LLMDT: LLM-Based Dual-Transformer for Scene-Aware Text-Driven 3D Human Motion Generation

Anonymous authors
Paper under double-blind review

Abstract

Text-driven 3D human motion generation has a wide range of applications across various domains, while scene-aware generation requires joint reasoning about language, objects, navigation, and human-scene interactions. However, existing methods often struggle to separate global scene reasoning from local motion generation, resulting in inaccurate object grounding, trajectories, and human-scene interactions. We propose LLMDT, a hierarchical framework that addresses spatial reasoning, scene consistency, and motion quality in scene-aware human motion generation. We introduce LLM-based spatial reasoning and object-centric scene understanding to establish structured relationships between human actions and scene elements to improve spatial accuracy. We develop a trajectory-aware generation strategy that integrates global navigation with local scene context to improve scene consistency and generate feasible motion trajectories in complex environments. We further propose a pose-aware motion generation strategy with hierarchical residual representations to improve motion quality by generating realistic and temporally coherent human motions. Experiments on HUMANISE and HumanML3D demonstrate improvements over existing methods in spatial accuracy, scene consistency, and motion quality.

1 Introduction

Human motion generation has gained attention due to its applications in virtual reality, gaming, animation, digital twins, robotics, and embodied artificial intelligence (Li et al., 2025; Dziadosz et al., 2024; Cella et al., 2026; Tran et al., 2025). Early approaches based on generative adversarial networks (GANs) and variational autoencoders (VAEs) demonstrate the potential of deep generative models but often struggle with training stability, diversity, and long-range temporal dependencies (Men et al., 2022; Petrovich et al., 2021). Text-driven motion generation has subsequently emerged as an active research direction, aiming to synthesize 3D human motions from natural language. Early methods explore language-pose alignment and motion retrieval (Messina et al., 2023; Wang et al., 2022; Petrovich et al., 2023), while later approaches learn bidirectional text-motion mappings using tokenized representations, Transformers, and compact motion embeddings (Guo et al., 2022b;a; Zhang et al., 2023a; Petrovich et al., 2021). Diffusion models further advance motion synthesis by modeling complex motion distributions and supporting controllable generation (Tevet et al., 2023; Chen et al., 2023; Yuan et al., 2023; Wei et al., 2024; Zhang et al., 2024; Wang et al., 2019). However, diffusion over high-dimensional motion sequences remains computationally expensive, motivating latent and discrete motion representations (Hong et al., 2025; Guo et al., 2024).

Recent methods perform diffusion in compact latent spaces or discrete motion representations to improve computational efficiency (Hong et al., 2025; Petrovich et al., 2023). T2M-GPT (Zhang et al., 2023a) and MoMask (Guo et al., 2024) demonstrate that discrete tokens can capture high-level motion structures. However, these approaches mainly focus on text-motion mappings and largely overlook the surrounding 3D environment. They struggle to reason about spatial relationships, object interactions, and feasible motion trajectories in complex scenes. Existing studies have extended text-driven motion generation by incorporating scene context, motion intent, and object affordances. HUMANISE (Wang et al., 2022) presents text-driven motion generation in 3D scenes, while the approach in (Song et al., 2026) considers high-level motion intent from language. Afford-Motion (Wang et al., 2024) utilizes scene affordances for object-aware motion generation, and the method in (Cen et al., 2024) explores scene-aware motion synthesis in complex indoor environments. However, existing

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

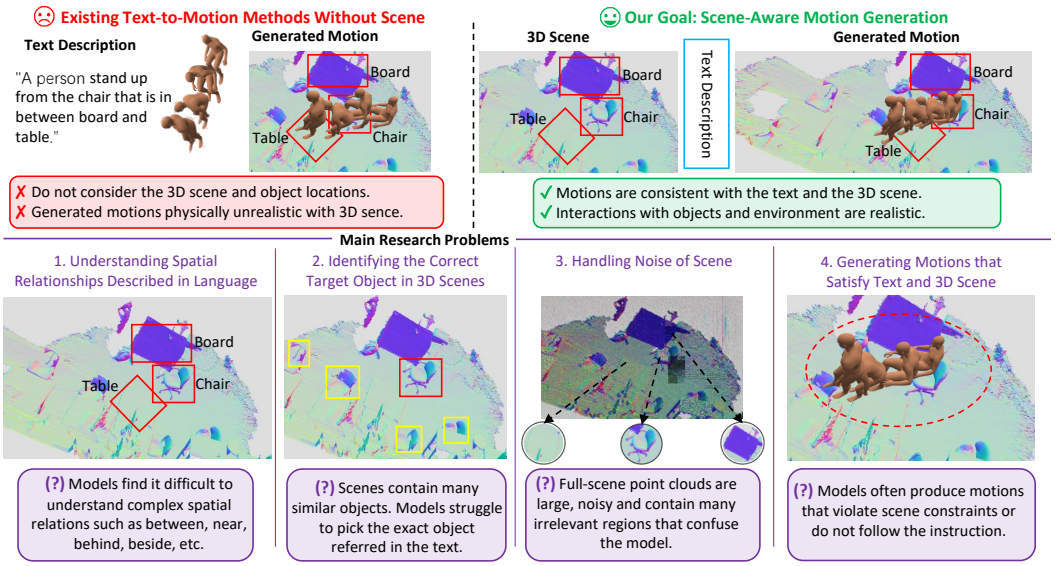


Figure 1: Existing problems and the goal of the proposed methodology to address these problems.

scene-aware motion methods often struggle to separate global scene reasoning from local motion generation, leading to inaccurate object grounding, trajectories, and human-scene interactions.

To address these problems, we propose LLMDT, which separates semantic grounding, global navigation, dynamic local scene sensing, and hierarchical pose synthesis, enabling each stage to operate at an appropriate spatial and temporal scale, as shown in Figure 1. LLMDT integrates LLM-based spatial reasoning with Transformer-based scene understanding to jointly model language, 3D environments, and human motion. LLM extracts structured semantics, including actions, target and anchor objects, and spatial relations, while an object-centric representation provides fine-grained scene understanding. A trajectory Transformer then generates scene-consistent navigation paths, and a pose-aware dual Transformer generates smooth and temporally consistent human motions. We employ hierarchical residual vector quantization to further improve motion realism by capturing both global motion patterns and fine-grained motion details. We evaluate LLMDT on the HUMANISE (Wang et al., 2022) and HumanML3D (Guo et al., 2022a) datasets. We additionally evaluate its cross-dataset generalization on the PROX (Hassan et al., 2019) dataset without fine-tuning. Experimental results demonstrate that LLMDT outperforms state-of-the-art methods in terms of spatial accuracy, scene consistency, and motion quality. The main contributions are summarized as follows:

- We introduce LLM-based spatial reasoning and object-centric scene understanding to establish structured relationships between human actions and scene elements, improving spatial accuracy in scene-aware motion generation.
- We design trajectory- and pose-aware Transformers with hierarchical residual vector quantization to generate scene-consistent navigation paths and realistic, temporally coherent human motions, improving scene consistency and motion quality.
- Experiments on the HUMANISE and HumanML3D datasets demonstrate that LLMDT outperforms existing methods in spatial accuracy, scene consistency, and motion quality.

