Learning Generatable Mutual Distance for Scene-Aware Human Motion Generation

Zonglin Yang[1,3], Chaoyue Xing[2], Yixuan Yin[1], Miaomiao Liu[2], and Liyuan Pan[1,3†]

[1] Beijing Institute of Technology, Beijing, China
[2] Australian National University, Canberra, Australia
[3] Yangtze Delta Region Academy of Beijing Institute of Technology, Jiaxing, China

Abstract. Generating human motion from language in 3D scenes requires both semantic alignment with textual intent and physically plausible human-scene interactions. Existing methods struggle because implicit language-scene fusion entangles motion synthesis with scene reasoning, while explicit grounding via fixed-vocabulary detectors or affordance maps lacks generalization and fails to capture full-body, temporally coherent interactions. We address this representation gap by identifying mutual distance as an effective interaction representation for modeling full-body spatiotemporal human-scene relations, and propose MDNet, a two-stage diffusion framework that explicitly generates and grounds such interactions. Since mutual distance encodes only relative geometry and is ambiguous for stochastic generation, we introduce VLM-based semantic anchors to provide absolute spatial grounding. We further model mutual distance in the frequency domain to enhance temporal stability before synthesizing final motions. Experiments on HUMANISE demonstrate state-of-the-art semantic alignment, physical plausibility, and temporal coherence, with cross-dataset generalization to unseen scenes.

Keywords: Scene-Aware Human Motion Generation · Mutual Distance · Interaction Representation

1 Introduction

Language-guided scene-aware human motion generation aims to synthesize full-body 3D motions that follow natural language instructions while remaining physically consistent with complex 3D scene geometry [47].

This task is fundamentally shaped by two intertwined challenges. First, natural language often provides only coarse or ambiguous intent, requiring the model to infer precise spatial and temporal motion goals within a complex environment. Second, the generated motion must satisfy strict physical interaction constraints, including joint limits, balance, collision avoidance, and coherent contact evolution with the scene. Together, these requirements severely restrict the feasible

† Corresponding author.

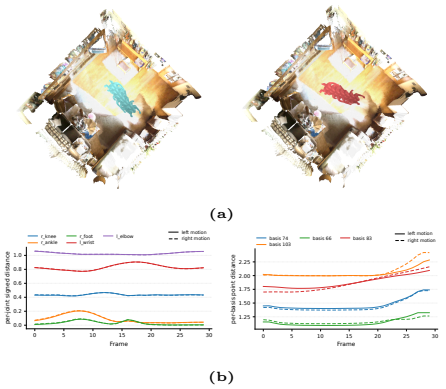

Fig. 1: Mutual distance ambiguity for stochastic motion generation. (a) Two semantically distinct walking motions in the same scene. (b) Corresponding mutual distance trajectories; solid and dashed curves denote different motions. Despite semantic differences, the trajectories are nearly identical, showing that mutual distance alone is insufficient for language-specified stochastic generation.

Fig. 2: Examples of generated motions. Motions generated by our MDNet (blue) and Wang *et al.* [40] (pink). Our MDNet produces physically plausible and temporally coherent motions, while the baseline exhibits penetration (walk) or floating/jittering (lie down), highlighting the impact of explicit interaction representation. Lighter colors indicate earlier frames (best viewed in color).

motion space and demand an interaction representation that can simultaneously support semantic grounding and physical consistency.

Recent generative models [3,40,41] attempt to address these challenges using conditional VAEs or diffusion models, yet they largely entangle motion synthesis with scene reasoning and lack an explicit interaction representation that can simultaneously encode global human-scene relations, support language grounding, and preserve temporal interaction dynamics. Implicit fusion strategies (e.g., HUMANISE [41]) often obscure task-relevant scene regions, leading to semantic misalignment [29]. Explicit grounding via affordance maps [40] focuses on sparse contact points and ignores full-body and temporal interaction structure, resulting in penetration, floating, and jittering (Fig. 2). These limitations expose a fundamental representation gap: existing methods fail to capture how the entire human body evolves relative to the scene in a structured, temporally coherent, and semantically grounded manner.

Mutual distance [42] provides properties that effectively address this representation gap. By combining per-joint signed distances to scene geometry with per-basis-point distances from scene anchors to the human body, mutual distance captures full-body human-scene relations and their temporal evolution in a unified and continuous manner, offering strong structural cues for physical plausibility and interaction coherence. However, mutual distance encodes only relative spatial configurations and lacks an absolute reference frame. While this

ambiguity can be resolved in motion forecasting using historical context, no such reference exists in stochastic generation. As a result, distinct motions performed in the same scene may yield nearly identical mutual distance trajectories, as illustrated in Fig. 1(a,b), where solid and dashed curves correspond to different motions. This inherent ambiguity prevents mutual distance from uniquely determining motions that satisfy language-specified goals (e.g., "walk from the chair to the table"), rendering it insufficient as a standalone representation for generation.

To resolve this ambiguity, we introduce semantic anchors that provide the missing absolute spatial reference, making mutual distance generatable for text-conditioned motion synthesis. By linking language-specified motion goals to concrete scene locations, semantic anchors provide absolute spatial conditions for relative human-scene interactions. We obtain these anchors using vision-language models (VLMs), which provide open-vocabulary understanding of object descriptions and scene context. This capability allows more flexible spatial grounding than fixed-vocabulary detector-based approaches such as Cen *et al.* [3], especially when target objects are unseen or ambiguously described.

Building on this insight, we propose MDNet, a two-stage diffusion framework that makes mutual distance a practical interaction representation for scene-aware motion generation. **First**, MDNet explicitly generates mutual distance to model full-body human-scene interactions, decoupling interaction reasoning from motion synthesis. **Second**, it resolves the inherent ambiguity of mutual distance by conditioning on VLM-derived semantic anchors that provide absolute spatial grounding for language-guided generation. **Third**, MDNet models mutual distance in the frequency domain using the discrete cosine transform (DCT), which suppresses high-frequency noise and facilitates temporally stable interaction synthesis within the diffusion process [7].

With these components, the Mutual Distance Generator (MDG) produces smooth, physically consistent mutual distance trajectories conditioned on scene geometry, language, and semantic anchors. The Mutual Distance-based Motion Generator (MDMG) then synthesizes full-body motions guided by the generated mutual distances. Experiments on HUMANISE [41] demonstrate that MDNet achieves superior semantic alignment, physical plausibility, and temporal coherence, with cross-dataset generalization on PROX [15]. Code will be released to support reproducibility.

