Move as You Say, Interact as You Can: Language-guided Human Motion Generation with Scene Affordance

Zan Wang[1,2], Yixin Chen[2], Baoxiong Jia[2], Puhao Li[2,3], Jinlu Zhang[2,4], Jingze Zhang[2,3], Tengyu Liu[2], Yixin Zhu[5✉], Wei Liang[1,6✉], Siyuan Huang[2✉]

✉ indicates corresponding authors   [1] School of Computer Science & Technology, Beijing Institute of Technology
[2] National Key Laboratory of General Artificial Intelligence, BIGAI   [3] Dept. of Automation, Tsinghua University
[4] CFCS, School of Computer Science, Peking University   [5] Institute for AI, Peking University
[6] Yangtze Delta Region Academy of Beijing Institute of Technology, Jiaxing

https://afford-motion.github.io

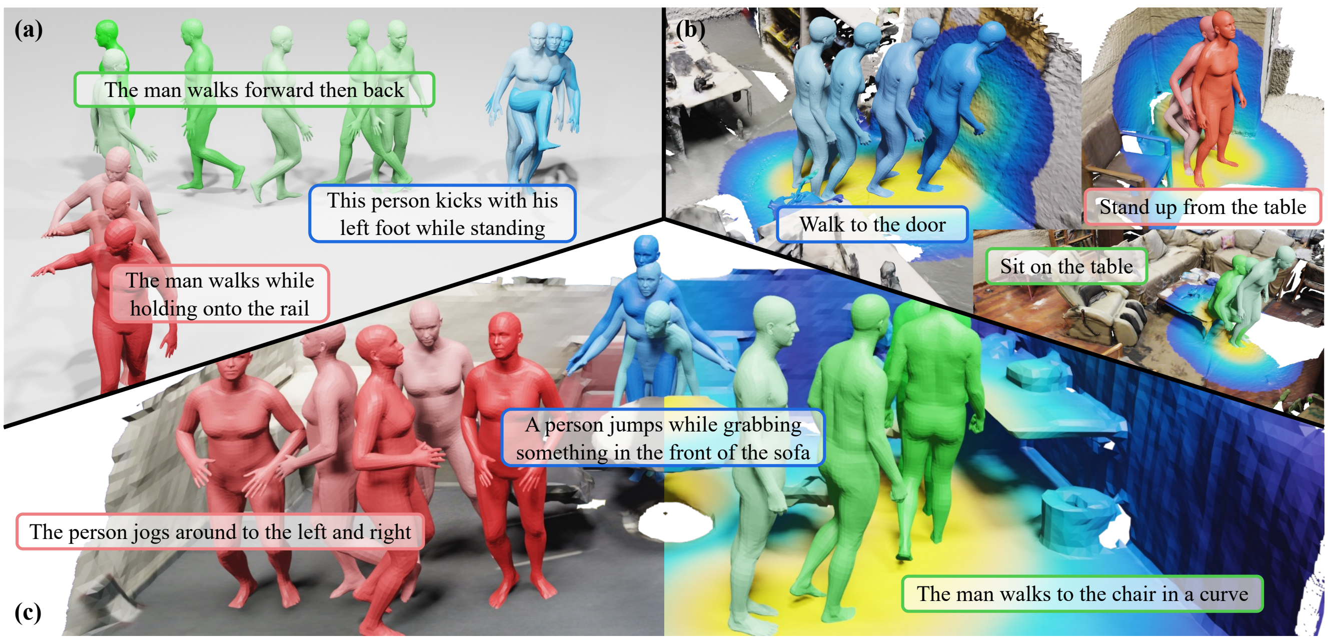


Figure 1. **Language-guided human motion generation in 3D scenes via scene affordance.** Employing scene affordance as an intermediate representation enhances motion generation capabilities on benchmarks (a) HumanML3D and (b) HUMANISE, and significantly boosts the model's ability to generalize to (c) unseen scenarios.