2 Related Work

2.1 Text-Driven Human Motion Generation

Text-driven human motion generation synthesizes human motions from natural language descriptions (Liu et al., 2026; Wang et al., 2025a). Early approaches explore conditional generative models (Guo et al., 2020), followed by Transformer-based methods that learn discrete motion representa-

Table 1: Comparison of LLMDT with existing representative scene-aware and text-driven human motion generation methods.

Method	LLM/MLLM Reasoning	Object Grounding	Trajectory Planning	Local Scene Sensing	Hierarchical Pose Generation	RVQ
HUMANISE (Wang et al., 2022)	–	✓	–	–	–	–
Afford-Motion (Wang et al., 2024)	–	✓	–	–	–	–
Cen *et al.* (Cen et al., 2024)	✓	✓	✓	–	–	–
IntentMotion (Song et al., 2026)	–	✓	✓	–	–	–
HSI-GPT (Wang et al., 2025b)	✓	✓	–	–	–	–
SceMoS (Ghosh et al., 2026)	–	✓	✓	–	✓	–
MLLM Planning (Qian et al., 2025)	✓	✓	✓	–	–	–
TSTMotion (Guo et al., 2025)	–	✓	–	–	–	–
UniHM (Geng et al., 2025)	–	✓	–	–	✓	–
LLMDT (Ours)	✓	✓	✓	✓	✓	✓

tions and text-motion alignment, including T2M (Guo et al., 2022a), TM2T (Guo et al., 2022b), and T2M-GPT (Zhang et al., 2023a). Diffusion-based methods such as MDM (Tevet et al., 2023) and MotionDiffuse (Zhang et al., 2024) further improve motion quality and diversity, while MoMask (Guo et al., 2024) and SALAD (Hong et al., 2025) explore body-part-aware and latent representations. Despite these advances, most methods generate motions without modeling scene geometry and human-scene interactions (Hong et al., 2025; Gao et al., 2024; Chen et al., 2023). The proposed framework extends text-driven motion generation to scene-aware scenarios by factorizing semantic grounding, navigation, local scene sensing, and pose synthesis.

2.2 Scene-Aware Human Motion Generation

Scene-aware motion generation incorporates environmental geometry and human-object interactions to produce physically and semantically plausible motions. HUMANISE (Wang et al., 2022) jointly models textual instructions and 3D scenes, while Afford-Motion (Wang et al., 2024) introduces scene affordances as an intermediate representation for motion generation. The method in (Cen et al., 2024) employs LLM-based language grounding and object-centric scene representations, followed by trajectory and motion generation. TeSMo (Song et al., 2026) explores text-controlled human interaction motions using diffusion models, while recent approaches further investigate scene-aware motion generation through training-free foundation-model guidance (Guo et al., 2025), unified motion-language modeling (Geng et al., 2025), and large scene-motion-language models such as HSI-GPT (Wang et al., 2025b). SceMoS (Ghosh et al., 2026) further separates global planning from local execution using geometry-grounded scene representations. Despite these advances, jointly reasoning about semantics, navigation, scene context, and fine-grained motion remains challenging. The proposed trajectory-conditioned local scene sensing provides frame-wise scene context along the generated trajectory, enabling the generated motion to better adapt to local scene geometry and improving scene consistency.

2.3 Trajectory Planning and Motion Synthesis

Realistic scene-aware motion requires coherent navigation together with temporally consistent body poses. Existing methods explore spatio-temporal motion modeling (Su et al., 2023; Dang et al., 2021), interaction-aware generation (Qiao et al., 2024), and physically constrained synthesis, such as PhysDiff (Yuan et al., 2023) and InterGen (Liang et al., 2024). Recent scene-aware approaches increasingly separate high-level planning from low-level motion generation. The method proposed in (Cen et al., 2024) first generates human trajectories before synthesizing local poses, while SceMoS (Ghosh et al., 2026) disentangles global planning from local execution. MLLM-based planning also explores open-ended textual instructions in 3D scenes (Qian et al., 2025). However, existing methods lack frame-wise local scene sensing along the trajectory. The proposed LLMDT combines global trajectory planning, local scene sensing, and pose-aware dual Transformers to generate scene-consistent and temporally coherent motions, as summarized in Table 1.

3 Proposed Methodology

We propose LLMDT, a hierarchical framework for scene-aware text-driven 3D human motion generation that factorizes semantic grounding, global navigation, trajectory-conditioned local scene sensing, and fine-grained pose synthesis, as shown in Figure 2.

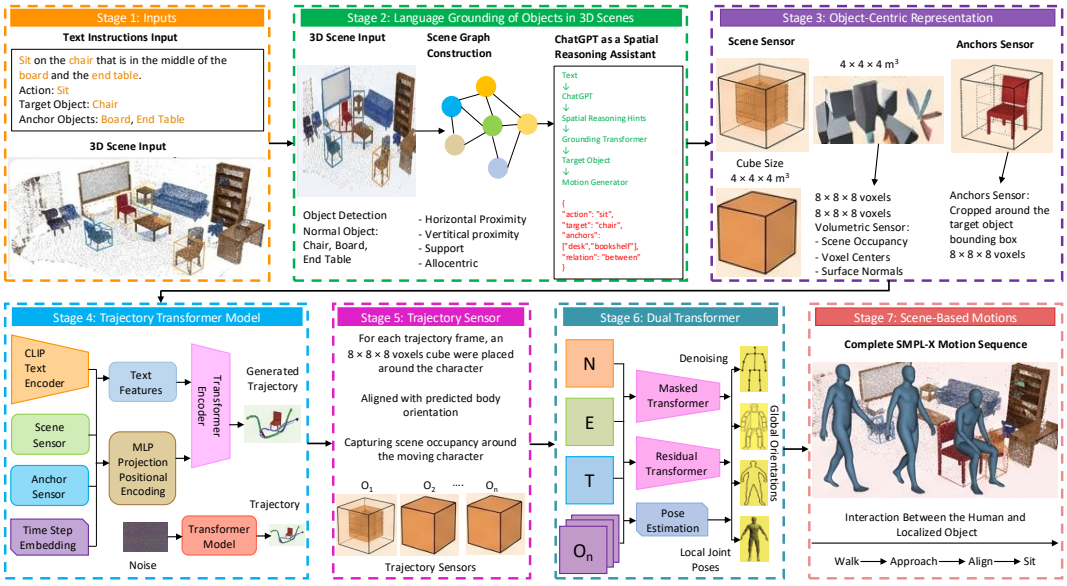


Figure 2: Overview of the proposed methodology.