Our main contributions are as follows:

- We make mutual distance an effective and expressive interaction representation for language-guided, scene-aware human motion generation.
- We resolve the inherent ambiguity of mutual distance for stochastic generation through VLM-based semantic anchors for absolute spatial grounding and frequency-domain modeling for temporal stability.
- We propose MDNet, a two-stage diffusion framework that generates anchor-grounded interaction trajectories and synthesizes physically plausible, semantically aligned full-body motions.

2 Related work

3D Human Motion Generation. Conditional 3D motion generation and understanding has been explored using action labels [4, 13], text [34, 44, 48, 49], music [24], objects [8, 45], and scene context [14, 22, 35–37, 43]. Early single-modality approaches [1, 13, 24] lacked adaptability in complex scenes. Text-to-motion models [11, 12] relied on latent alignment but struggled with scene constraints [28,40]. Recent methods combine signals, e.g., text with objects [8,26] or scenes [3,39,40] to improve physical realism, though joint modeling remains difficult due to modality gaps. In this work, we introduce a scene-aware, text-driven motion generation framework that explicitly models human-scene interactions via mutual distance, improving coherence and realism over latent-based methods [3].

Scene Understanding. Scene understanding involves recognizing objects and the overall semantic layout [9]. Most multi-modal visual grounding methods use a two-stage pipeline [18, 21, 32, 38, 52], which requires separate object detection and matching, resulting in longer inference times and limited generalization to open-world objects [3,20,27]. In contrast, we adopt a BEV-based semantic anchor prediction approach with the Vision Language Models (VLMs), whose structured and interpretable representations better support recognition and reasoning.

Scene-aware Motion Synthesis. For scene-aware human motion synthesis, it is necessary to consider the constraints imposed by the scene on human movement to generate more realistic motions. Previous methods [2,3,16,35,37,41,46] lack explicit constraints between humans and the scene, thus, the human motions generated by these methods often suffer from issues such as "ghost motion". SemGeoMo [6] generates dynamic contextual human motions from sequential point clouds of interactive targets. In contrast, MoMaps [25] studies semantics-aware 3D scene motion generation from a single image. Besides this, [40] adopts the contact map as an explicit representation of human-scene interaction, which explicitly encodes the contact relationship between the scene surface and human joints. However, this representation can only constrain the contact part, leaving other parts unconstrained, which is not a suitable constraint for the global trajectory of human motion.

In this paper, we introduce the mutual distance representation into human motion generation. In contrast to [40], mutual distance representation effectively provides comprehensive temporal and full-body constraints, enabling more realistic and coherent motion generation.

3 Preliminary

3.1 Motion Representations

Following Guo *et al.* [11] and Wang *et al.* [40], we represent the human pose for each frame using the joint positions from the SMPL-X body model, specifically denoted as $\mathbf{x}_u \in \mathbb{R}^{J\times 3}$ for the pose at motion timestep u, where J refers to the total number of joints. The output of our model consists of a pose sequence

represented as $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_U\}$, with U being the total number of motion timesteps. For visualization purposes, these poses are converted into body meshes parameterized by SMPL-X [30], by adjusting the body parameters to fit the joint positions.

3.2 Scene and Text Representations

We represent the 3D scene using a Signed Distance Field Volume as $\mathbf{S} \in \mathbb{R}^{N \times N \times N}$ following [42]. Each value in this volume shows how far a voxel is from the closest surface in the scene. The sign of the value indicates whether the voxel is inside or outside the surface.

The language description $\mathbf{L}$ is composed of W words. We extract its feature representation using a pretrained CLIP model [31], which encodes the entire sentence into a global text feature vector $\mathbf{F}_{\text{text}} \in \mathbb{R}^{512}$.

3.3 Diffusion Models

Diffusion models [17, 33] are generative models that synthesize data through a two-step process: a forward noising step and a reverse denoising step. The forward process gradually adds Gaussian noise to real data x_0, forming a sequence $\{x_t\}_{t=1}^{T}$, where each step follows the distribution $q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{\alpha_t}x_{t-1}, (1-\alpha_t)\mathbf{I})$. As t increases, the distribution converges to $\mathcal{N}(0, \mathbf{I})$.

The reverse process learns to reconstruct x_0 from $x_T \sim \mathcal{N}(0, \mathbf{I})$ by modeling $p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t))$, where μ_θ and Σ_θ are predicted by a neural network. In conditional generation, additional context C is incorporated into the reverse process as $p_\theta(x_{t-1}|x_t, C)$, enabling generation guided by text, scene, or other signals.

4 Method

Overview. We first formulate the text-driven scene-aware 3D human motion generation task. Given a 3D scene $\mathbf{S}$ and language description $\mathbf{L}$, our objective is to generate a 3D pose sequence $\mathbf{X}$ with U frames. The pipeline of our MD-Net is illustrated in Fig. 3. It consists of a Mutual Distance Generator (MDG) and a Mutual Distance-based Motion Generator (MDMG). In section 4.2, our MDG first transfers the noisy mutual distance to the frequency domain, then synthesizes mutual distance $\hat{\mathbf{M}}$ conditioned on scene SDF volume $\mathbf{S}$, language description $\mathbf{L}$, and semantic anchors $\mathbf{P}_{\text{se}}$. The MDMG synthesizes human motions $\hat{\mathbf{X}}$ conditioned on $\hat{\mathbf{M}}$, $\mathbf{S}$, $\mathbf{L}$, and $\mathbf{P}_{\text{se}}$. Here, the semantic anchors (start and end positions) are predicted by a VLM, Gemini [10] (refer to Sec. 4.3), denoted as $\mathbf{P}_{\text{se}} \in \mathbb{R}^{2\times 2}$, where each row corresponds to a 2D spatial coordinate.

4.1 Mutual Distance

Mutual distance consists of two complementary measures: (1) the *per-joint signed distance*, encoding detailed local interactions, and (2) the *per-basis point distance*, capturing global positional dynamics.

