IntentMotion: Learning Intent-Aware Human Motion from Language in 3D Scene

Wenfeng Song[1]*, Shi Zheng[1], Xinyu Zhang[2], Xingliang Jin[3], Aimin Hao[2], Fei Hou[4,5], Xia Hou[1], Shuai Li[2,6]†

[1]College of Computer Science, Beijing Information Science and Technology University
[2]State Key Laboratory of Virtual Reality Technology and Systems, Beihang University
[3]School of Computer Science and Technology, East China Normal University
[4]Key Laboratory of System Software (CAS), State Key Laboratory of Computer Science, Institute of Software, Chinese Academy of Sciences, China
[5]University of Chinese Academy of Sciences, China
[6]Zhongguancun Laboratory, China
{songwenfenga, zhengshi1109, xingliangjin276}@gmail.com, {zhangxinyu1, lishuai, ham}@buaa.edu.cn, houfei@ios.ac.cn, houxia@bistu.edu.cn

Abstract

Generating human motion in complex 3D scenes from text is a challenging task with broad applications. However, existing methods often overlook realistic physical contact, resulting in visually plausible but physically unrealistic motion, e.g., penetration. To alleviate this, we propose IntentMotion, a novel framework that generates human motion in 3D scenes from natural language instructions by explicitly modeling intent. We first introduce the Intention-Guided Contact Field (IGCF). This differentiable voxel-based contact region representation explicitly aligns parsed language roles with spatial contact regions through a hierarchical attention mechanism. IGCF is jointly trained with a diffusion-based motion generator, allowing contact predictions to adapt dynamically through gradient feedback. To improve the controllability and physics-aware motion, we further propose an Intention-Aware Diffusion Model (IADM), which decouples the high-level semantic planning from the low-level contact refinement in a coarse-to-fine process. The optimized contact cues are utilized to guide the synthesis of a coarse trajectory, followed by refining detailed pose sequences under IGCF supervision. Experiments on the HUMANISE and LINGO datasets demonstrate that our IntentMotion outperforms recent baselines in contact accuracy, semantic alignment, and generalization to unseen scenes.

Code — https://github.com/zhengshi119/IntentMotion

1 Introduction

Human-scene interaction (HSI) focuses on generating natural and diverse motion for virtual humans in 3D environments, reducing the need for manual animation or expert design. This ability is crucial for applications such as gaming, virtual reality, and film. Given a 3D scene and a text description, recent methods (Wang et al. 2022b, 2024; Cen et al. 2024; Jiang et al. 2024a; Yi et al. 2025; Milacski et al. 2025) aim to produce motion that are both semantically relevant and physically plausible within the environment.

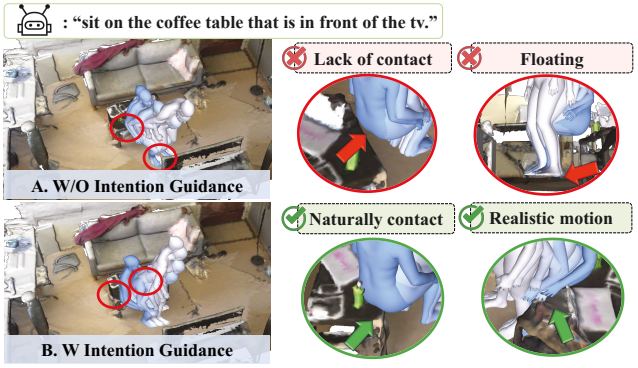

Figure 1: Impact of Intention-Guided Contact Refinement. (A) Lacking intention guidance may result in floating or detachment from the target objects. (B) Our IntentMotion integrates intention guidance, eliminating these issues and ensuring better interaction with the target objects.

Recent advances in human-scene motion generation often use scene-conditioned diffusion models with post hoc refinements (Huang et al. 2023; Chen et al. 2025; Wang et al. 2021) or Signed Distance Function (SDF) (Zhao et al. 2023; Xing, Mao, and Liu 2024; Chen et al. 2025) to reduce physical artifacts like penetration. However, these low-level cues are passive and lack structure, learnability, and awareness of semantic intent. They cannot answer a key question: where and how should contact occur to fulfill a language instruction. Although language provides high-level goals, it omits spatial specifics, making it difficult to ground semantics into physically coherent behavior. This gap stems from the lack of a structured and differentiable contact representation that aligns intent with localized physical interaction. For instance, when prompted to "*stand up from a table that is in front of the TV*", existing methods (Wang et al. 2022b; Cen et al. 2024; Wang et al. 2024) frequently result in physically

*Corresponding author.
†Corresponding author.