3.1 LLM-Based Semantic Reasoning and Object Grounding

Natural language instructions often contain implicit actions, object references, and spatial relationships that are difficult to resolve from a sentence-level representation. We therefore use a pretrained LLM to transform an instruction $\mathcal{T}$ into structured semantic representations that capture the action, relevant scene objects, and their spatial relationships. In parallel, objects in the input scene are detected with their semantic labels and geometric attributes. The detected objects are represented as a scene graph $G = (V, E)$, where nodes denote object instances and edges encode spatial relations such as proximity, support, containment, and relative direction. The LLM-derived semantics are combined with the scene graph to identify the most likely interaction object. Specifically, the grounded object is selected as:

$$o^* = \arg\max_{o_i \in V} \mathcal{J}(o_i), \tag{1}$$

where $\mathcal{J}(o_i)$ denotes the grounding score of candidate object o_i. The score integrates semantic confidence, language-object similarity, and relational consistency:

$$\mathcal{J}(o_i) = \alpha P(o_i|\mathcal{T}) + \beta\psi(e_T, e_{o_i}) + \gamma\phi(G, e_R, o_i), \tag{2}$$

where $P(o_i|\mathcal{T})$ denotes the semantic confidence of candidate o_i, e_T and e_{o_i} are language and object embeddings, $\psi(\cdot)$ measures language-object similarity, and $\phi(\cdot)$ evaluates relational consistency with the scene graph. The coefficients α, β, and γ balance the three terms. The grounded object o^* provides the interaction target for subsequent scene representation and trajectory planning.

3.2 Object-Centric Scene Representation

A complete 3D scene contains substantial geometric information, much of which is irrelevant to the intended interaction. We therefore construct an object-centric representation around the grounded target object to reduce redundant scene information and focus on interaction-relevant geometry. After grounding, we extract a $4 \times 4 \times 4$ m^3 region centered on the target object and discretize it into an $8 \times 8 \times 8$ voxel grid. Each voxel contains occupancy, spatial coordinates, and local surface normals. We further construct an anchor sensor around the target object's bounding box to capture fine-grained interaction geometry. These representations provide complementary scene-level and

local geometric information. The two representations are independently encoded and projected into a unified feature space:

$$F_{obj} = \sigma\left(W_f\left[\mathrm{Enc}(V_s, N, C) \oplus \mathrm{Enc}(V_a, N_a, C_a)\right] + b_f\right), \tag{3}$$

where V_s and V_a denote scene and anchor-sensor voxel occupancies, N and N_a are surface normals, and C and C_a are voxel-center coordinates. $\mathrm{Enc}(\cdot)$ denotes the geometric encoder, $\oplus$ denotes feature concatenation, and W_f, b_f, and $\sigma(\cdot)$ denote learnable projections and nonlinear activation, respectively. The resulting F_{obj} captures environmental structure and interaction-specific geometry while filtering out irrelevant scene information.

3.3 Trajectory Planning and Conditioned Local Scene Sensing

We separate global navigation from fine-grained pose synthesis by jointly modeling trajectory planning and trajectory-conditioned local scene sensing. The trajectory Transformer predicts a scene-consistent path from the initial human position to the grounded interaction location based on the semantic representation and object-centric scene features. The predicted trajectory then determines the locations at which local scene geometry is queried, providing frame-wise geometric observations for subsequent pose synthesis. We formulate trajectory generation as conditional diffusion. The trajectory Transformer progressively denoises a trajectory latent under semantic and geometric conditions, starting from a noisy latent $z_T \sim \mathcal{N}(0, I)$:

$$\hat{\Gamma} = D_\theta\left(z_T, [E_L\ F_{obj}\ E_\tau]\right), \tag{4}$$

where D_θ denotes the trajectory denoising network, E_L is the LLM-derived semantic embedding, F_{obj} is the object-centric scene feature, and E_τ is the diffusion timestep embedding. The Transformer uses temporal self-attention to model dependencies between trajectory points and cross-attention to integrate semantic and geometric conditions. The predicted trajectory is represented as:

$$\hat{\Gamma} = p_1, p_2, \ldots, p_N, \qquad p_i = (x_i, y_i, \theta_i), \tag{5}$$

where (x_i, y_i) denotes the human position and θ_i denotes the body orientation at waypoint i. This trajectory provides global spatial guidance from the initial position toward the grounded interaction location. The predicted trajectory is further used to obtain local geometric observations along the planned path. The cubic voxel region centered at the human pelvis is extracted at each waypoint p_i and aligned with the predicted body orientation. Each voxel contains occupancy, surface normals, and spatial coordinates. The resulting trajectory-conditioned observations are defined as:

$$O = \left(\mathcal{V}(\mathcal{S}, p_i, R_i)\right)_{i=1}^{N}, \tag{6}$$

where $\mathcal{V}(\cdot)$ denotes the voxel sampling operator, $\mathcal{S}$ is the input scene, p_i is the predicted body position, and R_i represents its orientation. Therefore, O provides a sequence of local geometric observations corresponding to successive locations along the predicted trajectory. The trajectory and local observations are subsequently provided to the pose-aware dual Transformer together with the semantic and object-centric representations. Therefore, $\hat{\Gamma}$ provides global navigation guidance, while O captures fine-grained scene geometry along the planned path. This joint representation links global trajectory planning with local geometric reasoning for scene-consistent motion synthesis.

3.4 Pose-Aware Dual Transformer

The pose-aware dual Transformer generates fine-grained human motion while preserving temporal coherence and scene consistency. High-dimensional motion sequence prediction remains challenging due to the complex dependencies between global body dynamics and local joint articulations. We therefore use complementary masked and residual branches for coarse-to-fine motion generation, as shown in Figure 3. The masked Transformer first predicts coarse motion tokens from partially masked motion sequences, capturing global action structure and temporal dependencies.

The residual Transformer then refines these representations to recover fine-grained dynamics, including body orientation, hand-object interaction, and foot adjustments. Hierarchical residual vector quantization (RVQ) provides discrete representations at multiple levels of motion detail. The masked Transformer predicts coarse motion tokens given the conditioning representation $[E_L; F_{obj}; O; \hat{\Gamma}]$:

$$Q^{(0)} = \mathcal{T}_M \left(X_M, \mathrm{CA} \left(X_M, \left[E_L \; F_{obj} \; O \; \hat{\Gamma} \right] \right) \right), \tag{7}$$

where $\mathcal{T}_M$ denotes the masked Transformer and X_M is the partially masked motion representation. The residual Transformer progressively predicts the residual representation at each RVQ level:

$$\Delta Q^{(l)} = \mathcal{T}_R^{(l)} \left(Q^{(l-1)}, \left[O \; \hat{\Gamma} \right] \right), \qquad l = 1, \ldots, L, \tag{8}$$

where $\Delta Q^{(l)}$ denotes the predicted residual at level l. The corresponding hierarchical representation is reconstructed through the RVQ codebooks:

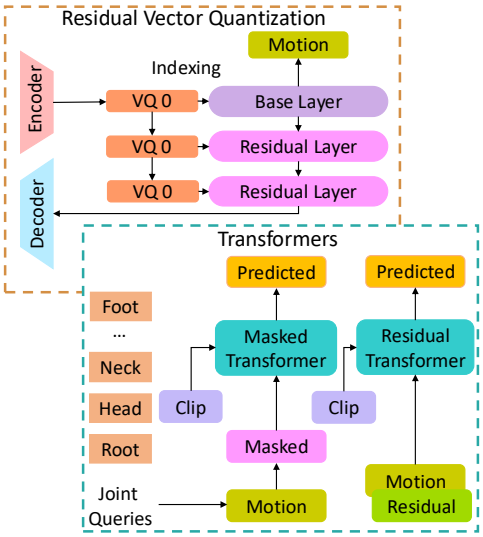


Figure 3: Pose-aware dual Transformer for coarse-to-fine motion generation.

$$Z = \sum_{l=0}^{L} \mathcal{C}_l \left(Q^{(l)} \right), \tag{9}$$

where $\mathcal{C}_l(\cdot)$ denotes the l-th RVQ codebook. Lower levels capture coarse motion structure, while higher levels progressively encode finer motion details. The reconstructed motion representation is decoded into SMPL-X parameters:

$$\hat{M} = \mathcal{D}_{\mathrm{SMPL-X}} \left(Z, \hat{\Gamma} \right), \tag{10}$$

where $\mathcal{D}_{\mathrm{SMPL-X}}(\cdot)$ produces the corresponding body pose and joint rotations. The predicted trajectory provides global spatial guidance, while the trajectory-conditioned observations provide local geometric context for scene-consistent motion synthesis.

4 Experiments

4.1 Implementation Details

We evaluate scene-aware motion generation on HUMANISE, which provides language descriptions, 3D scene geometry, and human motion, and general text-to-motion generation on HumanML3D, which provides paired text-motion data. Each scene $\mathcal{S} \in R^{H \times W \times D}$ is converted into an object-centric voxel representation with $8 \times 8 \times 8$ resolution, where each voxel encodes occupancy, surface normals, and spatial coordinates. A pretrained large language model transforms an instruction $\mathcal{T}$ into structured semantic representations capturing the action, relevant scene objects, and their spatial relationships. The trajectory Transformer predicts a navigation trajectory $\hat{\Gamma} \in R^{N \times 3}$, where each waypoint $p_i = (x_i, y_i, \theta_i)$ represents the human position and orientation. A trajectory-conditioned local scene sensing module then queries local scene geometry along the predicted path to obtain frame-wise observations $O = o_1, \ldots, o_N$ for motion synthesis.

The pose-aware dual Transformer uses hierarchical residual vector quantization with L discrete codebooks to encode a motion sequence $M \in R^{T \times J \times d}$ into hierarchical token representations $Q = Q^{(0)}, \ldots, Q^{(L)}$. The Transformer modules employ multi-head self- and cross-attention followed

Table 2: Performance comparison on the HUMANISE dataset. The best results are written in red and the second-best in blue.

Methods	Venue	Scene-conditional			Action-conditional			Pure Motion Quality	
		Goal Dist. ↓	Contact ↑	Scene-consistent ↑	Accuracy ↑	Diversity ↑	Multimodality ↑	FID ↓	Quality Score ↑
HUMANISE (Wang et al., 2022)	NeurIPS	$0.322_{\pm.004}$	$79.54_{\pm.006}$	$99.86_{\pm.003}$	$0.889_{\pm.003}$	$4.318_{\pm.051}$	$2.721_{\pm.039}$	$1.126_{\pm.046}$	$5.472_{\pm.045}$
MDM (Tevet et al., 2023)	ICLR	$0.081_{\pm.001}$	$94.61_{\pm.006}$	$99.71_{\pm.004}$	$0.898_{\pm.005}$	$4.187_{\pm.110}$	$1.964_{\pm.106}$	$0.091_{\pm.108}$	$5.271_{\pm.098}$
Afford-Motion (Wang et al., 2024)	CVPR	$0.139_{\pm.003}$	$94.18_{\pm.005}$	$99.84_{\pm.004}$	$0.913_{\pm.007}$	$4.391_{\pm.109}$	$2.467_{\pm.101}$	$0.592_{\pm.117}$	$4.821_{\pm.107}$
Cen et al. (Cen et al., 2024)	CVPR	$0.137_{\pm.002}$	$89.21_{\pm.003}$	$99.91_{\pm.006}$	$0.930_{\pm.005}$	$4.821_{\pm.116}$	$2.365_{\pm.099}$	$0.391_{\pm.111}$	$5.281_{\pm.102}$
Song et al. (Song et al., 2026)	AAAI	$0.376_{\pm.003}$	$96.82_{\pm.005}$	$98.77_{\pm.006}$	$0.966_{\pm.007}$	$4.931_{\pm.118}$	$2.472_{\pm.103}$	$0.173_{\pm.115}$	$5.932_{\pm.109}$
Ours	-	$0.053_{\pm.001}$	$96.48_{\pm.003}$	$99.95_{\pm.004}$	$0.986_{\pm.005}$	$4.824_{\pm.111}$	$2.566_{\pm.100}$	$0.079_{\pm.106}$	$5.941_{\pm.112}$

Table 3: Performance comparison on the HumanML3D dataset. The best results are written in red and the second-best in blue.

Methods	Venue	R-Precision ↑			FID ↓	MM DIST ↓	Diversity ↑	MultiModality ↑
		Top-1	Top-2	Top-3				
MLD Chen et al. (2023)	CVPR	$0.485_{\pm.002}$	$0.678_{\pm.003}$	$0.777_{\pm.004}$	$0.482_{\pm.012}$	-	$9.748_{\pm.095}$	$2.432_{\pm.069}$
MDM Tevet et al. (2023)	ICLR	-	-	$0.612_{\pm.005}$	$0.548_{\pm.039}$	$5.598_{\pm.033}$	$9.581_{\pm.093}$	$2.805_{\pm.070}$
ReMoDiffuse Zhang et al. (2023b)	ICCV	$0.511_{\pm.002}$	$0.704_{\pm.004}$	$0.792_{\pm.007}$	$0.108_{\pm.009}$	$2.971_{\pm.014}$	$9.045_{\pm.070}$	$1.806_{\pm.036}$
DT3DPE Abbas et al. (2026)	CAD	$0.527_{\pm.005}$	$0.721_{\pm.007}$	$0.811_{\pm.006}$	$0.049_{\pm.006}$	$2.968_{\pm.003}$	$9.689_{\pm.093}$	$1.249_{\pm.051}$
Ours	-	$0.535_{\pm.006}$	$0.728_{\pm.006}$	$0.823_{\pm.007}$	$0.058_{\pm.007}$	$2.959_{\pm.003}$	$9.701_{\pm.091}$	$1.257_{\pm.053}$

by two-layer feed-forward networks with GELU activation. LLMDT is trained with a joint objective consisting of trajectory diffusion loss $\mathcal{L}_{traj}$, masked motion-token prediction loss $\mathcal{L}_{mask}$, hierarchical residual-token prediction loss $\mathcal{L}_{res}$, and motion reconstruction loss $\mathcal{L}_{rec}$. The total loss is $\mathcal{L}_{total} = \lambda_{traj}\mathcal{L}_{traj} + \lambda_{mask}\mathcal{L}_{mask} + \lambda_{res}\mathcal{L}_{res} + \lambda_{rec}\mathcal{L}_{rec}$. We train the model using AdamW with an initial learning rate of 1×10^{-4} and cosine learning-rate decay. Detailed formulations of the individual loss terms and the training procedure are provided in the supplementary material. Experiments are conducted on a workstation with two NVIDIA RTX 3090 GPUs using Python, PyTorch, CUDA, SMPL-X, and Hugging Face Transformers. ChatGPT is used for semantic and spatial reasoning.