Its formal definition is as follows. Given a set of human joints $\mathcal{J}_u = \{\mathbf{J}_{ju}\}_{j=1}^{J}$ at motion timestep u, we define the per-joint signed distance d_{ju} for joint j as the shortest Euclidean distance from joint position $\mathbf{J}_{ju}$ to the nearest scene surface point, assigning a negative sign when the joint lies inside the solid volume enclosed by the scene. We use $\mathbf{S}_{in}$ to represent the volume inside the actual solid scene. The $\partial\mathbf{S}$ is the set of points on the scene surface, and $\mathbf{y} \in \partial\mathbf{S}$ denotes a surface point. Formally, the d_{ju} is defined as:

$$d_{ju} = \begin{cases} -\min\limits_{\mathbf{y}\in\partial\mathbf{S}} \|\mathbf{J}_{ju} - \mathbf{y}\|_2, & \text{if } \mathbf{J}_{ju} \in \mathbf{S}_{in}, \\ \min\limits_{\mathbf{y}\in\partial\mathbf{S}} \|\mathbf{J}_{ju} - \mathbf{y}\|_2, & \text{if } \mathbf{J}_{ju} \notin \mathbf{S}_{in}, \end{cases} \tag{1}$$

where $\mathbf{d}_u = [d_{1u}, d_{2u}, \cdots, d_{Ju}]^\top \in \mathbb{R}^J$, is the per-joint signed distance at motion timestep u.

To capture global positional context, we define a set of K basis points $\{\mathbf{p}_k\}_{k=1}^{K}$ fixed relative to the scene origin. At each motion timestep u, the per-basis point distance b_{ku} is computed as the minimum Euclidean distance from each basis point $\mathbf{p}_k$ to all joints in the current pose:

$$b_{ku} = \min_{\mathbf{y}\in\mathcal{J}_u} \|\mathbf{p}_k - \mathbf{y}\|_2, \tag{2}$$

where $\mathbf{b}_u = [b_{1u}, b_{2u}, \cdots, b_{Ku}]^\top \in \mathbb{R}^K$, is the per-basis point distance at motion timestep u. We first pre-compute the mutual distance for training, denoted as,

$$\mathbf{D} = [\mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_U] \in \mathbb{R}^{J\times U}, \mathbf{B} = [\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_U] \in \mathbb{R}^{K\times U}. \tag{3}$$

For consistency and compactness, we concatenate them along the spatial dimension to form the final mutual distance:

$$\mathbf{M} = [\mathbf{D}; \mathbf{B}] = [\mathbf{m}_1, \mathbf{m}_2, \ldots, \mathbf{m}_U]^\top \in \mathbb{R}^{U\times(J+K)} \ . \tag{4}$$

4.2 MDNet Architecture

In the first stage, we propose the Mutual Distance Generator (MDG), a diffusion-based framework that synthesizes mutual distance sequences conditioned on scene geometry $\mathbf{S}$, language descriptions $\mathbf{L}$, and semantic anchors $\mathbf{P}_{\text{se}}$. The generation process is formulated as:

$$p_\theta(\mathbf{M}^{0:T} \mid \mathbf{S}, \mathbf{L}, \mathbf{P}_{\text{se}}) = p(\mathbf{M}^T) \prod_{t=1}^{T} p_\theta(\mathbf{M}^{(t-1)} \mid \mathbf{M}^{(t)}, \mathbf{S}, \mathbf{L}, \mathbf{P}_{\text{se}}), \tag{5}$$

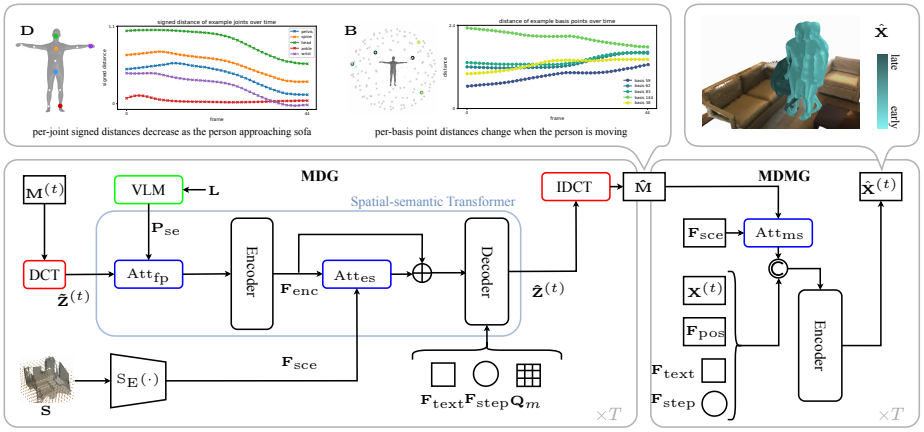

Fig. 3: Overview of our two-stage MDNet. The first stage, **Mutual Distance Generator (MDG)**, synthesizes mutual distance sequences by conditioning on scene geometry, text input, and semantic anchors. Temporal modeling is improved using the Discrete Cosine Transform (DCT). Next, in the second stage, the **Mutual Distance-based Motion Generator (MDMG)** generates full-body motion sequences guided by the predicted mutual distances and contextual features. On the top row, we select five joints and basis points, and show the corresponding per-joint signed distance and per-basis point distance curves over time for the example "sit on the sofa."

where $\mathbf{M}^{(t)}$ denotes the mutual distance at diffusion timestep t. As illustrated in Fig. 3, our MDG comprises three modules: (1) a *DCT Module*, (2) a *Spatial-semantic Transformer*, and (3) an *IDCT Module*. The modules are described below.

DCT Module. We apply DCT along temporal dimension, i.e., independently on each column of the mutual distance $\mathbf{M}^{(t)}$, yielding a smooth frequency-domain representation $\mathbf{Z}$. Formally, for each index $i \in \{1, 2, \ldots, J+K\}$, the transformation is defined as:

$$\mathbf{Z}_{if}^{(t)} = \sqrt{\frac{2}{U}} \sum_{u=1}^{U} M_{iu}^{(t)} \frac{1}{\sqrt{1+\delta_{f1}}} \cos\left(\frac{\pi}{2U}(2u+1)(f-1)\right), \tag{6}$$

where $M_{iu}^{(t)}$ represents the mutual distance of the i-th joint or basis point at motion timestep u, and $Z_{if}^{(t)}$ is the corresponding DCT coefficient at frequency index f. The δ_{f1} is a Kronecker delta function. With all cosine bases precomputed, the DCT can be represented as a linear transformation, $\mathbf{Z}^{(t)} = \mathbf{C}\mathbf{M}^{(t)}$, where $\mathbf{C} \in \mathbb{R}^{U \times U}$ is an orthogonal transformation matrix storing the precomputed cosine bases. Since most energy lies in low frequencies, we retain only the first d coefficients:

$$\tilde{\mathbf{Z}}^{(t)} = \mathbf{C}_d \mathbf{M}^{(t)}, \tag{7}$$

where $\mathbf{C}_d \in \mathbb{R}^{d \times U}$ contains the first d rows of $\mathbf{C}$, yielding a compact representation $\tilde{\mathbf{Z}}^{(t)} \in \mathbb{R}^{d \times (J+K)}$.