Abstract

Despite significant advancements in text-to-motion synthesis, generating language-guided human motion within 3D environments poses substantial challenges. These challenges stem primarily from (i) the absence of powerful generative models capable of jointly modeling natural language, 3D scenes, and human motion, and (ii) the generative models' intensive data requirements contrasted with the scarcity of comprehensive, high-quality, language-scene-motion datasets. To tackle these issues, we introduce a ***novel two-stage framework that employs scene affordance as an intermediate representation****, effectively linking 3D scene grounding and conditional motion generation. Our framework comprises an Affordance Diffusion Model (ADM) for predicting explicit affordance map and an Affordance-to-Motion Diffusion Model (AMDM) for generating plausible human motions. By leveraging scene affordance maps, our method overcomes the difficulty in generating human motion under multimodal condition signals, especially when training with limited data lacking extensive language-scene-motion pairs. Our extensive experiments demonstrate that our approach consistently outperforms all baselines on established benchmarks, including HumanML3D and HUMANISE. Additionally, we validate our model's exceptional generalization capabilities on a specially curated evaluation set featuring previously unseen descriptions and scenes.*

1. Introduction

Prior efforts in the field have investigated the integration of diverse modalities, such as textual descriptions [3, 5, 6, 14, 24, 29, 67, 75, 76, 92, 93], audio signals [49, 51, 89], and 3D scenes [4, 32, 37, 80–82] for guiding human motion generation. The significant strides in single-modality conditioned motion generation have been complemented by the introduction of Human-Scene Interaction (HSI) through language descriptions by Wang et al. [84], highlighting the demand for controllable motion generation in diverse applications such as animation synthesis [35], film production [78], and synthetic data generation [97, 100]. However, the task of effectively generating semantically driven and scene-aware motions remains daunting due to two principal challenges.

The first challenge entails ensuring that generated motions are descriptive-faithful, physically plausible within the scene, and accurately grounded in specific locations. Though direct application of conditional generative models like conditional Variational Autoencoder (cVAE) [66, 80, 84] and conditional diffusion models [14, 37, 76, 93] has been attempted, the inherent complexity of marrying 3D scene grounding with conditional motion generation presents a significant obstacle. This complexity impedes the model's ability to generalize across various scenes and descriptions, making it challenging to adapt specific motions (*e.g.*, *"lie down on the bed"*) to analogous actions in new contexts (*e.g.*, *"lie down on the floor"*) within unfamiliar 3D environments.

The second challenge arises from the generative models' dependency on large volumes of high-quality paired data. Existing HSI datasets [4, 9, 31] lack in both motion quality and diversity, featuring a limited number of scene layouts and, most critically, devoid of HSI descriptions. Although the HUMANISE dataset [84] attempts to address this gap, it is constrained by a narrow scope of action types and the use of fixed-form utterances, limiting the generation of diverse HSIs from varied and free-form language descriptions.

In response to these challenges, we propose to utilize the **scene affordance maps as an intermediate representation**, as depicted in Fig. 1. This representation is calculated from the distance field between human skeleton joints and the scene's surface points. The use of the affordance map presents two primary benefits for the generation of language-guided motion in 3D environments. First, it precisely delineates the region grounded in the language description, thereby significantly enhancing the 3D scene grounding essential for motion generation, even in scenarios characterized by limited training data availability. Second, the affordance map, rooted in distance measurements, provides a sophisticated understanding of the geometric interplay between scenes and human motions. This understanding aids in the generation of HSI and facilitates the model's ability to generalize across unique scene geometries.

Expanding upon this intermediate representation, we propose a novel two-stage model aimed at seamlessly integrating the 3D scene grounding with the language-guided motion generation. The first stage involves the development of an **Affordance Diffusion Model (ADM)**, which employs the Perceiver architecture [38, 39] to predict an affordance map given a specific 3D scene and description. The second stage introduces an **Affordance-to-Motion Diffusion Model (AMDM)**, comprising an affordance encoder and a Transformer backbone, to synthesize human motions by considering both the language descriptions and the affordance maps derived in the first stage.

We conduct extensive evaluations on established benchmarks, including HumanML3D [29] and HUMANISE [84], demonstrating superior performance in text-to-motion generation tasks and highlighting our model's advanced generalization capabilities on a specially curated evaluation set featuring unseen language descriptions and 3D scenes. These results underscore the utility of our approach in harnessing scene affordances for enriched 3D scene grounding and enhanced conditional motion generation.

Our contributions are summarized as follows:

- We introduce a novel two-stage model that incorporates scene affordance as an intermediate representation, bridging the gap between 3D scene grounding and conditional motion generation, and facilitating language-guided human motion synthesis in 3D environments.
- Through extensive quantitative and qualitative evaluations, we demonstrate our method's superiority over existing motion generation models across the HumanML3D and HUMANISE benchmarks.
- Our model showcases remarkable generalization capabilities, achieving impressive performance in generating human motions for novel language-scene pairs, despite the limited availability of language-scene-motion datasets.

