participants answer three forced-choice questions

Table S14: **Perceptual study**

Criterion	Baseline	All	Clear	Subtle	Almost a guess
Semantic correctness	InterDiff	96.0%	96.9% (65)	97.6% (41)	89.5% (19)
	FlowMDM	70.4%	77.2% (57)	68.3% (41)	59.3% (27)
Visual realism	InterDiff	95.2%	97.7% (88)	89.7% (29)	87.5% (8)
	FlowMDM	80.0%	95.5% (66)	67.3% (49)	40.0% (10)
Physical plausibility	InterDiff	96.0%	100.0% (78)	90.0% (30)	88.2% (17)
	FlowMDM	78.4%	94.1% (68)	65.8% (38)	47.4% (19)
Average	InterDiff	95.7%	98.3% (231)	93.0% (100)	88.6% (44)
	FlowMDM	76.3%	89.5% (191)	67.2% (128)	51.8% (56)

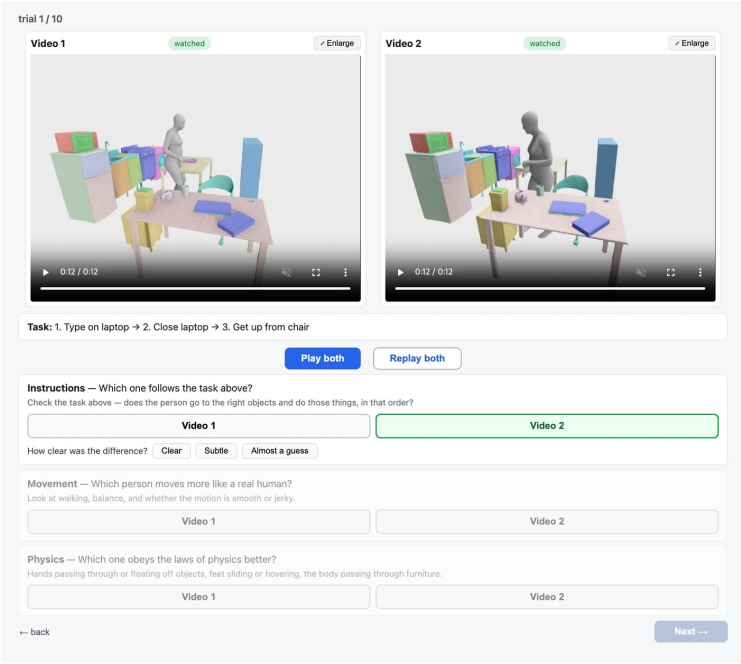


Figure S1: **Trial screen of the perceptual study**

on (1) *semantic correctness*, whether the motion follows the task list with the right objects and order; (2) *visual realism*, whether it moves like a real human; and (3) *physical plausibility*, whether it respects physics (*e.g.*, hand–object penetration, foot sliding, or passing through furniture). After each choice, they rate their confidence as *clear*, *subtle*, or *almost a guess*. Each participant judged 10 pairs, five against each baseline, giving 750 judgments in total (125 per criterion and baseline). The full instructions and the trial screen are provided in Sec. D.3.

D.2 Results

Table S14 reports the preference rate for LT-HOI, overall and split by confidence (number of judgments in parentheses). LT-HOI is preferred over both baselines across all three criteria, in 95.2–96.0% of the comparisons against InterDiff and 70.4–80.0% against FlowMDM. Against FlowMDM, the preference reaches 89.5% for clear judgments and decreases for subtle (67.2%) and near-guess (51.8%) ones, indicating that the preference for LT-HOI is strongest when participants perceive a clear difference. Against InterDiff, LT-HOI is preferred at all confidence levels (98.3% / 93.0% / 88.6%).

D.3 Instructions to Participants

Before the study, participants were shown the following instructions.

> **Video comparison study.**
> **What this is.** You'll watch short animations of a person doing everyday tasks in a room. Each screen shows two versions of the same task, made by two different systems. You pick which one is better in three ways. About 12 minutes, 10 screens.
> **What to look for.**
>
> - Does the person do the listed actions, in order, with the right objects?
> - Does the person move naturally, without jerky or sudden changes?
> - Do hands hold objects, feet stay on the floor, and the body avoid furniture?
>
> **What to ignore.** Colors, the plain grey body, blocky furniture, sharpness, lighting, camera angle. Both videos are rendered the same way—only the movement differs.
> **Rules.** Watch both videos fully; the buttons unlock once they've played. You must pick one every time—there's no "they're the same" option.