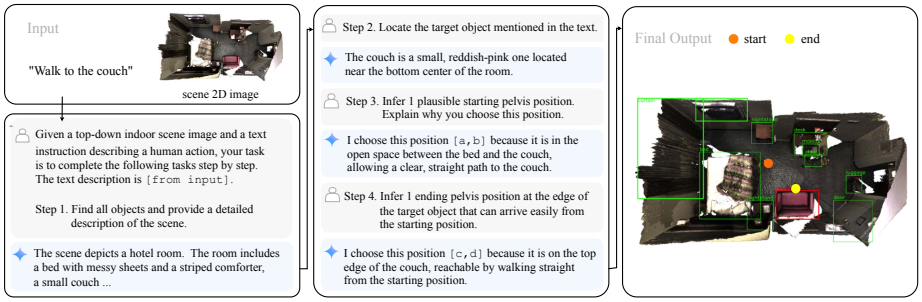

Fig. 4: VLM-guided anchor prediction pipeline. Given a BEV image and a text instruction (e.g., "Walk to the couch"), a VLM with chain-of-thought (CoT) reasoning performs the following steps: (1) identifies scene objects, (2) localizes the target, and (3–4) predicts plausible start/end pelvis positions. Green boxes indicate detected objects; orange/yellow dots denote start/end anchors.

Spatial-semantic Transformer. We propose a hierarchical transformer that refines mutual distance via structured multimodal interactions, integrating temporal, spatial, and semantic contexts. Starting from $\tilde{\mathbf{Z}}^{(t)}$, we incorporate positional information $\mathbf{F}_{\text{pos}}$ using an attention fusion block Att_{fp}, and then process the result with a transformer encoder. $\mathbf{F}_{\mathbf{pos}}$ is extracted by an MLP from $\mathbf{P}_{\text{se}}$. To incorporate scene geometry, we apply another attention module Att_{es} to fuse encoder output $\mathbf{F}_{\text{enc}}$ with scene spatial features $\mathbf{F}_{\text{sce}}$ extracted by a 3D-CNN-based encoder $\text{S}_{\text{E}}(\cdot)$. A multimodal transformer decoder then generates the final representation. Its queries are composed of textual features $\mathbf{F}_{\text{text}}$, diffusion timestep embeddings $\mathbf{F}_{\text{step}}$ of t, and learnable query tokens $\mathbf{Q}_m$. The keys and values are derived from the output of Att_{es}, which encodes the mutual distance with both scene and positional information.

IDCT Module. To reconstruct mutual distance sequences from latent frequency-domain embeddings $\hat{\mathbf{Z}}^{(t)}$, we apply the IDCT (Inverse DCT). The reconstruction process is defined as,

$$\hat{\mathbf{M}}^{(t)} = \mathbf{C}_d^{\top} \hat{\mathbf{Z}}^{(t)}, \tag{8}$$

where $\mathbf{C}_d^{\top} \in \mathbb{R}^{U \times d}$ denotes the truncated inverse cosine basis, and $\hat{\mathbf{M}}^{(t)}$ is the reconstructed mutual distance sequence.

Our model directly estimates denoised mutual distances in the frequency domain. We use an ℓ_1 loss to enforce temporal fidelity:

$$\mathcal{L}_{\text{MDG}} = \mathbb{E}_{\tilde{\mathbf{Z}}^0, t}[\|\tilde{\mathbf{Z}}^0 - G_\theta(\tilde{\mathbf{Z}}^{(t)}, t, \mathbf{S}, \mathbf{L}, \mathbf{P}_{\text{se}})\|_1], \tag{9}$$

where G_θ is our MDG model and $\tilde{\mathbf{Z}}^0$ is the GT DCT representation.

In the second stage, we use the Mutual Distance-based Motion Generator (MDMG) to synthesize the final human motion sequence based on the predicted mutual distance from MDG. The motion generation process follows a conditional diffusion formulation,

$$p_\phi(\mathbf{X}^{0:T}|\hat{\mathbf{M}},\mathbf{S},\mathbf{L},\mathbf{P}_{\text{se}}) = p(\mathbf{X}^T)\prod_{t=1}^{T} p_\phi(\mathbf{X}^{(t-1)} \mid \mathbf{X}^{(t)},\hat{\mathbf{M}},\mathbf{S},\mathbf{L},\mathbf{P}_{\text{se}}), \tag{10}$$

where $\mathbf{X}^{(t)}$ is the human motion sequence at diffusion timestep t. The MDMG model first integrates the predicted mutual distance $\hat{\mathbf{M}}$ with the scene feature $\mathbf{F}_{\text{sce}}$ using an attention-based fusion module Att_{ms}. The fused representation is then concatenated with additional inputs. This combined input is passed to a transformer encoder, and the output, $\hat{\mathbf{X}}$, is the reconstructed motion sequence.

We primarily optimize the model using an ℓ_1 loss, defined as:

$$\mathcal{L}_{\text{motion}} = \mathbb{E}_{\mathbf{X}^0,t}\left[\|\mathbf{X}^0 - G_\phi(\mathbf{X}^{(t)}, t, \hat{\mathbf{M}}, \mathbf{S}, \mathbf{L}, \mathbf{P}_{\text{se}})\|_1\right], \tag{11}$$

where G_ϕ represents our proposed MDMG model and $\mathbf{X}^0$ represents ground truth motions.