2. Related Work

2.1. Language, Human Motion, and 3D Scene

We seek to bridge the modalities of language, human motion, and 3D scenes, an area where prior research has often focused on combining just two of these elements. In the realm of 3D Vision-Language (3D-VL), tasks such as 3D object grounding [1, 16, 41, 77, 98, 105], reasoning [7, 21, 57, 88], and captioning [12, 13, 18, 36, 91] have intersected language with 3D scenes. Recent advancements in this area have focused on enhancing open-vocabulary scene understanding by integrating features from foundational models like CLIP [68] into 3D scene analysis [40, 45, 63, 74]. The interaction between language and human motion has been explored through efforts to guide motion generation with semantic cues, including text-to-motion [3, 6, 24, 29, 67] and action-to-motion synthesis [28, 66].

Existing HSI works focus on populating static human figures into 3D scenes [15, 17, 33, 97] and generating temporal human motions within these contextual environments [37, 43, 80–82]. A growing body of research [19, 56, 64, 65, 96, 101] has aimed at creating policies for continuous motion synthesis in virtual spaces, treating the challenge as a Reinforcement Learning (RL) task. The pioneering works of Zhao et al. [100] and Wang et al. [84] ventured into the simultaneous modeling of language, 3D scenes, and human motion, integrating semantics (*e.g.*, action labels and descriptive language) into the generation of HSI, requiring interactions to be both physically plausible and semantically consistent. Following this, Xiao et al. [86] leveraged a Large Language Model (LLM) to convert language prompts into sub-task plans, represented as Chain of Contacts (CoC), to facilitate motion planning within 3D scenes.

In our contribution, we present a novel two-stage framework that employs scene affordance [26] as an intermediary to effectively bridge 3D scene grounding with conditioned motion generation. This approach not only enhances multimodal alignment but also improves the generative model's ability to generalize across scenarios, even when trained on the limited paired data available in current datasets [29, 31, 84].

2.2. Conditional Human Motion Generation

The past few years have marked significant advancements in the domain of human motion modeling conditioned on diverse signals [102], including past motion [8, 9, 50, 59, 87, 90], audio [49, 51, 89], action labels [28, 66], natural language descriptions [3, 5, 6, 14, 24, 29, 67, 75, 76, 92, 93], objects [11, 25, 52, 73, 95], and 3D scenes [4, 32, 37, 44, 80–82]. These approaches, predominantly designed for single-modal conditioning, encounter difficulties in scenarios necessitating the simultaneous consideration of both scene and language cues. For example, methods that seek to align the conditional signal's latent space with that of human motions [2, 6, 67, 75] struggle in this intricate context due to the distinct and complementary nature of 3D scenes and language descriptions in motion generation. The former provides spatial boundaries while the latter offers semantic direction, rendering direct alignment approaches less effective.

Moreover, attempts at directly learning the conditional distribution with models such as cVAE [66, 80, 84] and diffusion models [14, 76, 93] often lead to suboptimal outcomes. This is attributed to the complex entanglement of the joint distribution across the three modalities, which complicates the development of an efficient multimodal embedding space, particularly when data is scarce. In response, our research proposes the utilization of scene affordance as an intermediate representation. This strategy aims to simplify the process of generating motion under multiple conditions, thereby enhancing the model's capacity to interpret and generate multimodal HSI more effectively.

2.3. Scene Affordance

The concept of "affordance," initially introduced by Gibson [26], describes the potential actions that the environment offers for interaction. Early investigations into affordances primarily focused on understanding scenes and object affordances through 2D observations [23, 27, 30, 47, 48, 54, 61, 83, 103, 104]. Transitioning to 3D, initial HSI research implicitly incorporated affordances in scene understanding [80–82, 84], with more recent work exploring explicit 3D visual affordances [22, 60, 104]. These advancements often represent affordances as contact maps for grasping [42, 46, 53, 55, 85] and scene-conditioned motion synthesis [33, 80, 81, 94, 97]. In our approach, we redefine the affordance map as a generalized distance field between human skeleton joints and surface points of 3D scenes. This model first refines the affordance map using the provided 3D scene and language description. It then utilizes the refined affordance map's grounding and geometric information to improve the subsequent conditional motion generation.

3. Preliminaries

Diffusion Model Diffusion models [34, 70, 71] are a class of generative models that operate through an iterative denoising process to learn and sample data distributions. They include a forward process and a reverse process.