4.2 Quantitative Results

The quantitative comparison on the HUMANISE dataset demonstrates the effectiveness of the proposed LLMDT framework in scene understanding, semantic alignment, and motion quality, as presented in Table 2. In the scene-conditional evaluation, LLMDT achieves the lowest goal distance of 0.053, improving over the previous best of 0.081 by MDM, indicating accurate localization of the target interaction position. It also obtains the highest scene-consistent rate of 99.95%, outperforming Cen *et al.* (99.91%) and HUMANISE (99.86%), demonstrating effective obstacle avoidance. The contact score of 96.48% is also competitive with the best result of 96.82% achieved by Song *et al.*. LLMDT achieves the highest action accuracy of 0.986, outperforming the previous best of 0.966 and demonstrating improved semantic consistency in action-conditioned motion generation. It also achieves a diversity score of 4.824 and multimodality of 2.566, indicating that the generated motions remain diverse while preserving multiple realistic motion modes. Furthermore, LLMDT obtains the lowest FID of 0.079 and the highest quality score of 5.941, slightly exceeding Song *et al.* (5.932). The results on the HumanML3D dataset, shown in Table 3, further demonstrate LLMDT's effectiveness. Our method achieves the best Top-1 and Top-3 R-Precision, MM DIST, and FID scores, while maintaining competitive multimodality and diversity, showing text-motion alignment and the ability to generate diverse and realistic motions.

4.3 Qualitative Results

The proposed method synthesizes diverse human motions across various indoor environments, including standing, walking, sitting, and lying. LLMDT interprets complex spatial expressions such as *"far away from the TV"*, *"near the trash can"*, *"in front of the bookshelf"*, and *"farthest from the door"*, identifying the intended interaction object among multiple candidates, as shown in Figure 4. The generated motions navigate toward the target while preserving the specified spatial relationships. The scene representation and trajectory Transformer guide scene-consistent navigation, while the dual Transformer generates natural poses and realistic human-scene interactions. Figure 5 presents qualitative comparisons with state-of-the-art methods. Existing methods produce incorrect object identification, inaccurate interaction locations, or unrealistic poses for challenging instructions such as *"Stand up from the chair that is in between the board and the table"*. In contrast, LLMDT

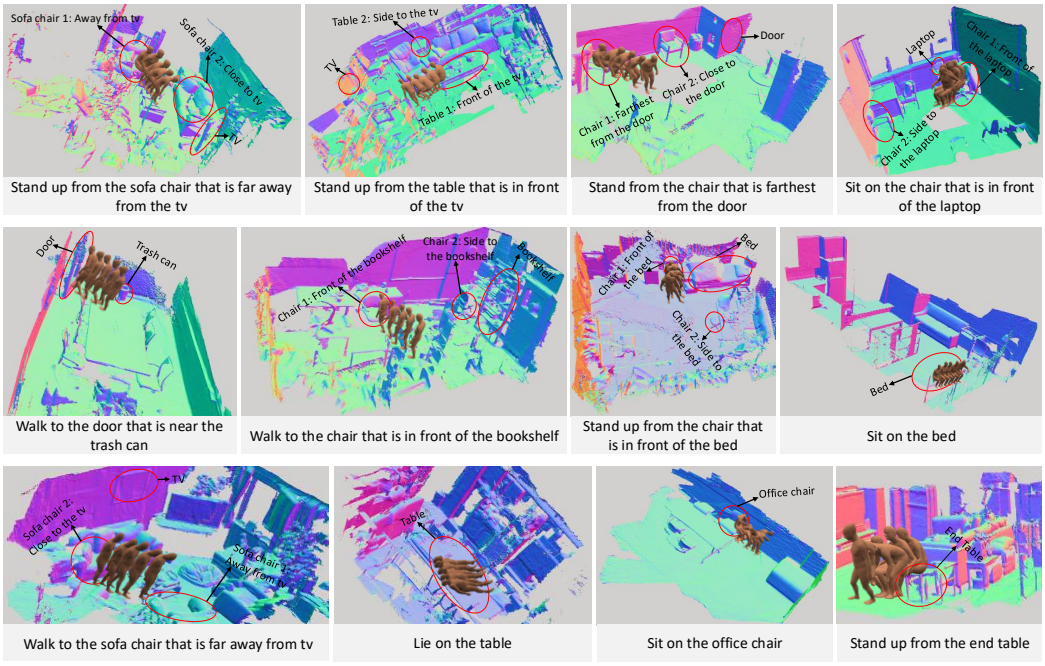


Figure 4: Qualitative results of the proposed LLMDT on the HUMANISE dataset.

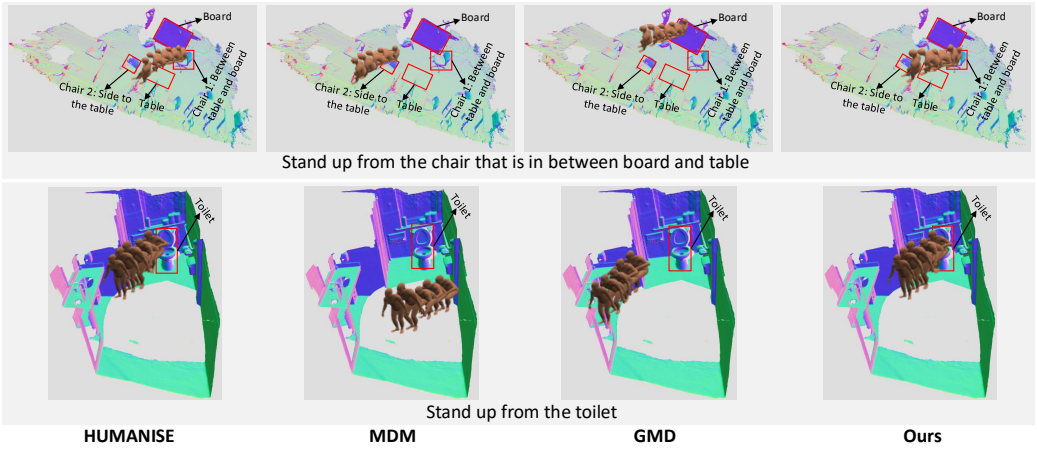


Figure 5: Qualitative comparison with existing methods on the HUMANISE dataset.

jointly reasons over semantic and scene cues to identify the interaction object and generate precise motions. LLMDT generates smoother transitions, consistent postures, and realistic scene contact than HUMANISE, MDM, and GMD for *"Stand up from the toilet"*. We further evaluate the cross-dataset generalization of LLMDT on the PROX dataset without fine-tuning, as shown in Figure 6.