In addition, we introduce two consistency losses. Given that our 3D scene representation is modeled as a signed distance volume, we query the signed distance of each human joint by interpolating within this volume. This enables a per-joint signed distance consistency loss, defined as,

$$\mathcal{L}_{\text{joint}} = \frac{1}{UJ}\sum_{u=1}^{U}\sum_{j=1}^{J}\left|\hat{d}_{ju} - d_{ju}\right|, \tag{12}$$

where $\hat{d}_{ju}$ and d_{ju} denote the signed distances of the j-th joint in the predicted and GT human pose at motion timestep u, respectively. We also introduce a per-basis point distance consistency loss to enforce spatial alignment between the predicted human motion and the scene. Formally, it is defined as,

$$\mathcal{L}_{\text{basis}} = \frac{1}{UK}\sum_{u=1}^{U}\sum_{k=1}^{K}\left|\hat{b}_{ku} - b_{ku}\right|, \tag{13}$$

where $\hat{b}_{ku}$ and b_{ku} represent the predicted and GT distances for the k-th basis point at timestep u, respectively. Ground-truth motions are used only during training to compute these losses. The overall training loss is:

$$\mathcal{L}_{\text{MDMG}} = \mathcal{L}_{\text{motion}} + \mathcal{L}_{\text{joint}} + \mathcal{L}_{\text{basis}}. \tag{14}$$

4.3 Semantic Anchor Prediction

We adopt a BEV-based semantic anchor localization approach (predicting the start and end positions of the motion sequence) that efficiently integrates with our mutual distance-based, language-guided motion generation pipeline. Given the inferred semantic anchors, the generated motion is constrained to be close to the target object. The pipeline includes two main steps:

(i) We render a BEV image by projecting the 3D scene from a virtual overhead camera, as it efficiently captures the full scene and walkable areas for motion

generation, while multi-view representations require costly occlusion reasoning and rendering.

(ii) We apply a VLM to the rendered BEV image. Unlike 3D-scene models such as 3D-LLM [19] or LL3DA [5], which require complex 3D feature construction (e.g., SLAM) or task-specific tuning, a general 2D VLM is sufficient and far easier to deploy, making it more suitable for pixel-level anchor prediction. Using chain-of-thought prompting, the VLM interprets natural language instructions, identifies relevant objects, and predicts 2D semantic anchors aligned with the intended action. An example of this process is shown in Fig. 4. The predicted anchors are then lifted back into 3D space. The complete procedure is provided as pseudocode in the supplementary A. Prompts are provided in supplementary D.

5 Experiments

5.1 Implementation Details

We adopt a frozen CLIP-ViT-B/32 [31] to extract textual features for both stages. All transformer-based modules are implemented using the native PyTorch implementation. Both diffusion models are trained with the AdamW optimizer using a learning rate of 1×10^{-4} and a batch size of 32 on a single NVIDIA A100 GPU. We utilize Gemini-1.5-Pro [10] to generate semantic anchors. For each motion-language pair, we generate 5 anchor candidates to enhance the diversity of generated motions. To compute the per-basis point distance, we follow the protocol of [42] and uniformly sample 150 basis points located 2.0 meters away from the scene origin. We retain the first 50 DCT coefficients to capture the dominant low-frequency components.

5.2 Datasets

Experiments are conducted on the **HUMANISE** [41], a large-scale benchmark for language-conditioned motion generation in 3D scenes. It includes four action categories, each paired with language descriptions of the action and target object. Following [40], we exclude spatially referring descriptions. We also use **PROX** [15], which captures human motions interacting with real 3D indoor environments for studying human-scene physical interactions.

5.3 Metrics and Baselines

We assess semantic alignment and diversity using goal distance (goal dist.) and Average Pairwise Distance (APD) [40, 41]. For physical realism, we report non-collision and contact scores [40, 50], measuring the ratio of body meshes without penetration and with meaningful contact. We additionally report $\text{pene}_{\text{mean}}$ and pene_{max} [23, 51] to evaluate the average penetration over time and the maximum penetration in one sequence, respectively. To assess motion expressiveness and

Table 1: Quantitative results on the HUMANISE dataset. We evaluate semantic alignment (**goal dist.**), physical plausibility (**contact**, **non-collision**, **pene$_{\text{mean}}$**, **pene$_{\text{max}}$**), and motion quality (**APD**, **TMV**). ↓ indicates the lower the better, ↑ the higher the better, and → the closer to GT the better. The best and the second-best are highlighted in **bold** and underlined.

Method	goal dist.↓	contact↑	non-collision↑	pene$_{\text{mean}}$ ↓	pene$_{\text{max}}$ ↓	APD↑	TMV→
G.T.	0.014	-	-	-	-	0.000	0.1875
Humanise [41]	0.422	84.06	<u>99.77</u>	1.283	2.656	**4.094**	0.2055
Cen *et al.* [3]	0.384	86.36	99.60	<u>1.144</u>	<u>2.551</u>	2.073	<u>0.1908</u>
Wang *et al.* [40]	<u>0.156</u>	**95.86**	99.69	1.367	14.768	<u>2.597</u>	0.1486
Ours	**0.057**	<u>95.05</u>	**99.81**	**0.635**	**1.849**	2.235	**0.1870**

temporal smoothness, we propose the Temporal Motion Variance (TMV) metric:

$$\text{TMV} = \text{std}\left(\frac{\gamma}{|i-j|}\left\|\mathbf{x}_i - \mathbf{x}_j\right\|_2\right)_{1\le i<j\le U,} \tag{15}$$

where $\mathbf{x}_i$ is the pose at timestep i, and $\gamma = 10$ balances scale. The weight $\frac{\gamma}{|i-j|}$ emphasizes short-range variations while retaining global structure. A TMV close to that of ground truth (GT) indicates natural, coherent motion; higher values imply jitter, and lower ones suggest underactuated or static sequences.

We compare our method with the following baselines, (i) Humanise [41]: a cVAE-based approach. We directly use their released models; (ii) Cen *et al.* [3]: a diffusion-based generation method with 3D object detection. Note that its goal dist. evaluation includes spatially referring descriptions; (iii) Wang *et al.* [40]: a two-stage diffusion-based method using affordance map for scene awareness.

5.4 Results

Tab. 1 shows a comparison across multiple evaluation metrics. Our method achieves the best goal distance, indicating improved semantic grounding. We also outperform all baselines in physical plausibility, with higher non-collision rate and lower average and maximum penetration. Despite reduced penetration, our method maintains a competitive contact score, showing that it preserves meaningful body-scene contact. These results highlight stronger spatial awareness, enabled by signed distances in mutual distance.

For motion quality, we obtain a lower APD than Humanise [41] and Wang *et al.* [40], which is due to its higher controllability and the generation of more semantically aligned, object-aware actions. This balance is further reflected in TMV, where our score (0.1870) nearly matches the GT (0.1875), indicating our sequences exhibit realistic temporal variation while avoiding excessive jitter or static frames.