The forward process starts with the real data $\mathbf{X}_0$ at step 0, iteratively adds Gaussian noise $\boldsymbol{\epsilon}_t$, and converts $\mathbf{X}_0$ to $\mathbf{X}_t$ over t steps in a Markovian manner. The one-step forward process can be described as $q(\mathbf{X}_t \mid \mathbf{X}_{t-1}) = \mathcal{N}(\mathbf{X}_t; \sqrt{1-\beta_t}\mathbf{X}_{t-1}, \beta_t \mathbf{I})$, where $\{\beta_t \in (0,1)\}_{t=1}^{T}$ is the pre-defined variance schedule. When $t \to \infty$, $\mathbf{X}_t$ is equivalent to an isotropic Gaussian distribution. The entire forward process is given by $q(\mathbf{X}_{1:T} \mid \mathbf{X}_0) = \prod_{t=1}^{T} q(\mathbf{X}_t|\mathbf{X}_{t-1})$, where T is the total number of diffusion steps.

In the reverse process, the diffusion model learns to gradually remove noise for sampling from the Gaussian distribution $\mathbf{X}_T$: $p_\theta(\mathbf{X}_{0:T}) = p(\mathbf{X}_T)\prod_{t=1}^{T} p_\theta(\mathbf{X}_{t-1} \mid \mathbf{X}_t)$, $p_\theta(\mathbf{X}_{t-1} \mid \mathbf{X}_t) = \mathcal{N}(\mathbf{X}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{X}_t, t), \boldsymbol{\Sigma}_\theta(\mathbf{X}_t, t))$, where $\mathbf{X}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, $\boldsymbol{\mu}_\theta$ and $\boldsymbol{\Sigma}_\theta$ are estimated by the models with learnable parameters θ.

For learning a conditional distribution $p_\theta(\mathbf{X}_0 \mid \mathcal{C})$, the diffusion model can be adapted to include condition $\mathcal{C}$ in the reverse process: $p_\theta(\mathbf{X}_{t-1} \mid \mathbf{X}_t, \mathcal{C}) = \mathcal{N}(\mathbf{X}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{X}_t, t, \mathcal{C}), \boldsymbol{\Sigma}_\theta(\mathbf{X}_t, t, \mathcal{C}))$.

Problem Definition We tackle the task of *language-guided human motion generation in 3D scenes*. The 3D scene is represented as an RGB point cloud $\mathcal{S} \in \mathbb{R}^{N\times 6}$, while the language description is denoted as $\mathcal{L} = [w_1, w_2, \cdots, w_M]$, comprising M tokenized words. Our objective is to generate motion sequences $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^{F}$ that are both physically plausible and semantically consistent with the given descriptions, where each sequence consists of F frames. Diverging