4.4 Ablation Study

We conduct an ablation study by removing each proposed component, as presented in Table 4. Removing LLM-based semantic reasoning causes the largest degradation, increasing the goal distance from 0.053 to 0.112 and reducing action accuracy from 0.986 to 0.947, highlighting the importance of semantic reasoning and object grounding. Removing the object-centric scene representation increases the goal distance to 0.097 and FID to 0.184, demonstrating the benefit of localized geometric features. Removing the trajectory Transformer reduces the scene-consistent rate to 99.71%,

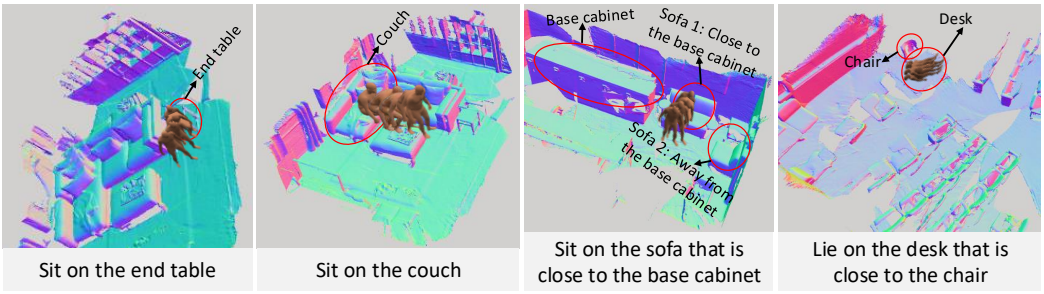


Figure 6: Qualitative results on the PROX dataset.

Table 4: Ablation study of the proposed LLMDT framework on the HUMANISE dataset.

Method	Goal Dist.↓	Accuracy↑	Scene-consistent↑	FID↓	Quality↑
w/o LLM Semantic Reasoning	0.112	0.947	99.79	0.231	5.624
w/o Object-Centric Scene Rep.	0.097	0.958	99.83	0.184	5.701
w/o Trajectory Transformer	0.088	0.972	99.71	0.146	5.786
w/o Trajectory Sensor	0.072	0.978	99.81	0.119	5.852
w/o Pose-Aware Dual Transformer	0.069	0.975	99.88	0.127	5.831
LLMDT (Full Model)	**0.053**	**0.986**	**99.95**	**0.079**	**5.941**

showing the importance of global trajectory planning. Removing the trajectory sensor further degrades performance, increasing the goal distance to 0.072 and FID to 0.119, demonstrating the benefit of trajectory-conditioned local scene sensing. Removing the pose-aware dual Transformer reduces action accuracy to 0.975, increases FID to 0.127, and decreases the quality score to 5.831, demonstrating the contribution of hierarchical motion refinement. The full LLMDT achieves the best performance, showing the complementary contributions of the proposed components. Figure 7 further visualizes the effect of individual components using the instruction *"Walk to the door that is near the trash can"*. Removing LLM semantic reasoning, object-centric scene representation, trajectory Transformer, trajectory sensor, or pose-aware dual Transformer leads to less accurate navigation or less coherent motion, whereas the full LLMDT generates the most consistent result.

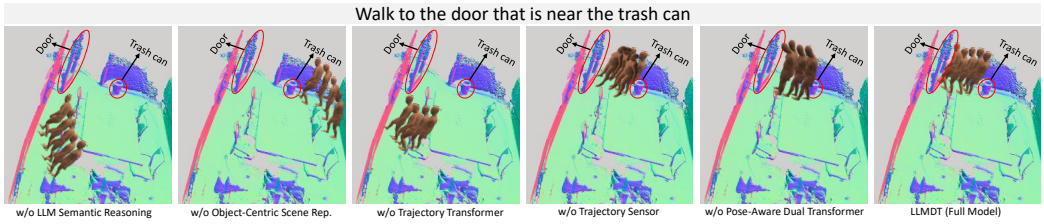


Figure 7: Qualitative comparison of individual component ablations.

5 Conclusion and Future Work

In this paper, we proposed LLMDT, a hierarchical framework for scene-aware text-driven 3D human motion generation. LLMDT integrates LLM-based spatial reasoning, object-centric scene understanding, trajectory-aware generation, and pose-aware motion synthesis to jointly address spatial accuracy, scene consistency, and motion quality. Experiments on HUMANISE and HumanML3D demonstrate competitive performance with existing methods.

Limitations and Future Work LLMDT remains limited by long-horizon language reasoning, computational cost, and generalization to challenging human-scene configurations. It mainly considers static indoor scenes and does not model dynamic changes or multi-agent interactions. Future work will explore lightweight reasoning, dynamic environments, and richer human-scene interactions.

5.1 AI Use Statement

In this work, ChatGPT was used as a spatial reasoning assistant within the proposed framework to interpret textual instructions and assist in extracting semantic and spatial relationships among objects. Generative AI tools were used to assist with grammar correction, language editing, text refinement, and improving the clarity and readability of the manuscript. Generative AI tools were not used to generate experimental results, perform quantitative analyses, create datasets, or produce figures. The model architecture, research methodology, experimental design, code implementation, evaluation procedures, and scientific conclusions were developed and verified by the authors. All AI-assisted outputs were carefully reviewed, edited, and validated by the authors to ensure their accuracy and consistency with the research. The authors take full responsibility for the final content of this work, including all scientific claims, analyses, results, figures, datasets, and text.

References

Rizwan Abbas, Hua Gao, and Xi Li. Text-driven 3d human motion generation for pose estimation using dual-transformer architecture. *Computer-Aided Design*, 190:103991, 2026. ISSN 0010-4485. doi: https://doi.org/10.1016/j.cad.2025.103991. URL https://www.sciencedirect.com/science/article/pii/S0010448525001526.

Christian Cella, M. Ristic, Marco Faroni, Andrea M. Zanchettin, and Paolo Rocco. Adaptive human-robot collaborative painting combining preference-based optimization and dynamic motion primitives. *IEEE Robotics Autom. Lett.*, 11(6):7023–7030, 2026. doi: 10.1109/LRA.2026.3683596. URL https://doi.org/10.1109/LRA.2026.3683596.