Fig. 5 showcases visual comparisons across different interaction scenarios. Humanise [41] and Cen *et al.* [3] produce floating artifacts or fail to reach the correct

Table 2: Quantitative comparison under two scene cropping strategies. GT denotes using GT anchors with random offsets, while VLM denotes using our predicted anchors for scene cropping.

Method	Cropping	goal dist.↓	contact↑	non-collision↑	$\text{pene}_{\text{mean}}$ ↓	pene_{max} ↓	APD↑	TMV→
Wang_{ori}	GT	0.156	95.86	99.69	1.367	14.768	**2.597**	0.1486
Wang_{se}	GT	0.057	**95.93**	99.78	1.057	10.133	1.969	0.1567
Ours	GT	**0.055**	95.11	**99.81**	**0.434**	**1.872**	2.144	**0.1893**
Wang_{ori}	VLM	0.158	96.27	99.65	1.642	17.556	**2.506**	0.1480
Wang_{se}	VLM	0.060	**96.32**	99.74	1.062	11.853	2.028	0.1502
Ours	VLM	**0.057**	95.05	**99.81**	**0.635**	**1.849**	2.235	**0.1870**

Table 3: Ablation results on the HUMANISE test set. We evaluate the contribution of mutual distance components, position anchors, and DCT encoding. ↓: lower is better, ↑: higher is better, →: closer to GT is better.

Method	goal dist.↓	contact↑	non-collision↑	$\text{pene}_{\text{mean}}$ ↓	pene_{max} ↓	APD↑	TMV→
w/o mutual distance	0.298	86.54	99.77	1.921	6.680	3.903	0.3573
w/o per-joint signed dist.	0.059	93.98	99.80	0.992	3.998	2.229	0.2524
w/o per-basis point dist.	0.158	94.66	99.76	0.701	2.829	2.361	0.2083
w/o semantic anchors	0.365	90.37	99.79	0.767	4.010	**4.433**	0.2383
w/o DCT & IDCT	0.059	94.96	99.80	0.741	3.246	2.178	0.2046
Rand-proj. mutual dist.	0.147	94.42	99.70	1.343	3.928	2.369	0.2145
Ours	**0.057**	**95.05**	**99.81**	**0.635**	**1.849**	2.235	**0.1870**

target object, such as sitting or lying in the air. Wang *et al.* [40] exhibits obvious penetrations, motion jitter, and limited dynamics, such as basically static or incomplete transitions. In contrast, our method generates stable, realistic motions with smooth temporal evolution and precise object interaction. More results can be found in supplementary F.

5.5 Ablation Study and Discussions

VLM Cropping. Tab. 2 compares the performance of Wang [40] and its variants under two scene cropping settings: *Ground-truth Scene Cropping*, which follows [40] using GT motion positions, and *VLM Scene Cropping*, which uses our predicted anchors from 2D reasoning. Wang_{ori} is the original model and Wang_{se} denotes we integrate semantic anchors for their model. The use of VLM-predicted anchors improves goal alignment and penetration performance for both Wang *et al.* and our method, confirming the general applicability and robustness of our anchor prediction strategy.

Model Component. Tab. 3 shows that removing mutual distance causes the largest performance drop across all metrics. Disabling per-joint or per-basis distances respectively degrades physical plausibility and goal alignment, confirming their complementarity. Without position anchors, goal distance and APD

(a) GT (b) Humanise (c) Cen *et al.* (d) Wang *et al.* (e) Ours

Fig. 5: Qualitative comparison of four interaction scenarios shows that our method (e, blue) generates motions that are physically plausible and semantically aligned, outperforming prior works [3, 40, 41] (b–d, pink). For instance, in "walk to the table," our model avoids penetration and unnatural poses. Ground truth is shown in (a, green); lighter colors indicate earlier frames.

increase notably, highlighting their role in semantic grounding. Removing DCT and relying on time-domain modeling raises penetration and TMV, showing its benefit for temporal coherence. We also replace mutual distance with $\mathbf{M}' = \mathbf{MR}$ using a fixed random orthogonal matrix $\mathbf{R}$, and retrain both stages under identical settings. The drop across metrics shows that its structure, rather than representation capacity alone, is important for generation.

VLM Model. We evaluate four popular VLMs for anchor prediction: Claude-3.5-Sonnet, GPT-o4-mini, GPT-4o, and Gemini-1.5-Pro (Tab. 4). Among them, Gemini-1.5-Pro consistently locates the correct target and produces valid anchors. Other models, while understanding the task intent, may sometimes struggle to pinpoint the correct object or yield semantically misaligned anchors, resulting in lower goal accuracy. Based on this, we adopt Gemini-1.5-Pro for reliable anchor generation and downstream performance. Despite weaker VLMs reducing performance, our model stays competitive in physical plausibility with methods in Tab. 1 due to our mutual distance and consistency losses.

DCT Coefficients. We evaluate how the number of DCT coefficients affects reconstruction accuracy using mean and max per-joint and per-basis point errors. As shown in Fig. 6, errors decrease as the number increases, with notable gains from 20 to 50 coefficients. Beyond 50, improvements are marginal. We thus choose 50 as a balance between compression and accuracy.

Table 4: Anchor prediction performance of different VLMs.

Method	goal dis.	conta.	n-col.	$\text{pene}_{\text{mean}}$	pene_{max}
Claude-3.5-Son.	0.629	92.84	99.61	1.150	2.939
GPT-o4-mini	0.390	92.83	99.63	1.052	2.046
GPT-4o	0.599	91.61	99.66	1.041	3.560
Gemini-1.5-Pro	**0.057**	**95.05**	**99.81**	**0.635**	**1.849**

Table 5: Generalization results on PROX dataset. Our method demonstrates generalization ability to unseen scenes.

Method	goal dis.	non-col.	conta.	$\text{pene}_{\text{mean}}$	pene_{max}
Cen *et al.*	0.281	99.65	86.62	0.963	1.921
Wang *et al.*	0.283	99.69	95.67	0.816	2.237
Ours	**0.054**	**99.87**	**95.93**	**0.270**	**0.654**

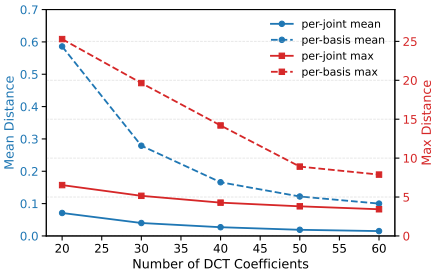

Fig. 6: Reconstruction error of mutual distance under different numbers of DCT coefficients. Blue curves show mean distances and red curves show maximum distances; solid and dashed lines denote per-joint and per-basis measures, respectively.