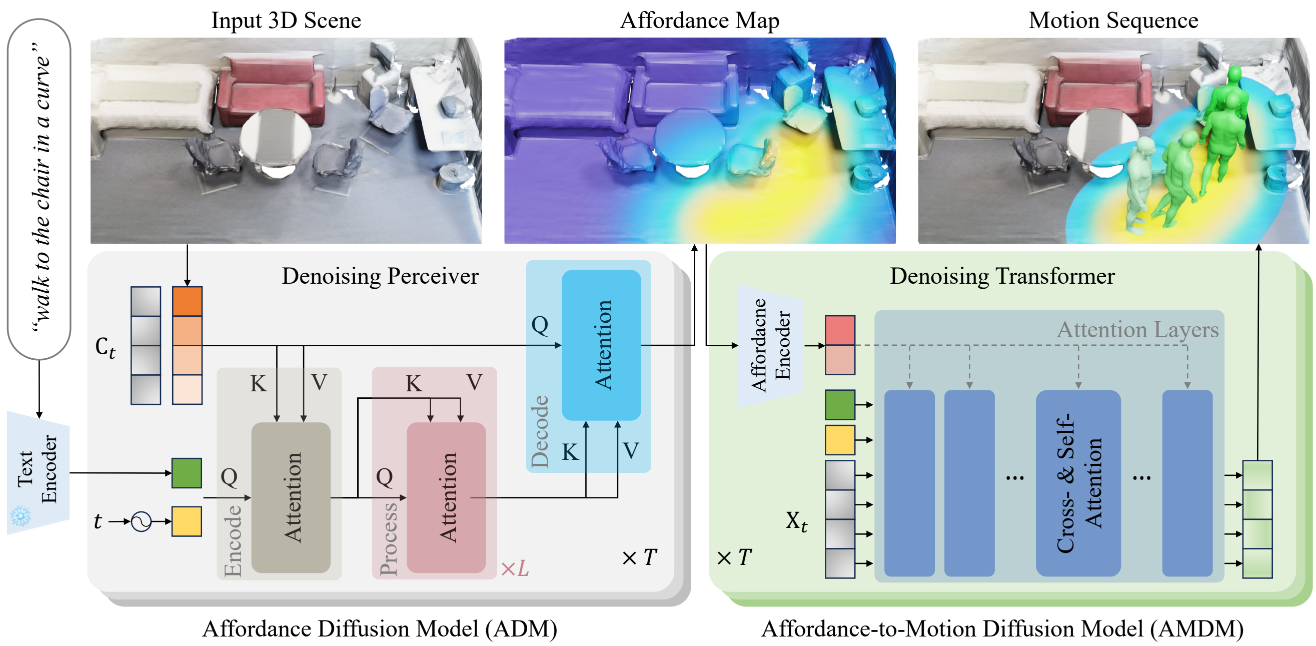


Figure 2. **Overview of our method.** To generate language-guided human motions in 3D scenes, our framework first predicts the scene affordance map in accordance with the language description using Affordance Diffusion Model (ADM). Next, it generates interactive human motions with Affordance-to-Motion Diffusion Model (AMDM) conditioned on the predicted affordance map.

from the redundant motion representation used by Guo et al. [29], we parameterize the per-frame human pose using the body joint positions of the SMPL-X body model [62], specifically $\mathbf{x}_i \in \mathbb{R}^{J\times 3}$, with J representing the total number of joints utilized. For visualization purposes, these poses are converted into body meshes by optimizing the SMPL-X parameters based on joint positions. Please refer to Appendix A for additional details on the optimization process.

4. Method

We propose a novel two-stage model for generating plausible human motions conditioned on the 3D scene and language descriptions. Fig. 2 illustrates the model's framework. The first stage introduces an Affordance Diffusion Model (ADM) to generate language-grounded affordance maps. The second stage takes input as the generated affordance map and the language description to synthesize plausible human motions from Gaussian noise via the proposed Affordance-to-Motion Diffusion Model (AMDM).

4.1. Affordance Map

The affordance map serves as an intermediate representation that abstracts essential details of a 3D indoor scene to support generalization, accurately ground interaction regions, and preserve vital geometric information. In this work, we derive such an affordance map from the distance field between the points in a 3D scene $\mathcal{S}$ and the human skeleton joints across a motion sequence $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^{F}$. We calculate the ℓ_2 distance between each scene point and the skeleton joints per frame, resulting in a per-frame distance field $\mathbf{d} \in \mathbb{R}^{N\times J}$; $\mathbf{d}(n, j)$ measures the distance between the n-th scene point and the j-th skeleton joint. Following Mao et al. [59], we transform this distance field into a normalized distance map $\mathbf{c} \in \mathbb{R}^{N\times J}$:

$$\mathbf{c}(n, j) = \exp\left(-\frac{1}{2}\frac{\mathbf{d}(n, j)}{\sigma^2}\right), \tag{1}$$

where σ is a constant normalizing factor. This operation assigns higher weights to points closer to the joints, thereby aiding in stabilizing the training procedure.

To compute the affordance map $\mathbf{C}$, we employ a max-pooling operation over the temporal dimension of the per-frame distance fields:

$$\mathbf{C} = \texttt{max-pool}(\mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_F). \tag{2}$$

The resulting paired data is denoted as $(\mathbf{C}, \mathbf{X}, \mathcal{S}, \mathcal{L})$, with $\mathcal{S}$ and $\mathcal{L}$ representing the scene's point cloud and the associated language description, respectively.

4.2. Affordance Diffusion Model

To learn the distribution of language-grounded affordance maps, we introduce an Affordance Diffusion Model (ADM) designed to process the 3D scene point cloud $\mathcal{S}$ and the corresponding language description $\mathcal{L}$, generating an affordance map $\mathbf{C}$. This process is formalized as follows:

$$p_\theta(\mathbf{C}_{0:T} \mid \mathcal{S}, \mathcal{L}) = p(\mathbf{C}_T) \prod_{t=1}^{T} p_\theta(\mathbf{C}_{t-1} \mid \mathbf{C}_t, \mathcal{S}, \mathcal{L}). \tag{3}$$

As depicted in Fig. 2, ADM's architecture is based on the Perceiver [38, 39], leveraging an attention mechanism to efficiently extract point-wise features.