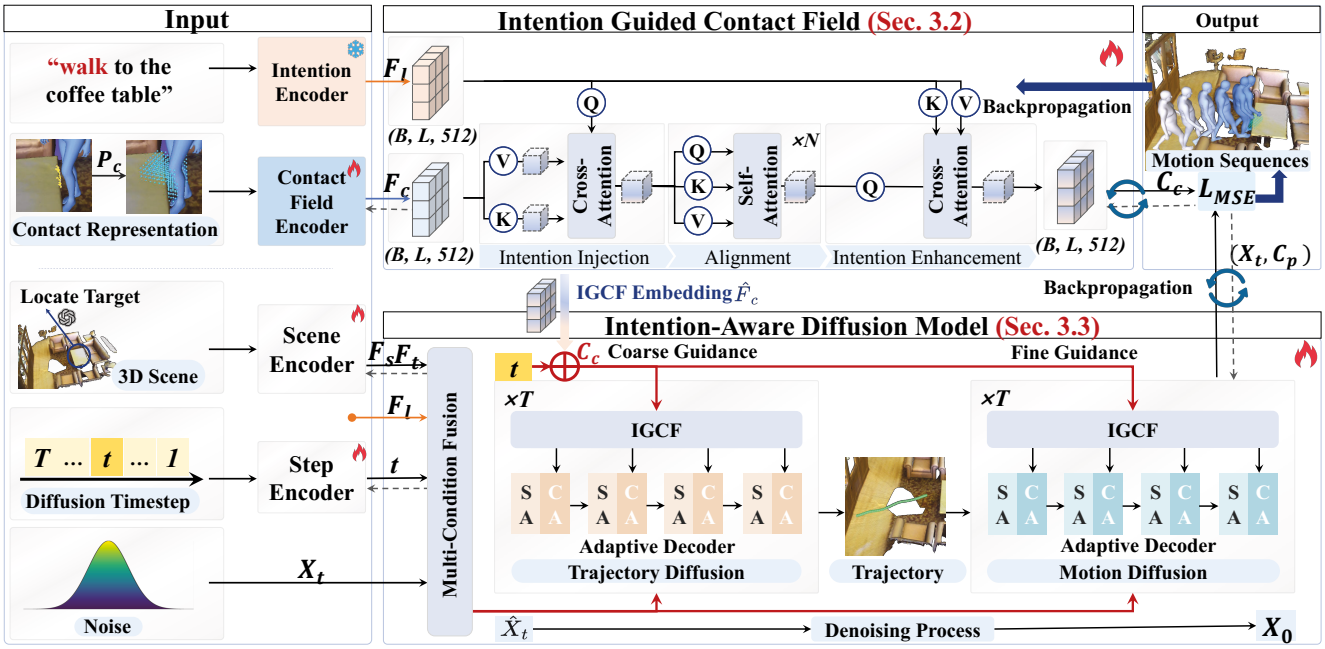

Figure 2: Method overview. Our IntentMotion has two components: (1) *Intention Guided Contact Field (IGCF)* (Sec. 3.2) that models human-scene contact as a structured spatial prior guided by user instructions. (2) *Intention-Aware Diffusion Model (IADM)* (Sec. 3.3) that separates primary conditions and contact constraints via a two-stage generation process, progressively enhancing motion realism and contact plausibility via our Adaptive decoder. During training, gradients from motion loss are backpropagated into IGCF, enabling learning between contact inference and motion refinement.

implausible motion, such as foot-floor floating or missing pelvis-table support (see Fig. 1-A).

This limitation is further compounded by the absence of a generative framework that unifies semantic intent, scenario context, and physical interaction within a single model. Although recent methods (Wang et al. 2022b; Cen et al. 2024; Wang et al. 2024; Jiang et al. 2024a) incorporate both language and scene geometry as conditioning signals for motion generation, they typically treat these signals independently and overlook contact as a structured and learnable constraint. As a result, motion may follow the intended semantics but fail to reflect the physical affordances of the environment. Without explicitly grounding linguistic intent in spatial contact, models struggle to produce coherent and context-aware human-scene interactions.

To tackle these limitations, we propose an **Intention-Guided Contact Field (IGCF)** to explicitly determine *where* and *how* contact should occur during human-scene interaction. Unlike prior methods (Song et al. 2024b; Bhatnagar et al. 2022) that rely on passive geometric proximity, IGCF constructs a fully differentiable voxel-space field by detecting body-object contact points via signed distance functions and encoding their spatial distribution and surface normals. Beyond geometry, IGCF introduces a novel semantic-contact fusion mechanism: it incorporates parsed language roles (action, target, anchor) into the contact field through a hierarchical attention mechanism (Shridhar, Manuelli, and Fox 2023; Jaegle et al. 2022, 2021), enabling structured alignment between linguistic intent and physical constraints. This transforms contact modeling from post hoc geometric filtering into a learnable, intent-driven interaction prior (see Fig. 1-B).

Building on IGCF, we introduce an **Intention-Aware Diffusion Model (IADM)** that separates semantic conditions (text and scene) from physical cues (contact) via a two-stage generation framework. The first stage synthesizes a coarse trajectory guided by global intent, which is then refined under IGCF supervision to enforce contact realism. In contrast to prior two-stage models (Pi et al. 2023; Wang et al. 2022a; Cen et al. 2024) that emphasize denoising stability, IADM is semantically driven: the contact-aware trajectory acts as an interaction scaffold, grounding motion refinement in a structured physical context. This hierarchical conditioning improves controllability, realism, and generalization across diverse interaction scenarios.

The key contributions of IntentMotion include:

- We propose a unified framework that jointly models linguistic intent, contact reasoning, and motion generation within a tightly integrated architecture. Our method enables gradient-based interaction across modalities, allowing contact representations to adapt dynamically based on motion quality and semantic alignment.
- We introduce the **Intention-Guided Contact Field**, a differentiable voxel-based representation that combines contact geometry with language semantic roles. IGCF is trained jointly with the motion generator and refined using feedback from motion loss, allowing contact predic-

tion to adapt to both intent and physical constraints.

- We design a two-stage **Intention-Aware Diffusion Model** that explicitly separates high-level semantic planning from low-level contact refinement. This coarse-to-fine architecture allows global intent to guide motion structure while enabling contact-aware local correction, improving both semantic consistency and physical plausibility.

2 Related Work

Human Motion Diffusion Models. Diffusion models (Sohl-Dickstein et al. 2015) generate human motion by iteratively denoising random noise into coherent sequences, and have shown strong performance in text-to-motion generation. Key methods include MDM (Tevet et al. 2023), MotionDiffuse (Zhang et al. 2024), MLD (Chen et al. 2023), GMD (Karunratanakul et al. 2023), FineMoGen (Zhang et al. 2023b), StableMoFusion (Huang et al. 2024), MotionLCM (Dai et al. 2024), ReMoDiffuse (Zhang et al. 2023a), PhysDiff (Yuan et al. 2023), and GraphMotion (Jin et al. 2023). For scene-conditioned motion generation, recent efforts include Cen et al.(Cen et al. 2024), Afford-Motion(Wang et al. 2024), LINGO (Jiang et al. 2024a), TRUMANS (Jiang et al. 2024b), CLoSD (Tevet et al. 2025) and TeSMo (Yi et al. 2025). While these models incorporate scene or affordance cues, capturing the fine-grained interplay between language intent, spatial layout, and physical contact remains challenging due to modality complexity and limited data. To address this, we propose an adaptive diffusion model that enhances the denoising process by explicitly incorporating structured contact and intent guidance.

Contact Modeling of Motion Generation. Contact modeling has been extensively studied in human-object interaction (HOI) tasks, such as BEHAVE (Bhatnagar et al. 2022), HOI-Diff (Peng et al. 2025), HOIAnimator (Song et al. 2024b), CG-HOI (Diller and Dai 2024), and GenHOI (Li et al. 2025). while these methods focus on isolated object contact, they offer valuable insights for scene-level interaction. In HSI tasks, PLACE (Zhang et al. 2020a), POSA (Hassan et al. 2021), DIMOS (Zhao et al. 2023), RICH-CAT (Ma et al. 2024), HIS-GPT (Zhao et al. 2025), Afford-Motion (Wang et al. 2024), and other works (Mao et al. 2022) introduce contact modeling into scene-aware motion. However, many of them rely on global contact maps or heuristic estimators (Zhang et al. 2020a; Wang et al. 2024; Mao et al. 2022), which lack part-level precision and adaptability to scene and language variations. Meanwhile, recent scene-conditioned motion generation methods (Wang et al. 2022b; Cen et al. 2024; Chen et al. 2025; Yi et al. 2025) often overlook explicit contact reasoning and semantic grounding, leading to unrealistic behaviors. To bridge this gap, we propose a **learnable contact field** that encodes geometry-aware contact cues conditioned on parsed language roles.