Each trial screen (Fig. S1) shows the two videos side by side together with the task list and asks the following three questions.

> **Instructions**—Which one follows the task above? Check the task above—does the person go to the right objects and do those things, in that order?
> **Movement**—Which person moves more like a real human? Look at walking, balance, and whether the motion is smooth or jerky.
> **Physics**—Which one obeys the laws of physics better? Hands passing through or floating off objects, feet sliding or hovering, the body passing through furniture.

After each choice, participants are asked "How clear was the difference?" and select one of *Clear*, *Subtle*, or *Almost a guess*.

E Discussions

Limitations. LT-HOI takes object motions as input, treating them as a high-level spatio-temporal plan for human motion synthesis. In practice, such motions can be obtained from planning [2, 45] or object pose estimation [39, 40], and LT-HOI remains effective under moderate object motion errors (Sec. C.3); jointly generating object and human motions from task-level goals is a promising future direction. In addition, the scene is represented as a 2D ground-plane map and the DNO objective acts on the pelvis trajectory, so fine-grained full-body collisions are not directly penalized. A per-joint objective reduces collisions further but degrades contact quality (Sec. C.4), and a 3D voxel representation improves motion quality at about $4\times$ the training cost (Sec. C.2). Finally, DNO requires test-time optimization, which takes 240 s per subsequence in our default setting and can be reduced to 34.0 s with fewer steps (Sec. C.4).

Failure cases. When the base model produces severe scene penetrations (*e.g.*, passing through furniture), DNO may introduce minor floating or sliding artifacts while resolving them. These artifacts are not systematic: across the test set, foot-related metrics remain comparable after DNO, motion smoothness is preserved, and MPJPE improves (Table 3).

Broader impact. By generating realistic and temporally coherent long-term human–object interactions, LT-HOI can benefit embodied AI, AR/VR content creation, and digital human simulation. Since it does not explicitly model full physics or safety constraints, generated motions require additional verification before deployment on real-world robots.

F Resources and License

Compute resources. We train and evaluate LT-HOI on a single NVIDIA RTX A6000 GPU; training for 200,000 iterations takes about 18.5 hours and requires about 29 GB of GPU memory. We implement LT-HOI in PyTorch.

Licenses of existing assets. We use the ParaHome dataset [18], which is released under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License (CC BY-NC-SA 4.0). For body and hand parameterization, we use the SMPL-X model [26], which is released under the Max Planck non-commercial scientific research license. For the experiments in Sec. C.5, we use the HIMO dataset [25], which is released under the HIMO license for non-commercial scientific research. We encode text with the pretrained CLIP model [33], which is released under the MIT License. We also build upon publicly available code releases of MDM [38], CHOIS [20], DecHOI [16], InterDiff [43], and FlowMDM [3] for our baselines, each used under the licenses provided in their respective repositories.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our main claims presented in the abstract and introduction accurately reflect the paper's contributions and scope.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations of our work in the supplementary discussion (Section E).

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: Our paper does not include theoretical results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide the model architecture, training procedure, and implementation details in Section 3.3, Section 3.4, and Section 4.3, with additional details in the supplementary (Section C).

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We release our code as open source on the project page (https://gongda0e.github.io/LTHOI/). The ParaHome dataset used for training and evaluation is publicly available.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: We specify the training and test details in Section 4.3 and the supplementary material.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: We do not report error bars due to the limited computational resources required for repeated training of long-horizon human-object interaction models.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We report the GPU type used for training and inference in the supplementary material (Section F).

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: We follow the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss the broader impact of our work in the supplementary discussion (Section E).

    Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: We do not release data or models that pose a high risk for misuse.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite all the papers that are relevant to the code, models, and datasets used in our work, and list the license of each existing asset in the supplementary material (Section F).

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We release our code as open source on the project page (`https://gongda0e.github.io/LTHOI/`).

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: The protocol of the perceptual study is described in Sec. D, and the full instructions given to participants are provided in Sec. D.3. Participants were unpaid volunteers.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [No]

    Justification: We did not obtain IRB approval for the perceptual study (Sec. D). The study involved voluntary participants watching generated human motion videos and providing forced-choice preference judgments. All participants were informed about the study and voluntarily agreed to participate. The study did not involve collecting sensitive personal information.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.