The Perceiver backbone within ADM consists of three primary components: an *Encode* block, a *Process* block, and a *Decode* block. Initially, the *Encode* utilizes an attention module to encode the extracted point features along with the noisy affordance map, termed as input features. We denote the concatenation of the language feature and diffusion step embeddings as the latent features. In this configuration, the input features act as the attention module's key and value, with the latent features acting as the query. Next, the *Process* block refines the latent features through multiple self-attention layers. Finally, the *Decode* block employs another attention module, allowing the input features to attend to the updated latent features, thereby achieving refined per-point feature refinement. We forward these per-point feature vectors into a linear layer for further processing. Contrary to approaches that predict the added noise $\boldsymbol{\epsilon}_t$, our model directly estimate the input signal [69, 76], allowing the end-to-end training of ADM, denoted as G_θ, with a simple objective:

$$L_{\text{MSE}} = \mathbb{E}_{\mathbf{C}_0,t}\left[\|\mathbf{C}_0 - G_\theta(\mathbf{C}_t, t, \mathcal{S}, \mathcal{L})\|_2^2\right]. \tag{4}$$

4.3. Affordance-to-Motion Diffusion Model

In the subsequent stage, our framework employs an Affordance-to-Motion Diffusion Model (AMDM) to generate plausible human motions, leveraging both the language descriptions and the previously generated affordance maps:

$$p_\phi(\mathbf{X}_{0:T} \mid \mathbf{C}, \mathcal{S}, \mathcal{L}) = p(\mathbf{X}_T)\prod_{t=1}^{T} p_\phi(\mathbf{X}_{t-1} \mid \mathbf{X}_t, \mathbf{C}, \mathcal{S}, \mathcal{L}). \tag{5}$$

The architecture of the model is illustrated in Fig. 2. The AMDM comprises an encoder specifically for the affordance map and a Transformer backbone that integrates multimodal features to facilitate motion generation. Utilizing a Point Transformer architecture [99], the affordance map encoder extracts feature maps of varying cardinalities, which are further processed by U-net decoder layers; the Transformer backbone stacks self-attention and cross-attention layers. We concatenate the noisy motion sequence with language features and diffusion timestep embeddings and forward this concatenation to the Transformer backbone. In each cross-attention layer, the concatenation attends to the affordance features to fuse multimodal information. A linear layer finally maps the fused features into the motion space.

Similar to ADM, we train the AMDM, denoted as G_ϕ, by optimizing a mean squared error objective:

$$L_{\text{MSE}} = \mathbb{E}_{\mathbf{X}_0,t}\left[\|\mathbf{X}_0 - G_\phi(\mathbf{X}_t, t, \mathbf{C}, \mathcal{S}, \mathcal{L})\|_2^2\right]. \tag{6}$$

4.4. Implementation Details

In our implementation, we use a frozen *CLIP-ViT-B/32* to extract text features in both stages. The normalization factor σ is set to 0.8. The Transformer models are constructed using the native PyTorch implementation. Both ADM and AMDM undergo training to convergence using the AdamW optimizer with a fixed learning rate of 10^{-4}. For the training of ADM, we leverage 2 NVIDIA A100 GPUs, assigning a batch size of 64 per GPU. The training of AMDM is conducted on 4 NVIDIA A100 GPUs, with a batch size of 32 per GPU. Refer to Appendix C for further implementation details.

5. Experiments

To demonstrate the efficacy of our methods, we conducted evaluations using the HumanML3D [29], HUMANISE [84], and a uniquely compiled evaluation set specifically curated for examining the generalization capability.

5.1. Datasets

We evaluate our model on HumanML3D [29], a modern text-to-motion dataset derived from annotating AMASS [58] motion sequences with sequential-level descriptions. As HumanML3D lacks 3D scenes, we augment it by adding a floor to support the training and evaluation of our two-stage model. We use the original motion representation and train-test splits in the task setting.

We also evaluate our model on HUMANISE [84], distinguished as the first extensive and semantic-rich HSI dataset that aligns motion sequences from AMASS with the 3D scene from ScanNet [20]. The synthesized results are automatically annotated with descriptions from Sr3D [1]. We exclude spatially referring descriptions and segment scenes into chunks while retaining the original motions and splits.