3 Method

Method Overview. We aim to generate 3D human motion that aligns with language instructions and respect physical contact within the scene. As illustrated in Fig. 2, we propose **IntentMotion**, a framework that combines semantic and physical reasoning through two components: (1) **Intention-Guided Contact Field (IGCF)** (Sec.3.2), a learnable voxel representation that links contact prediction to scene geometry and language semantics; and (2) **Intention-Aware Diffusion Model (IADM)** (Sec.3.3), a two-stage generator that first produces coarse motion and then refines them via IGCF for accurate, contact-aware synthesis.

3.1 IntentMotion Preliminaries

The language instruction follows a structured format (Wang et al. 2022b; Cen et al. 2024) including an action verb, a referenced object, and optionally a spatial anchor. This facilitates semantic role parsing for downstream alignment. The motion is represented as a sequence of SMPL-X (Pavlakos et al. 2019) body parameters of length L, including global translation $r_t \in \mathbb{R}^3$, orientation $\gamma_t \in \mathbb{R}^6$, body poses $\theta_t \in \mathbb{R}^{J\times 6}$, and shape parameters $\beta \in \mathbb{R}^{10}$. The scene is represented as a point cloud and encoded using a simple MLP-based scene encoder, producing scene feature F_s. To ground the 3D scene based on the text command (Cen et al. 2024), we first identify the target object and its spatial context, resulting in target feature F_t. Given input text V and scene S, our model predicts a motion sequence $\{X\}_{l=1}^{L} = \{(r_t, \gamma_t, \theta_t)\}_{l=1}^{L}$ that aligns with linguistic intent.

Diffusion model (Ho, Jain, and Abbeel 2020; Sohl-Dickstein et al. 2015; Song and Ermon 2019) generates data via iterative denoising over a Markov chain $\{x_t\}_{t=0}^{T}$ governed by $q(x_t|x_0)$, where $x_0 \sim q(x_0)$ is the original data, and x_t is the noised data at noising step t (encoded via a MLP). The forward process adds noise as:

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\alpha_t}x_0, (1-\alpha_t)\mathbf{I}), \tag{1}$$

with α_t following a decreasing schedule. When α_t becomes small, $x_T \sim \mathcal{N}(0, \mathbf{I})$, allowing the model to generate samples by reversing the process from noise to data.

3.2 Intention-Guided Contact Field

Inspired by prior contact modeling (Starke et al. 2019; Zhang et al. 2021; Cen et al. 2024; Jiang et al. 2024a), we introduce the **Intention-Guided Contact Field (IGCF)**. Unlike static contact maps or post-hoc proximity cues (Bhatnagar et al. 2022; Mao et al. 2022; Wang et al. 2024), our IGCF is a structured and differentiable representation that enables intent-aware contact reasoning by fusing spatial contact patterns with high-level language semantics. IGCF resolves the gap between textual intent and localized physical interaction by jointly modeling *where* contact occurs, *how* it happens. Importantly, IGCF is a parameterized module with learnable weights, trained jointly with the motion denoiser. Through end-to-end backpropagation, the contact field is continuously refined based on the downstream motion loss.

Voxel-Based Contact Representation. Voxel-based field enables dense region-level reasoning and serves as an effective bridge between unstructured 3D geometry and structured semantic attention, which is critical for contact-aware human-scene interaction synthesis. Extended from (Cen

et al. 2024; Starke et al. 2019), we take a step further to adopt a voxel-based contact representation for integrating geometric contact cues with language intent. Specifically, we construct a contact-aware voxel grid by computing the signed distance between the posed SMPL-X mesh M and the target point set P_t, and identifying contact points within a threshold $\epsilon = 0.1$:

$$P_c = \{P_t \mid |\delta(P_t, M)| < \epsilon\}, \tag{2}$$

where $\delta(\cdot)$ denotes the signed distance function (SDF). The resulting contact set is encoded into a voxel-based contact representation as follows:

$$F_c = \text{ConEncoder}(o_c, n_v, P_o), \tag{3}$$

where $\text{ConEncoder}(\cdot)$ denotes the Contact Field Encoder, implemented as a lightweight feedforward network composed of two linear layers with SiLU (Elfwing, Uchibe, and Doya 2018) activations. The input features include the contact occupancy score o_c, the local surface normal n_v, and the contact centroid $P_o = \frac{1}{|P_c|}\sum_{p\in P_c} p$. The output $F_c \in \mathbb{R}^{B\times L\times D}$ represents the encoded contact features.

Voxelized formulation provides part-aware contact signals with spatial continuity, enabling gradient backpropagation from motion supervision. It significantly improves contact localization and physical plausibility compared to static SDFs or global heatmaps, and plays a key role in grounding motion generation to geometry-aware contact semantics.

IGCF Construction. To condition contact reasoning on semantic intent, we employ a pretrained CLIP encoder (Radford et al. 2021) as the Intention Encoder to process the language instructions V, yielding language features $F_l \in \mathbb{R}^{B\times L\times D}$. These are fused with the voxelized contact features F_c through an attention mechanism that enables structured guidance from semantic roles to physical contact cues.

(1) Intention Injection. Language features are first used to guide contact understanding through cross-attention:

$$F_c^{(0)} = \underbrace{\varphi\left(\eta\left(\text{CA}(\underbrace{\tilde{F}_l}_{\eta(F_l)}, \underbrace{\tilde{F}_c}_{\eta(F_c)}, \tilde{F}_c)\right) \oplus F_l\right)}_{\text{Cross-modal injection (language} \rightarrow \text{contact)}}, \tag{4}$$

where $\tilde{F}_l = \eta(F_l)$ and $\tilde{F}_c = \eta(F_c)$ respectively denote the normalized language and contact features. $\text{CA}(\cdot)$ is the standard cross-attention mechanism. $\varphi(\cdot)$ is a position-wise feedforward network. $\oplus$ denotes element-wise addition for residual connection.

(2) Multi-Layer Alignment. We further refine $F_c^{(0)}$ via a network with N layers of self-attention, which is to capture long-range dependencies across contact-space regions:

$$F_c^{(i)} = \underbrace{\varphi\left(\text{SA}(\tilde{F}_c^{(i-1)}) \oplus \tilde{F}_c^{(i-1)}\right)}_{\text{Self-attention refinement}} \oplus F_c^{(i-1)}, \quad i \in [1, N], \tag{5}$$

where $\text{SA}(\cdot)$ denotes self-attention mechanism. This design aligns semantic and spatial cues across layers, allowing the model to learn structured and coherent contact intent.

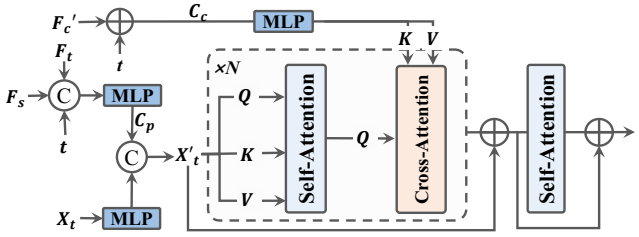

Figure 3: Adapt-Dec of IADM. Adapt-Dec consists of N stacked blocks, each containing a self-attention module for modeling intra-sequence motion structure and a cross-attention module for contact-guided refinement.