Zhi Cen, Huaijin Pi, Sida Peng, Zehong Shen, Minghui Yang, Shuai Zhu, Hujun Bao, and Xiaowei Zhou. Generating human motion in 3d scenes from text descriptions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 1855–1866. IEEE, 2024. doi: 10.1109/CVPR52733.2024.00182. URL https://doi.org/10.1109/CVPR52733.2024.00182.

Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. Executing your commands via motion diffusion in latent space. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pp. 18000–18010. IEEE, 2023. doi: 10.1109/CVPR52729.2023.01726. URL https://doi.org/10.1109/CVPR52729.2023.01726.

Lingwei Dang, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li. MSR-GCN: multi-scale residual graph convolution networks for human motion prediction. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pp. 11447–11456. IEEE, 2021. doi: 10.1109/ICCV48922.2021.01127. URL https://doi.org/10.1109/ICCV48922.2021.01127.

Marcin Dziadosz, Mariusz Mazurek, Tomasz Rymarczyk, Dariusz Wójcik, and Pawel Olszewski. Poster abstract: A computer vision system for human motion monitoring on a bicycle trainer. In Jie Liu, Yuanchao Shu, Jiming Chen, Yuan He, and Rui Tan (eds.), *Proceedings of the 22nd ACM Conference on Embedded Networked Sensor Systems, SenSys 2024, Hangzhou, China, November 4-7, 2024*, pp. 859–860. ACM, 2024. doi: 10.1145/3666025.3699407. URL https://doi.org/10.1145/3666025.3699407.

Xuehao Gao, Yang Yang, Zhenyu Xie, Shaoyi Du, Zhongqian Sun, and Yang Wu. GUESS: gradually enriching synthesis for text-driven human motion generation. *IEEE Trans. Vis. Comput. Graph.*, 30(12):7518–7530, 2024. doi: 10.1109/TVCG.2024.3352002. URL https://doi.org/10.1109/TVCG.2024.3352002.

Zichen Geng, Zeeshan Hayder, Wei Liu, and Ajmal Mian. Unihm: Universal human motion generation with object interactions in indoor scenes. *CoRR*, abs/2505.12774, 2025. doi: 10.48550/ARXIV.2505.12774. URL https://doi.org/10.48550/arXiv.2505.12774.

Anindita Ghosh, Vladislav Golyanik, Taku Komura, Philipp Slusallek, Christian Theobalt, and Rishabh Dabral. Scemos: Scene-aware 3d human motion synthesis by planning with geometry-grounded tokens. *CoRR*, abs/2602.20476, 2026. doi: 10.48550/ARXIV.2602.20476. URL https://doi.org/10.48550/arXiv.2602.20476.

Chuan Guo, Xinxin Zuo, Sen Wang, Shihao Zou, Qingyao Sun, Annan Deng, Minglun Gong, and Li Cheng. Action2motion: Conditioned generation of 3d human motions. In Chang Wen Chen, Rita Cucchiara, Xian-Sheng Hua, Guo-Jun Qi, Elisa Ricci, Zhengyou Zhang, and Roger Zimmermann (eds.), *MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020*, pp. 2021–2029. ACM, 2020. doi: 10.1145/3394171.3413635. URL https://doi.org/10.1145/3394171.3413635.

Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 5142–5151. IEEE, 2022a. doi: 10.1109/CVPR52688.2022.00509. URL https://doi.org/10.1109/CVPR52688.2022.00509.

Chuan Guo, Xinxin Zuo, Sen Wang, and Li Cheng. TM2T: stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXV*, volume 13695 of *Lecture Notes in Computer Science*, pp. 580–597. Springer, 2022b. doi: 10.1007/978-3-031-19833-5_34. URL https://doi.org/10.1007/978-3-031-19833-5_34.

Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, and Li Cheng. Momask: Generative masked modeling of 3d human motions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 1900–1910. IEEE, 2024. doi: 10.1109/CVPR52733.2024.00186. URL https://doi.org/10.1109/CVPR52733.2024.00186.

Ziyan Guo, Haoxuan Qu, Hossein Rahmani, De Wen Soh, Ping Hu, Qiuhong Ke, and Jun Liu. Tstmotion: Training-free scene-aware text-to-motion generation. In *IEEE International Conference on Multimedia and Expo, ICME 2025, Nantes, France, June 30 - July 4, 2025*, pp. 1–6. IEEE, 2025. doi: 10.1109/ICME59968.2025.11209881. URL https://doi.org/10.1109/ICME59968.2025.11209881.

Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J. Black. Resolving 3d human pose ambiguities with 3d scene constraints. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pp. 2282–2292. IEEE, 2019. doi: 10.1109/ICCV.2019.00237. URL https://doi.org/10.1109/ICCV.2019.00237.

Seokhyeon Hong, Chaelin Kim, Serin Yoon, Junghyun Nam, Sihun Cha, and Junyong Noh. SALAD: skeleton-aware latent diffusion for text-driven motion generation and editing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 7158–7168. Computer Vision Foundation / IEEE, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Hong_SALAD_Skeleton-aware_Latent_Diffusion_for_Text-driven_Motion_Generation_and_Editing_CVPR_2025_paper.html.

Zhengli Li, Liantao Wang, and Xueqing Wu. Artificial intelligence based virtual gaming experience for sports training and simulation of human motion trajectory capture. *Entertain. Comput.*, 52: 100828, 2025. doi: 10.1016/J.ENTCOM.2024.100828. URL https://doi.org/10.1016/j.entcom.2024.100828.

Han Liang, Wenqian Zhang, Wenxuan Li, Jingyi Yu, and Lan Xu. Intergen: Diffusion-based multi-human motion generation under complex interactions. *Int. J. Comput. Vis.*, 132(9):3463–3483, 2024. doi: 10.1007/S11263-024-02042-6. URL https://doi.org/10.1007/s11263-024-02042-6.

Sheng Liu, Jinghao Cao, Xiangxiang Wang, Rui Zhao, Yang Li, and Sidan Du. Motionlifter: Unsupervised text-driven 3d human motion generation with semantic and layout self-supervision. *Neurocomputing*, 700:134379, 2026. doi: 10.1016/J.NEUCOM.2026.134379. URL https://doi.org/10.1016/j.neucom.2026.134379.

Qianhui Men, Hubert P. H. Shum, Edmond S. L. Ho, and Howard Leung. Gan-based reactive motion synthesis with class-aware discriminators for human-human interaction. *Comput. Graph.*, 102: 634–645, 2022. doi: 10.1016/J.CAG.2021.09.014. URL https://doi.org/10.1016/j.cag.2021.09.014.

Nicola Messina, Jan Sedmidubský, Fabrizio Falchi, and Tomás Rebok. Text-to-motion retrieval: Towards joint understanding of human motion data and natural language. In Hsin-Hsi Chen, Wei-Jou (Edward) Duh, Hen-Hsen Huang, Makoto P. Kato, Josiane Mothe, and Barbara Poblete (eds.), *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23-27, 2023*, pp. 2420–2425. ACM, 2023. doi: 10.1145/3539618.3592069. URL https://doi.org/10.1145/3539618.3592069.