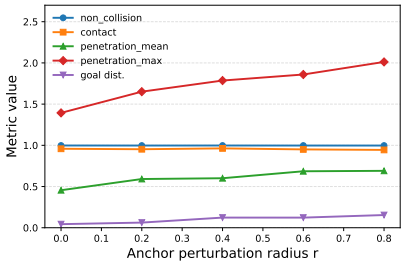

Fig. 7: Anchor perturbation robustness. As the perturbation radius r (in meters) increases, penetration-related metrics and goal distance generally increase, while non-collision and contact remain relatively stable.

Anchor-noise robustness. We evaluate the robustness of MDNet to imperfect grounding by perturbing the start and end semantic anchors in the bird's-eye-view (BEV) plane (Fig. 7). For each perturbation radius r, we apply a fixed-radius 2D offset with uniformly sampled direction to the semantic anchors. As shown in Fig. 7, the metrics remain relatively stable or degrade gradually as the perturbation increases, indicating that MDNet is robust to the quality and stability of the VLM. Details are provided in supplementary G.

(a) sit on the sofa (b) stand up from the end table (c) lie in an open area (d) I'm exhausted and sleepy. Where should I go?

Fig. 8: Qualitative results on PROX. Our model produces plausible interactions on unseen PROX scenes and diverse text formats.

5.6 Generalization

To assess cross-dataset generalization, we follow [3] and evaluate our HUMANISE-trained model on the PROX dataset [15] without finetuning. We construct motion-language pairs including novel instructions (e.g., "I'm exhausted and sleepy after a long day. Where should I go?"), where our VLM successfully grounds targets (e.g. bed). For unseen language descriptions, the VLM first identifies the referenced target object and subsequently maps the instruction to a corresponding HUMANISE action category (e.g., mapping "Where should I go?" to "walk to ..."). As shown in Tab. 5 and Fig. 8, our method outperforms [3,40] in both goal alignment and physical plausibility. In contrast, the 3D object detector in [3] sometimes fails under domain shift, while our VLM-based grounding remains robust. Details are provided in supplementary B.

6 Conclusion

We propose MDNet for language-guided, scene-aware human motion generation, built on mutual distance as an explicit interaction representation. By grounding mutual distance with VLM-predicted semantic anchors and modeling it in the frequency domain, MDNet enables physically plausible, semantically consistent, and temporally stable motion synthesis. Extensive experiments demonstrate superior performance and cross-dataset generalization compared with state-of-the-art methods.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (62302045), the Fundamental Research Funds for the Central Universities, and the BIT Special-Zone.

References

1. Ahn, H., Ha, T., Choi, Y., Yoo, H., Oh, S.: Text2action: Generative adversarial synthesis from language to action. In: 2018 IEEE International Conference on Robotics and Automation (ICRA). pp. 5915–5920. IEEE (2018)
2. Cao, Z., Gao, H., Mangalam, K., Cai, Q.Z., Vo, M., Malik, J.: Long-term human motion prediction with scene context. In: European Conference on Computer Vision. pp. 387–404. Springer (2020)
3. Cen, Z., Pi, H., Peng, S., Shen, Z., Yang, M., Zhu, S., Bao, H., Zhou, X.: Generating human motion in 3d scenes from text descriptions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1855–1866 (2024)
4. Chen, J., Xu, Q., Pan, L.: Darkshake-dvs: Event-based human action recognition under low-light and shaking camera conditions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20149–20159 (2026)

5. Chen, S., Chen, X., Zhang, C., Li, M., Yu, G., Fei, H., Zhu, H., Fan, J., Chen, T.: Ll3da: Visual interactive instruction tuning for omni-3d understanding reasoning and planning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 26428–26438 (2024)
6. Cong, P., Wang, Z., Ma, Y., Yue, X.: Semgeomo: Dynamic contextual human motion generation with semantic and geometric guidance. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 17561–17570 (2025)
7. Crabbé, J., Huynh, N., Stanczuk, J., Van Der Schaar, M.: Time series diffusion in the frequency domain. In: Proceedings of the 41st International Conference on Machine Learning. ICML'24, JMLR.org (2024)
8. Diller, C., Dai, A.: Cg-hoi: Contact-guided 3d human-object interaction generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19888–19901 (2024)
9. Gao, H., Mao, W., Liu, M.: Visfusion: Visibility-aware online 3d scene reconstruction from videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17317–17326 (2023)
10. Gemini Team, Georgiev, P., Lei, V.I., Burnell, R., Bai, L., Gulati, A., Tanzer, G., Vincent, D., Pan, Z., Wang, S., et al.: Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv preprint arXiv:2403.05530 (2024)
11. Guo, C., Zou, S., Zuo, X., Wang, S., Ji, W., Li, X., Cheng, L.: Generating diverse and natural 3d human motions from text. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5152–5161 (2022)
12. Guo, C., Zuo, X., Wang, S., Cheng, L.: Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts. In: European Conference on Computer Vision. pp. 580–597. Springer (2022)
13. Guo, C., Zuo, X., Wang, S., Zou, S., Sun, Q., Deng, A., Gong, M., Cheng, L.: Action2motion: Conditioned generation of 3d human motions. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 2021–2029 (2020)
14. Hassan, M., Ceylan, D., Villegas, R., Saito, J., Yang, J., Zhou, Y., Black, M.J.: Stochastic scene-aware motion prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11374–11384 (2021)
15. Hassan, M., Choutas, V., Tzionas, D., Black, M.J.: Resolving 3d human pose ambiguities with 3d scene constraints. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2282–2292 (2019)
16. Hassan, M., Ghosh, P., Tesch, J., Tzionas, D., Black, M.J.: Populating 3d scenes by learning human-scene interaction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14708–14718 (2021)
17. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems **33**, 6840–6851 (2020)
18. Hong, R., Liu, D., Mo, X., He, X., Zhang, H.: Learning to compose and reason with language tree structures for visual grounding. IEEE transactions on Pattern Analysis and Machine Intelligence **44**(2), 684–696 (2019)
19. Hong, Y., Zhen, H., Chen, P., Zheng, S., Du, Y., Chen, Z., Gan, C.: 3d-llm: Injecting the 3d world into large language models. Advances in Neural Information Processing Systems **36**, 20482–20494 (2023)
20. Hsu, J., Mao, J., Wu, J.: Ns3d: Neuro-symbolic grounding of 3d objects and relations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2614–2623 (2023)