(3) Intention Enhancement. Finally, $F_c^{(N)}$ re-attend to language tokens to further reinforce semantic grounding:

$$\hat{F}_c = \text{CA}(\tilde{F}_c^{(N)}, \tilde{F}_l, \tilde{F}_l) \oplus F_c, \tag{6}$$

where the embedding $\hat{F}_c \in \mathbb{R}^{B\times L\times D}$ serves as a unified representation that captures both the spatial likelihood and the semantic necessity of contact. The entire IGCF pipeline is fully differentiable, allowing gradient-based learning of spatial contact representation F_c under semantic guidance F_l. This enables the contact field $\hat{F}_c$ to be directly optimized via interaction-specific objectives, thereby tightly coupling semantic intent with physical grounding during training.

This hierarchical attention mechanism introduces two key innovations: (1) a **layer-wise attention design** for progressive intent alignment; and (2) a fully differentiable formulation that supports contact-aware learning under semantic constraints. Together, these elements enable the model to generate motion that is both physically grounded and semantically faithful, particularly in contact-sensitive scenarios such as sitting or leaning objects.

3.3 Intention-Aware Diffusion Model

To address the challenge of generating semantically coherent and physically plausible human motion under complex multi-modal conditions, we introduce an **Intention-Aware Diffusion Model**. At each timestep, the noisy input is concatenated with multi-source condition features, including language semantics, target-object grounding, scene geometry, and contact constraints. Our model applies a modality-aware integration strategy: while text, target, and scene features are globally fused, the contact condition is selectively injected via designated cross-attention layers.

Multi-Condition Fusion and Guidance. Motivated by the structured guidance scheme in (Song et al. 2024a), we decompose conditions into two streams: (1) the **contact condition** C_c, derived from the IGCF embedding $\hat{F}_c$ fused with the diffusion timestep t, and (2) the **primary condition** C_p, which combines language feature F_l, target object feature F_t, and scene feature F_s, also conditioned on t. This separation allows the model to apply differentiated attention across semantic and physical modalities.

The denoising process is split into coarse-to-fine stages: we first synthesize a coarse trajectory capturing global motion structure, then refine it into detailed pose sequences.

Methods	Scene-conditional			Action-conditional			Pure motion quality	
	goal dist. ↓	contact↑	non-collision↑	accuracy↑	diversity→	multimodality→	FID↓	APD↑
Real	$0.014^{\pm 0.001}$	$98.91^{\pm 0.004}$	$99.98^{\pm 0.000}$	$0.991^{\pm 0.003}$	$4.793^{\pm 0.092}$	$2.168^{\pm 0.041}$	$0.000^{\pm 0.000}$	-
HUAMNISE	$0.315^{\pm 0.013}$	$79.70^{\pm 0.009}$	$99.93^{\pm 0.001}$	$0.895^{\pm 0.003}$	$4.272^{\pm 0.039}$	$2.754^{\pm 0.079}$	$1.103^{\pm 0.031}$	$\mathbf{5.525}^{\pm 0.038}$
Afford-Motion	$0.146^{\pm 0.007}$	$93.92^{\pm 0.008}$	$99.90^{\pm 0.002}$	$0.920^{\pm 0.001}$	$4.462^{\pm 0.005}$	$2.430^{\pm 0.020}$	$0.569^{\pm 0.008}$	$4.879^{\pm 0.026}$
Ours (w/o Target)	$\mathbf{0.069}^{\pm 0.010}$	$\mathbf{95.63}^{\pm 0.007}$	$\mathbf{99.95}^{\pm 0.001}$	$\mathbf{0.970}^{\pm 0.001}$	$\mathbf{4.789}^{\pm 0.021}$	$\mathbf{2.205}^{\pm 0.043}$	$\mathbf{0.255}^{\pm 0.002}$	$4.860^{\pm 0.035}$
MDM*	$0.074^{\pm 0.011}$	$94.87^{\pm 0.008}$	$99.82^{\pm 0.002}$	$0.904^{\pm 0.001}$	$4.253^{\pm 0.028}$	$2.022^{\pm 0.069}$	$1.084^{\pm 0.016}$	$5.330^{\pm 0.041}$
Cen et al.	$0.130^{\pm 0.012}$	$89.44^{\pm 0.011}$	$99.95^{\pm 0.002}$	$0.937^{\pm 0.001}$	$\mathbf{4.768}^{\pm 0.036}$	$2.403^{\pm 0.087}$	$0.370^{\pm 0.011}$	$\mathbf{5.334}^{\pm 0.040}$
Ours (w/ Target)	$\mathbf{0.067}^{\pm 0.011}$	$\mathbf{95.30}^{\pm 0.007}$	$\mathbf{99.96}^{\pm 0.002}$	$\mathbf{0.976}^{\pm 0.000}$	$4.569^{\pm 0.009}$	$\mathbf{2.176}^{\pm 0.048}$	$\mathbf{0.163}^{\pm 0.001}$	$4.942^{\pm 0.034}$

Table 1: Quantitative Evaluation on HUMANISE dataset. The different interpretations of Methods can be found in Sec. 4.2. To ensure a fair comparison, we conducted 10 experiments, $x^{\pm y}$ denotes that x represents the average value of the metric, while y corresponds to the 95% confidence interval around this mean. ↑ means higher is better and ↓ means lower is better. → means closer to the real data is better. **Bold** highlights the best results. Underline indicates the second best.

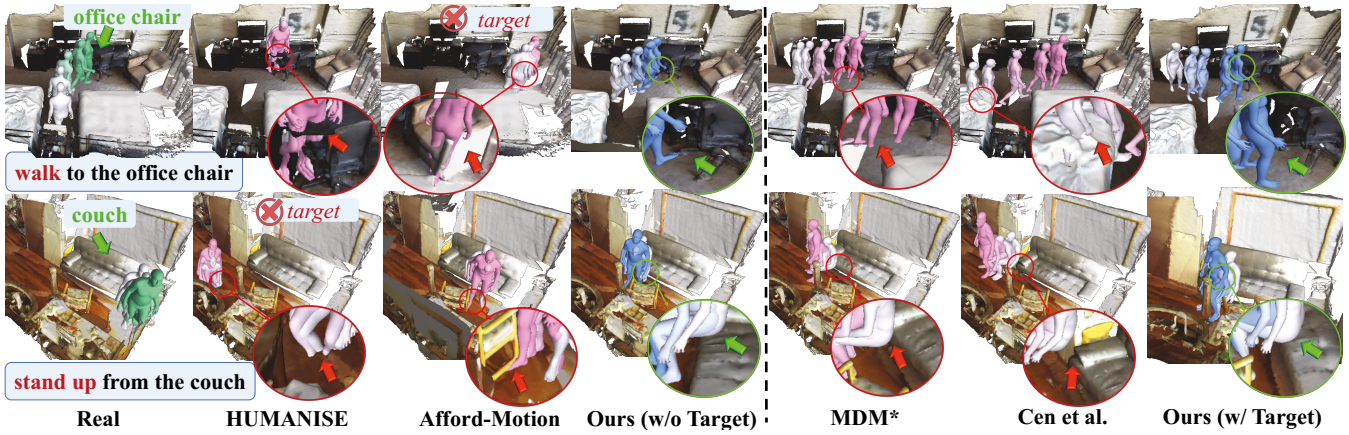

Figure 4: Visualization Comparison on HUMANISE dataset. Leveraging the IGCF module and Adapt-Dec, our method generates motion that accurately capture the fine-grained contact patterns observed in real interaction.