Mathis Petrovich, Michael J. Black, and Gül Varol. Action-conditioned 3d human motion synthesis with transformer VAE. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pp. 10965–10975. IEEE, 2021. doi: 10.1109/ICCV48922.2021.01080. URL https://doi.org/10.1109/ICCV48922.2021.01080.

Mathis Petrovich, Michael J. Black, and Gül Varol. TMR: text-to-motion retrieval using contrastive 3d human motion synthesis. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pp. 9454–9463. IEEE, 2023. doi: 10.1109/ICCV51070.2023.00870. URL https://doi.org/10.1109/ICCV51070.2023.00870.

Siyi Qian, Jian Fang, Yuzhou Mao, Yayun Zou, Wentao Zhang, and Haiwei Xue. Human motion generation in 3d scenes from open-ended textual instructions with MLLM planning. In Cathal Gurrin, Klaus Schoeffmann, Min Zhang, Luca Rossetto, Stevan Rudinac, Duc-Tien Dang-Nguyen, Wen-Huang Cheng, Phoebe Chen, and Jenny Benois-Pineau (eds.), *Proceedings of the 33rd ACM International Conference on Multimedia, MM 2025, Dublin, Ireland, October 27-31, 2025*, pp. 9404–9413. ACM, 2025. doi: 10.1145/3746027.3754850. URL https://doi.org/10.1145/3746027.3754850.

Shanbao Qiao, Neal N. Xiong, Yongbin Gao, Zhijun Fang, Wenjun Yu, Juan Zhang, and Xiaoyan Jiang. Self-supervised learning of depth and ego-motion for 3d perception in human computer interaction. *ACM Trans. Multim. Comput. Commun. Appl.*, 20(2):45:1–45:21, 2024. doi: 10.1145/3588571. URL https://doi.org/10.1145/3588571.

Wenfeng Song, Shi Zheng, Xinyu Zhang, Xingliang Jin, Aimin Hao, Fei Hou, Xia Hou, and Shuai Li. Intentmotion: Learning intent-aware human motion from language in 3d scenes. In Sven Koenig, Chad Jenkins, and Matthew E. Taylor (eds.), *Fortieth AAAI Conference on Artificial Intelligence, Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence, Sixteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2026, Singapore, January 20-27, 2026*, pp. 2065–2073. AAAI Press, 2026. doi: 10.1609/AAAI.V40I3.37188. URL https://doi.org/10.1609/aaai.v40i3.37188.

Haoyu Su, Shenglan Liu, Zewen Gao, Yifeng Dong, Junshi Yang, and Suhao Ding. Multi-scale spatio-temporal aggregation network for human motion prediction. In *18th International Conference on Intelligent Systems and Knowledge Engineering, ISKE 2023, Fuzhou, China, November 17-19, 2023*, pp. 410–417. IEEE, 2023. doi: 10.1109/ISKE60036.2023.10481288. URL https://doi.org/10.1109/ISKE60036.2023.10481288.

Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit Haim Bermano. Human motion diffusion model. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=SJ1kSyO2jwu.

Thien Tran, Quang Nguyen, Toan Luu, Minh Tran, Jonathan Kua, Thuong Hoang, and Man Dien. Empowering robotic training with kinesthetic learning and digital twins in human-centric industrial systems. *J. Ind. Inf. Integr.*, 43:100743, 2025. doi: 10.1016/J.JII.2024.100743. URL https://doi.org/10.1016/j.jii.2024.100743.

Lichen Wang, Zhengming Ding, and Yun Fu. Low-rank transfer human motion segmentation. *IEEE Transactions on Image Processing*, 28(2):1023–1034, 2019. doi: 10.1109/TIP.2018.2870945.

Yin Wang, Mu Li, Jiapeng Liu, Zhiying Leng, Frederick W. B. Li, Ziyao Zhang, and Xiaohui Liang. Fg-t2m++: Llms-augmented fine-grained text driven human motion generation. *Int. J. Comput. Vis.*, 133(7):4277–4293, 2025a. doi: 10.1007/S11263-025-02392-9. URL https://doi.org/10.1007/s11263-025-02392-9.

Yuan Wang, Yali Li, Xiang Li, and Shengjin Wang. HSI-GPT: A general-purpose large scene-motion-language model for human scene interaction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 7147–7157. Computer Vision Foundation / IEEE, 2025b. doi: 10.1109/CVPR52734.2025.00670. URL https://openaccess.thecvf.com/content/CVPR2025/html/Wang_HSI-GPT_A_General-Purpose_Large_Scene-Motion-Language_Model_for_Human_Scene_Interaction_CVPR_2025_paper.html.

Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. HUMANISE: language-conditioned human motion generation in 3d scenes. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/6030db5195150ac86d942186f4abdad8-Abstract-Conference.html.

Zan Wang, Yixin Chen, Baoxiong Jia, Puhao Li, Jinlu Zhang, Jingze Zhang, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. Move as you say, interact as you can: Language-guided human motion generation with scene affordance. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 433–444. IEEE, 2024. doi: 10.1109/CVPR52733.2024.00049. URL https://doi.org/10.1109/CVPR52733.2024.00049.

Dong Wei, Xiaoning Sun, Huaijiang Sun, Shengxiang Hu, Bin Li, Weiqing Li, and Jianfeng Lu. Enhanced fine-grained motion diffusion for text-driven human motion synthesis. In Michael J. Wooldridge, Jennifer G. Dy, and Sriraam Natarajan (eds.), *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pp. 5876–5884. AAAI Press, 2024. doi: 10.1609/AAAI.V38I6.28401. URL https://doi.org/10.1609/aaai.v38i6.28401.

Ye Yuan, Jiaming Song, Umar Iqbal, Arash Vahdat, and Jan Kautz. Physdiff: Physics-guided human motion diffusion model. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pp. 15964–15975. IEEE, 2023. doi: 10.1109/ICCV51070.2023.01467. URL https://doi.org/10.1109/ICCV51070.2023.01467.

Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Shaoli Huang, Yong Zhang, Hongwei Zhao, Hongtao Lu, and Xi Shen. T2M-GPT: generating human motion from textual descriptions with discrete representations. *CoRR*, abs/2301.06052, 2023a. doi: 10.48550/ARXIV.2301.06052. URL https://doi.org/10.48550/arXiv.2301.06052.

Mingyuan Zhang, Xinying Guo, Liang Pan, Zhongang Cai, Fangzhou Hong, Huirong Li, Lei Yang, and Ziwei Liu. Remodiffuse: Retrieval-augmented motion diffusion model. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 364–373, 2023b. doi: 10.1109/ICCV51070.2023.00040.

Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(6):4115–4128, 2024. doi: 10.1109/TPAMI.2024.3355414. URL https://doi.org/10.1109/TPAMI.2024.3355414.