21. Hu, R., Rohrbach, M., Andreas, J., Darrell, T., Saenko, K.: Modeling relationships in referential expressions with compositional modular networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1115–1124 (2017)
22. Huang, S., Wang, Z., Li, P., Jia, B., Liu, T., Zhu, Y., Liang, W., Zhu, S.C.: Diffusion-based generation, optimization, and planning in 3d scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16750–16761 (2023)
23. Jiang, N., Zhang, Z., Li, H., Ma, X., Wang, Z., Chen, Y., Liu, T., Zhu, Y., Huang, S.: Scaling up dynamic human-scene interaction modeling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1737–1747 (2024)
24. Lee, H.Y., Yang, X., Liu, M.Y., Wang, T.C., Lu, Y.D., Yang, M.H., Kautz, J.: Dancing to music. Advances in Neural Information Processing Systems **32** (2019)
25. Lei, J., Genova, K., Kopanas, G., Snavely, N., Guibas, L.: Momaps: Semantics-aware scene motion generation with motion maps. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10022–10031 (2025)
26. Li, J., Clegg, A., Mottaghi, R., Wu, J., Puig, X., Liu, C.K.: Controllable human-object interaction synthesis. In: European Conference on Computer Vision. pp. 54–72. Springer (2024)
27. Li, M., Sigal, L.: Referring transformer: A one-step approach to multi-task visual grounding. Advances in Neural Information Processing Systems **34**, 19652–19664 (2021)
28. Mao, W., Hartley, R.I., Salzmann, M., et al.: Contact-aware human motion forecasting. Advances in Neural Information Processing Systems **35**, 7356–7367 (2022)
29. Mir, A., Puig, X., Kanazawa, A., Pons-Moll, G.: Generating continual human motion in diverse 3d scenes. In: 2024 International Conference on 3D Vision (3DV). pp. 903–913. IEEE (2024)
30. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A., Tzionas, D., Black, M.J.: 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. pp. 10975–10985 (2019)
31. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
32. Roh, J., Desingh, K., Farhadi, A., Fox, D.: Languagerefer: Spatial-language model for 3d visual grounding. In: Conference on Robot Learning. pp. 1046–1056. PMLR (2022)
33. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
34. Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-Or, D., Bermano, A.H.: Human motion diffusion model. arXiv preprint arXiv:2209.14916 (2022)
35. Wang, J., Xu, H., Xu, J., Liu, S., Wang, X.: Synthesizing long-term 3d human motion and interaction in 3d scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9401–9411 (2021)
36. Wang, J., Rong, Y., Liu, J., Yan, S., Lin, D., Dai, B.: Towards diverse and natural scene-aware 3d human motion synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20460–20469 (2022)

37. Wang, J., Yan, S., Dai, B., Lin, D.: Scene-aware generative network for human motion synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12206–12215 (2021)
38. Wang, L., Li, Y., Huang, J., Lazebnik, S.: Learning two-branch neural networks for image-text matching tasks. IEEE Transactions on Pattern Analysis and Machine Intelligence **41**(2), 394–407 (2018)
39. Wang, Y., Li, Y., Li, X., Wang, S.: Hsi-gpt: A general-purpose large scene-motion-language model for human scene interaction. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 7147–7157 (2025)
40. Wang, Z., Chen, Y., Jia, B., Li, P., Zhang, J., Zhang, J., Liu, T., Zhu, Y., Liang, W., Huang, S.: 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. pp. 433–444 (2024)
41. Wang, Z., Chen, Y., Liu, T., Zhu, Y., Liang, W., Huang, S.: Humanise: Language-conditioned human motion generation in 3d scenes. Advances in Neural Information Processing Systems **35**, 14959–14971 (2022)
42. Xing, C., Mao, W., Liu, M.: Scene-aware human motion forecasting via mutual distance prediction. In: European Conference on Computer Vision. pp. 128–144. Springer (2024)
43. Xing, C., Mao, W., Liu, M.: Interphys: Physics-aware human motion synthesis in a dynamic scene. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 30729–30739 (2026)
44. Xu, C., Tan, R.T., Tan, Y., Chen, S., Wang, X., Wang, Y.: Auxiliary tasks benefit 3d skeleton-based human motion prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9509–9520 (2023)
45. Yang, Y., Zhai, W., Luo, H., Cao, Y., Zha, Z.J.: Lemon: Learning 3d human-object interaction relation from 2d images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16284–16295 (2024)
46. Ye, S., Wang, Y., Li, J., Park, D., Liu, C.K., Xu, H., Wu, J.: Scene synthesis from human motion. In: SIGGRAPH Asia 2022 Conference Papers. pp. 1–9 (2022)
47. Ye, Y., Liu, L., Hu, L., Xia, S.: Neural3points: Learning to generate physically realistic full-body motion for virtual reality users. In: Computer Graphics Forum. vol. 41, pp. 183–194. Wiley Online Library (2022)
48. Zhang, J., Zhang, Y., Cun, X., Zhang, Y., Zhao, H., Lu, H., Shen, X., Shan, Y.: Generating human motion from textual descriptions with discrete representations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14730–14740 (2023)
49. Zhang, M., Cai, Z., Pan, L., Hong, F., Guo, X., Yang, L., Liu, Z.: Motiondiffuse: Text-driven human motion generation with diffusion model. IEEE transactions on Pattern Analysis and Machine Intelligence **46**(6), 4115–4128 (2024)
50. Zhang, Y., Hassan, M., Neumann, H., Black, M.J., Tang, S.: Generating 3d people in scenes without people. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6194–6204 (2020)
51. Zhao, K., Zhang, Y., Wang, S., Beeler, T., Tang, S.: Synthesizing diverse human motions in 3d indoor scenes. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14738–14749 (2023)
52. Zhao, L., Cai, D., Sheng, L., Xu, D.: 3dvg-transformer: Relation modeling for visual grounding on point clouds. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2928–2937 (2021)