This two-stage design helps mitigate semantic drift and improves contact precision. The conditioning and input process is defined as:

$$\begin{aligned} C_c &= \phi(t + F_c), \\ C_p &= \phi(\mathcal{F}_{cat}(t, F_l, F_t, F_s)), \\ \hat{X}_t &= \mathcal{F}_{cat}(C_p, \phi(X_t)), \end{aligned} \tag{7}$$

where $\phi(\cdot)$ is a lightweight MLP, and $\mathcal{F}_{cat}(\cdot)$ denotes feature concatenation. The same formulation applies to both trajectory and motion diffusion stages, with the key difference being that motion generation further conditions on the coarse trajectory from the first stage.

Denoiser for IADM. Inspired by Afford-Motion (Wang et al. 2024), we propose the **Adaptive Decoder (Adapt-Dec)**, a contact-aware diffusion backbone designed to integrate semantic intent, scene geometry, and physical contact constraints in a unified manner. Adapt-Dec dynamically fuses global semantics and local contact cues at each decoding layer through dual attention pathways. As shown in Fig. 3, Adapt-Dec consists of N stacked blocks (e.g., 4), each containing a Self-Attention (SA) module for modeling intra-sequence motion structure and a Cross-Attention (CA) module for condition-guided refinement. The attention mechanism in each block is defined as:

$$\begin{aligned} &\text{Attn}(\underset{\Uparrow}{Q, K, V}) = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V, \\ &\text{where} \begin{cases} Q = \hat{X}_t, K = \hat{X}_t, V = \hat{X}_t & \text{(if Attn is SA)} \\ Q = \hat{X}_t, K = C_c, V = C_c & \text{(if Attn is CA)} \end{cases} \end{aligned} \tag{8}$$

We directly regress the clean motion signal X_0 rather than the residual noise, enabling more stable optimization and faster convergence:

$$\mathcal{L}_{\text{MSE}} = \mathbb{E}_{X_0,t}\left[|X_0 - G_\alpha(X_t \mid C_p, C_c)|_2^2\right], \tag{9}$$

where $t \sim U[1, T]$ and $X_0 \sim q(X_0)$. G_α denotes the generative model. This $\mathcal{L}_{MSE}$ ensures stable training while preserving the theoretical underpinnings of the diffusion framework. Importantly, since IGCF is directly involved in the CA computation, it receives gradient feedback from the motion decoder, allowing contact predictions to be refined jointly with motion quality.

Backpropagation of IGCF. While the IGCF provides semantically grounded contact priors $\hat{F}_c$ to guide pose refine-

Variants.	$\text{Pene}_{\text{rate}} \downarrow$	$\text{Pene}_{\text{mean}} \downarrow$	$\text{Pene}_{\text{max}} \downarrow$	FID$\downarrow$
Real	$0.352^{\pm 0.018}$	$0.786^{\pm 0.014}$	$0.926^{\pm 0.012}$	$0.000^{\pm 0.000}$
Cen et al.*	$0.261^{\pm 0.013}$	$0.517^{\pm 0.014}$	$0.764^{\pm 0.013}$	$1.545^{\pm 0.022}$
Ours	$\mathbf{0.119}^{\pm 0.012}$	$\mathbf{0.382}^{\pm 0.017}$	$\mathbf{0.631}^{\pm 0.014}$	$\mathbf{1.254}^{\pm 0.009}$

Table 2: Quantitative results on LINGO dataset. Our IGCF module achieves the best motion quality.

ment, we further establish a training-time feedback mechanism by **backpropagating the denoising loss from the diffusion-based motion generator into the IGCF module**. This gradient flow enables the C_c to adapt dynamically based on motion prediction errors, allowing it to evolve throughout training. Rather than serving as a static prior, the IGCF is jointly optimized to enhance both motion plausibility and contact accuracy. This bidirectional dependency, where contact representations guide motion generation and motion supervision is in turn used to refine contact inference, forms a closed learning loop within our framework.

4 Experiments

This section presents both quantitative and qualitative experimental results, and ablation studies.

4.1 Evaluation Metrics

To ensure fair comparisons, we follow the evaluation protocols of PSI (Zhang et al. 2020b), HUMANISE (Wang et al. 2022b), Cen et al. (Cen et al. 2024), Afford-Motion(Wang et al. 2024), and LINGO (Jiang et al. 2024a). We evaluate the motion in three aspects: (1) **Scene-conditional Motion Quality.** For the HUMANISE dataset, we evaluate scene alignment via body-to-goal distance (goal dist.) and motion plausibility using contact and non-collision scores. For the LINGO dataset, we measure scene penetration via penetration rate ($\text{Pene}_{\text{rate}}$), average penetration distance ($\text{Pene}_{\text{mean}}$), and maximum penetration distance (Pene_{max}). (2) **Action-conditional Motion Quality.** To measure how the generated motion is aligned with the text, we assess the action recognition performance, evaluating the accuracy, diversity, and multimodality of the results. (3) **Pure Motion Quality.** We evaluate the realism of generated motion using the Frechet Inception Distance (FID), while Average Pairwise Distance (APD) is used to assess motion diversity across all motion pairs. We further conduct a user study to assess motion quality, scene consistency, and semantic alignment from a human perception perspective (details in supplementary).

4.2 Comparisons With State-of-the-Art Methods

Baselines. We conduct our experiments mainly on the HUMANISE dataset and compare our method with four baselines: (1) HUMANISE (Wang et al. 2022b) uses a conditional VAE with auxiliary tasks for object center regression. (2) Afford-Motion (Wang et al. 2024) employs a two-stage framework, first predicting affordance maps, then generating human motion. The released HUMANISE and Afford-Motion models are used directly for motion generation. (3)

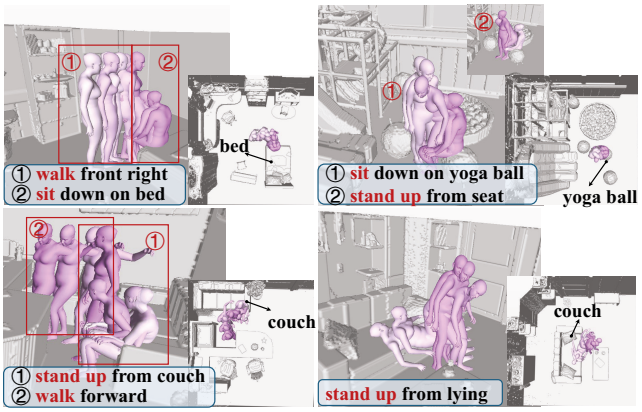

Figure 5: Qualitative Results on LINGO dataset. Our method generates realistic and coherent motion with natural contact.

MDM* (Tevet et al. 2023) is retrained with scene and target features using the same inputs as our method. (4) Cen et al. (Cen et al. 2024) adopt a two-step approach using LLMs for target grounding and an object-centric scene representation for motion generation. (1) and (2) do not infer the target object from text instructions, while (3) and (4) do. For fair comparison, we design two variants of our method: Ours (w/o Target), which does not explicitly use target information, and Ours (w/ Target), which incorporates it.

Quantitative Analysis and Results. (1) For the HUMANISE dataset, we conduct extensive experiments to evaluate the effectiveness of our proposed framework. As summarized in Tab. 1, Ours (w/ Target) achieves the best goal distance, contact and non-collision, indicating precise spatial alignment and robust physical plausibility. Ours (w/ Target) achieves an action accuracy of 0.976 and an FID of 0.163, indicating strong semantic consistency with intended actions and high motion realism. Our APD of 4.942 is slightly lower than Cen et al., yet our results remain natural and contextually coherent. We achieve a diversity score of 4.569 and a multi-modality score of 2.176, with the multi-modality closest to real human motion among all methods. Although our diversity is not the highest, the qualitative results still demonstrate diverse and plausible motion patterns. Compared to HUMANISE and Afford-Motion, Ours (w/o Target) achieves higher scores on all metrics except APD. Nevertheless, the quantitative results validate that Ours (w/o Target) generates motion that are more semantically accurate and contact-plausible. (2) For the LINGO dataset, base on "Cen et al." (Cen et al. 2024), we implement the AdaptDec architecture to construct the "Cen et al.*" baseline. Ours is trained with IGCF module to assess intention conditioning, whereas the "Cen et al.*" variant omits this module. As shown in Tab. 2, Ours better captures contact dynamics, with lower penetration rates and improved FID scores, demonstrating enhanced physical plausibility and motion quality.

Qualitative Analysis and Results. Understanding human motion relies on visual analysis, so we provide visual comparisons to demonstrate the effectiveness of our method.

Variants	goal dist. ↓	non-collision↑	accuracy↑	FID↓
w/o Contact	$0.082^{\pm 0.011}$	$\underline{99.88}^{\pm 0.000}$	$0.958^{\pm 0.001}$	$0.233^{\pm 0.007}$
w/o HA	$0.070^{\pm 0.011}$	$99.87^{\pm 0.000}$	$0.971^{\pm 0.001}$	$0.203^{\pm 0.004}$
w/ SA (×2)	$\underline{0.068}^{\pm 0.011}$	$99.86^{\pm 0.001}$	$\underline{0.976}^{\pm 0.001}$	$\mathbf{0.156}^{\pm 0.002}$
w/ SA (×6)	$0.069^{\pm 0.011}$	$99.86^{\pm 0.001}$	$\mathbf{0.979}^{\pm 0.000}$	$0.216^{\pm 0.002}$
w/ SA (×4)	$\mathbf{0.067}^{\pm 0.011}$	$\mathbf{99.96}^{\pm 0.002}$	$\underline{0.976}^{\pm 0.001}$	$\underline{0.163}^{\pm 0.001}$

Table 3: Ablation for the IGCF. The results demonstrate that SA4 successfully captures realistic and physically plausible human-object contact interaction.

Arch.	goal dist. ↓	contact↑	accuracy↑	FID↓
DiT	$0.076^{\pm 0.011}$	$94.47^{\pm 0.008}$	$\underline{0.967}^{\pm 0.001}$	$0.247^{\pm 0.003}$
U-Net	$0.137^{\pm 0.013}$	$87.93^{\pm 0.011}$	$0.886^{\pm 0.001}$	$0.994^{\pm 0.007}$
AD (w/o Guid)	$\underline{0.075}^{\pm 0.012}$	$\underline{94.88}^{\pm 0.008}$	$0.961^{\pm 0.001}$	$\underline{0.200}^{\pm 0.003}$
Adapt-Dec	$\mathbf{0.067}^{\pm 0.011}$	$\mathbf{95.30}^{\pm 0.007}$	$\mathbf{0.976}^{\pm 0.000}$	$\mathbf{0.163}^{\pm 0.001}$

Table 4: Ablation for IADM. The results verify that Adapt-Dec with guidance achieves the best motion quality.

(1) For the HUMANISE dataset (Fig. 4), HUMANISE and Afford-Motion frequently deviate from targets and occasionally penetrate objects. MDM* benefits from our scene input but still generates imprecise motion. Cen et al. improve localization, yet suffer from pose-object misalignment in actions like stand up and walk. In contrast, Ours (w/ Target and w/o Target), using semantic contact representation, achieves accurate contact alignment and avoids penetration, demonstrating superior spatial reasoning. (2) For the LINGO dataset, the results are presented in Fig. 5. We visualize longer motion sequences, which demonstrate that our method produces realistic and coherent motion with consistent contact behavior over time.

4.3 Ablation Studies

We conduct ablation studies on both the IGCF and the IADM to validate the rationale behind our design choices.

IGCF. (1) "w/o Contact" excludes contact conditioning entirely. (2)"w/o HA" uses a simple MLP to incorporate contact information, omitting the hierarchical attention mechanism. (3) "w/ SA" integrates contact and language features via the hierarchical attention mechanism, with "×2", "×4", "×6" indicating alignment depth. Among these, the "SA ×4" configuration achieves the best trade-off between effectiveness and efficiency (see Tab.3 and Fig.6-top). The IGCF module is instrumental in capturing fine-grained and physically plausible contact interaction between the human body and surrounding objects. Benefiting from IGCF, our method enables precise and realistic control of contact behaviors, enhancing physical plausibility for task-driven animation and embodied AI (see Fig. 7).

IADM. We explore three denoisers and conditional guidance strategies (see Tab. 4 and Fig. 6-bottom). (1) U-Net (Dabral et al. 2023; Karunratanakul et al. 2023), Diffusion Transformer (DiT) (Peebles and Xie 2023; Song et al. 2024a), and **Adapt-Dec**. We extend U-Net (Huang et al. 2024) and DiT (Song et al. 2024a; Huang et al. 2024), originally designed for text-to-motion, by incorporating language and scene geometry into the latent space, and embedding IGCF into the self-attention layers. U-Net and DiT struggle with multi-modal fusion, often generating incoherent motions, whereas Adapt-Dec better captures contact dynamics and produces more coherent, physically plausible results. (2) "AD (w/o Guid)" corresponds to the Adapt-Dec variant where both coarse and fine guidance mechanisms are removed. The observed performance gap demonstrates that guidance plays a critical role in enhancing motion quality.

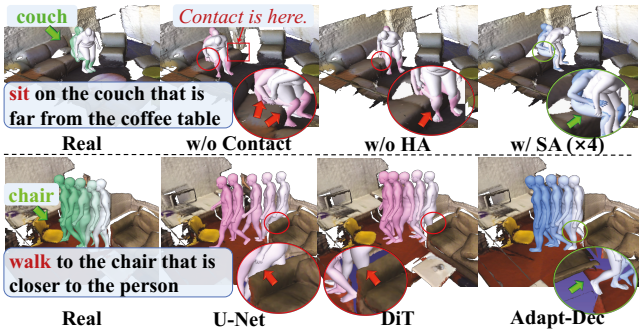

Figure 6: Ablation on IGCF and Denoiser. Results validate the effectiveness of the IGCF in producing the most accurate condition offsets, and confirm that our Adapt-Dec achieves superior motion quality.

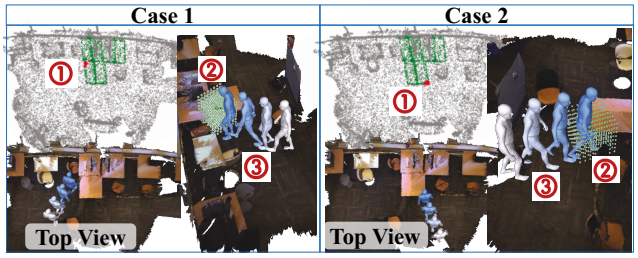

Figure 7: Controllable contact generation. Given the instruction "*walk to the table that is closer to the board*": ① selects the contact area on the table, ② generates contact field, and ③ synthesizes the motion sequence.

5 Conclusion, Limitation and Future Work

We present IntentMotion, a novel framework that bridges high-level linguistic intent and low-level physical contact for realistic motion generation in 3D scenes. By introducing the Intention-Guided Contact Field (IGCF) and the Intention-Aware Diffusion Model (IADM), IntentMotion enables structured contact reasoning and semantically grounded motion synthesis through a unified and differentiable architecture. Experiments demonstrate that IntentMotion excels at contact accuracy, semantic alignment, and generalization. However, IGCF still lacks fine-scale contact and scalability in clutter, and is confined to single-agentand static scenes; future work will extend to multi-scale, multi-agent, and temporal dynamics for real-world deployment.

Acknowledgments

This paper is supported by the National Natural Science Foundation of China (62572062, 62525204, 62441201, 62272021), the Beijing Natural Science Foundation (L232102), the Open Project Program of the State Key Laboratory of CAD&CG (Grant No. A2406), Zhejiang University, and the Research Projects of ISCAS (ISCAS-JCMS-202303, ISCAS-ZD-202401, ISCAS-JCZD-202402, ISCAS-JCMS-202403).

References

Bhatnagar, B. L.; Xie, X.; Petrov, I. A.; Sminchisescu, C.; Theobalt, C.; and Pons-Moll, G. 2022. Behave: Dataset and method for tracking human object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15935–15946.

Cen, Z.; Pi, H.; Peng, S.; Shen, Z.; Yang, M.; Shuai, Z.; Bao, H.; and Zhou, X. 2024. Generating Human Motion in 3D Scenes from Text Descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1855–1866.

Chen, J.; Hu, P.; Chang, X.; Shi, Z.; Kampffmeyer, M.; and Liang, X. 2025. Sitcom-crafter: A plot-driven human motion generation system in 3d scenes. *International Conference on Learning Representations*.

Chen, X.; Jiang, B.; Liu, W.; Huang, Z.; Fu, B.; Chen, T.; and Yu, G. 2023. Executing your commands via motion diffusion in latent space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18000–18010.

Dabral, R.; Mughal, M. H.; Golyanik, V.; and Theobalt, C. 2023. Mofusion: A framework for denoising-diffusion-based motion synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9760–9770.

Dai, W.; Chen, L.-H.; Wang, J.; Liu, J.; Dai, B.; and Tang, Y. 2024. Motionlcm: Real-time controllable motion generation via latent consistency model. In *Proceedings of the European Conference on Computer Vision*, 390–408.

Diller, C.; and Dai, A. 2024. Cg-hoi: Contact-guided 3d human-object interaction generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19888–19901.

Elfwing, S.; Uchibe, E.; and Doya, K. 2018. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural networks*, 107: 3–11.

Hassan, M.; Ghosh, P.; Tesch, J.; Tzionas, D.; and Black, M. J. 2021. Populating 3D scenes by learning human-scene interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14708–14718.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33: 6840–6851.

Huang, S.; Wang, Z.; Li, P.; Jia, B.; Liu, T.; Zhu, Y.; Liang, W.; and Zhu, S.-C. 2023. Diffusion-based generation, optimization, and planning in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16750–16761.

Huang, Y.; Yang, H.; Luo, C.; Wang, Y.; Xu, S.; Zhang, Z.; Zhang, M.; and Peng, J. 2024. Stablemofusion: Towards robust and efficient diffusion-based motion generation framework. In *Proceedings of the ACM International Conference on Multimedia*, 224–232.

Jaegle, A.; Borgeaud, S.; Alayrac, J.-B.; Doersch, C.; Ionescu, C.; Ding, D.; Koppula, S.; Zoran, D.; Brock, A.; Shelhamer, E.; Henaff, O. J.; Botvinick, M.; Zisserman, A.; Vinyals, O.; and Carreira, J. 2022. Perceiver IO: A General Architecture for Structured Inputs & Outputs. In *International Conference on Learning Representations*.

Jaegle, A.; Gimeno, F.; Brock, A.; Vinyals, O.; Zisserman, A.; and Carreira, J. 2021. Perceiver: General perception with iterative attention. In *International Conference on Machine Learning*, 4651–4664.

Jiang, N.; He, Z.; Wang, Z.; Li, H.; Chen, Y.; Huang, S.; and Zhu, Y. 2024a. Autonomous character-scene interaction synthesis from text instruction. In *SIGGRAPH Asia Conference Papers*, 1–11.

Jiang, N.; Zhang, Z.; Li, H.; Ma, X.; Wang, Z.; Chen, Y.; Liu, T.; Zhu, Y.; and Huang, S. 2024b. Scaling up dynamic human-scene interaction modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1737–1747.

Jin, P.; Wu, Y.; Fan, Y.; Sun, Z.; Yang, W.; and Yuan, L. 2023. Act as you wish: Fine-grained control of motion diffusion model with hierarchical semantic graphs. *Advances in Neural Information Processing Systems*, 36: 15497–15518.

Karunratanakul, K.; Preechakul, K.; Suwajanakorn, S.; and Tang, S. 2023. Guided Motion Diffusion for Controllable Human Motion Synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2151–2162.

Li, S.; Zhang, H.; Chen, X.; Wang, Y.; and Ban, Y. 2025. GenHOI: Generalizing Text-driven 4D Human-Object Interaction Synthesis for Unseen Objects. *arXiv preprint arXiv:2506.15483*.

Ma, S.; Cao, Q.; Zhang, J.; and Tao, D. 2024. Contact-aware human motion generation from textual descriptions. *arXiv preprint arXiv:2403.15709*.

Mao, W.; Hartley, R. I.; Salzmann, M.; et al. 2022. Contact-aware human motion forecasting. *Advances in Neural Information Processing Systems*, 35: 7356–7367.

Milacski, Z. Á.; Niinuma, K.; Kawamura, R.; De La Torre, F.; and Jeni, L. A. 2025. GHOST: Grounded Human Motion Generation with Open Vocabulary Scene-and-Text Contexts. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, 4108–4118.

Pavlakos, G.; Choutas, V.; Ghorbani, N.; Bolkart, T.; Osman, A. A.; Tzionas, D.; and Black, M. J. 2019. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10975–10985.

Peebles, W.; and Xie, S. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4195–4205.

Peng, X.; Xie, Y.; Wu, Z.; Jampani, V.; Sun, D.; and Jiang, H. 2025. Hoi-diff: Text-driven synthesis of 3d human-object interactions using diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2878–2888.

Pi, H.; Peng, S.; Yang, M.; Zhou, X.; and Bao, H. 2023. Hierarchical generation of human-object interactions with diffusion probabilistic models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15061–15073.

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

Shridhar, M.; Manuelli, L.; and Fox, D. 2023. Perceiver-actor: A multi-task transformer for robotic manipulation. In *Conference on Robot Learning*, 785–799.

Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; and Ganguli, S. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, 2256–2265.

Song, W.; Jin, X.; Li, S.; Chen, C.; Hao, A.; Hou, X.; Li, N.; and Qin, H. 2024a. Arbitrary motion style transfer with multi-condition motion latent diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 821–830.

Song, W.; Zhang, X.; Li, S.; Gao, Y.; Hao, A.; Hou, X.; Chen, C.; Li, N.; and Qin, H. 2024b. HOIAnimator: Generating Text-prompt Human-object Animations using Novel Perceptive Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 811–820.

Song, Y.; and Ermon, S. 2019. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, 32: 11918–11930.

Starke, S.; Zhang, H.; Komura, T.; and Saito, J. 2019. Neural state machine for character-scene interactions. *ACM Transactions on Graphics*, 38(6): 178.

Tevet, G.; Raab, S.; Cohan, S.; Reda, D.; Luo, Z.; Peng, X. B.; Bermano, A. H.; and van de Panne, M. 2025. CLoSD: Closing the Loop between Simulation and Diffusion for multi-task character control. In *International Conference on Learning Representations*.

Tevet, G.; Raab, S.; Gordon, B.; Shafir, Y.; Cohen-or, D.; and Bermano, A. H. 2023. Human Motion Diffusion Model. In *International Conference on Learning Representations*, 1–16.

Wang, J.; Rong, Y.; Liu, J.; Yan, S.; Lin, D.; and Dai, B. 2022a. Towards diverse and natural scene-aware 3d human motion synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20460–20469.

Wang, J.; Xu, H.; Xu, J.; Liu, S.; and Wang, X. 2021. Synthesizing long-term 3d human motion and interaction in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9401–9411.

Wang, Z.; Chen, Y.; Jia, B.; Li, P.; Zhang, J.; Zhang, J.; Liu, T.; Zhu, Y.; Liang, W.; and Huang, S. 2024. Move as You Say Interact as You Can: Language-guided Human Motion Generation with Scene Affordance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 433–444.

Wang, Z.; Chen, Y.; Liu, T.; Zhu, Y.; Liang, W.; and Huang, S. 2022b. Humanise: Language-conditioned human motion generation in 3d scenes. *Advances in Neural Information Processing Systems*, 35: 14959–14971.

Xing, C.; Mao, W.; and Liu, M. 2024. Scene-aware human motion forecasting via mutual distance prediction. In *Proceedings of the European Conference on Computer Vision*, 128–144.

Yi, H.; Thies, J.; Black, M. J.; Peng, X. B.; and Rempe, D. 2025. Generating human interaction motions in scenes with text control. In *Proceedings of the European Conference on Computer Vision*, 246–263.

Yuan, Y.; Song, J.; Iqbal, U.; Vahdat, A.; and Kautz, J. 2023. Physdiff: Physics-guided human motion diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 16010–16021.

Zhang, H.; Ye, Y.; Shiratori, T.; and Komura, T. 2021. Manipnet: neural manipulation synthesis with a hand-object spatial representation. *ACM Transactions on Graphics*, 40(4): 1–14.

Zhang, M.; Cai, Z.; Pan, L.; Hong, F.; Guo, X.; Yang, L.; and Liu, Z. 2024. MotionDiffuse: Text-Driven Human Motion Generation With Diffusion Model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(6): 4115–4128.

Zhang, M.; Guo, X.; Pan, L.; Cai, Z.; Hong, F.; Li, H.; Yang, L.; and Liu, Z. 2023a. Remodiffuse: Retrieval-augmented motion diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 364–373.

Zhang, M.; Li, H.; Cai, Z.; Ren, J.; Yang, L.; and Liu, Z. 2023b. Finemogen: Fine-grained spatio-temporal motion generation and editing. *Advances in Neural Information Processing Systems*, 36: 13981–13992.

Zhang, S.; Zhang, Y.; Ma, Q.; Black, M. J.; and Tang, S. 2020a. PLACE: Proximity learning of articulation and contact in 3D environments. In *International Conference on 3D Vision*, 642–651.

Zhang, Y.; Hassan, M.; Neumann, H.; Black, M. J.; and Tang, S. 2020b. Generating 3d people in scenes without people. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6194–6204.

Zhao, J.; Hou, R.; Tian, Z.; Chang, H.; and Shan, S. 2025. HIS-GPT: Towards 3D Human-In-Scene Multimodal Understanding. *arXiv preprint arXiv:2503.12955*.

Zhao, K.; Zhang, Y.; Wang, S.; Beeler, T.; and Tang, S. 2023. Synthesizing diverse human motions in 3d indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 14738–14749.