To probe the model's generalization prowess, we curate a novel evaluation set that comprises 16 scenes from diverse sources, including ScanNet [20], PROX [31], Replica [72], and Matterport3D [10], along with 80 HSI descriptions crafted by Turkers. Furthermore, we construct a training set that connects language, 3D scene, and motion by incorporating data from HumanML3D, HUMANISE, and PROX. We leverage the annotations to unify the representation as joint positions across different datasets; we augment the HumanML3D by randomly positioning furniture [79] around the motion to boost 3D scene awareness. This consolidated dataset comprises 63,770 HSIs, with 48,470 featuring language annotations. Refer to Appendix E for more details.

5.2. Metrics and Baselines

Metrics For the evaluation on **HumanML3D**, we adopt the metrics proposed by Guo et al. [29], including *Diversity*, measuring the variation within generated motions; *Multi-Modality*, quantifying the average variation relative to text descriptions; *R-Precision* and *Multimodal-Dist*, assessing the relevance between generated motions and language descriptions; and *FID*, evaluating the discrepancy between the distributions of generated results and the original dataset. On

Table 1. **Quantitative results of generation on HumanML3D.** "Real" denotes the results computed with GT motions. "→" indicates metrics that are better when closer to "Real" distribution. Our model uses Perceiver in ADM and encoder-based architecture in AMDM.

Model	R-Precision ↑			FID ↓	MultiModal Dist. ↓	Diversity →	MultiModality ↑
	Top 1	Top 2	Top 3				
Real	$0.511^{\pm.003}$	$0.703^{\pm.003}$	$0.797^{\pm.002}$	$0.002^{\pm.000}$	$2.974^{\pm.008}$	$9.503^{\pm.065}$	-
Language2Pose [3]	$0.246^{\pm.002}$	$0.387^{\pm.002}$	$0.486^{\pm.002}$	$11.02^{\pm.046}$	$5.296^{\pm.008}$	$7.676^{\pm.058}$	-
T2M [29]	$\mathbf{0.457}^{\pm.002}$	$\mathbf{0.639}^{\pm.003}$	$\mathbf{0.740}^{\pm.003}$	$1.067^{\pm.002}$	$\mathbf{3.340}^{\pm.008}$	$9.188^{\pm.002}$	$2.090^{\pm.083}$
MDM [76]	$0.319^{\pm.005}$	$0.498^{\pm.004}$	$0.611^{\pm.007}$	$0.544^{\pm.044}$	$5.566^{\pm.027}$	$\mathbf{9.559}^{\pm.086}$	$2.799^{\pm.072}$
Ours	$0.341^{\pm.010}$	$0.514^{\pm.016}$	$0.625^{\pm.011}$	$\mathbf{0.352}^{\pm.109}$	$5.455^{\pm.073}$	$9.772^{\pm.117}$	$\mathbf{2.835}^{\pm.075}$
MDM† [76]	$0.418^{\pm.005}$	$0.604^{\pm.005}$	$0.707^{\pm.004}$	$0.489^{\pm.025}$	$3.631^{\pm.023}$	$\mathbf{9.449}^{\pm.066}$	$\mathbf{2.873}^{\pm.111}$
Ours†	$\mathbf{0.432}^{\pm.007}$	$\mathbf{0.629}^{\pm.007}$	$\mathbf{0.733}^{\pm.006}$	$\mathbf{0.352}^{\pm.109}$	$\mathbf{3.430}^{\pm.061}$	$9.825^{\pm.159}$	$2.835^{\pm.075}$

Table 2. **Quantitative results of human motion generation on HUMANISE dataset. Bold** indicates the best result.

Model	goal dist.↓	APD↑	contact↑	non-collision↑	quality score↑	action score↑
cVAE [84]	$0.422^{\pm.011}$	$4.094^{\pm.013}$	$84.06^{\pm.716}$	$\mathbf{99.77}^{\pm.004}$	2.25 ± 1.26	3.66 ± 1.38
one-stage @ Enc	$0.326^{\pm.013}$	$\mathbf{5.510}^{\pm.019}$	$76.11^{\pm.684}$	$99.71^{\pm.014}$	2.60 ± 1.24	3.88 ± 1.32
one-stage @ Dec	$0.185^{\pm.014}$	$4.063^{\pm.020}$	$86.43^{\pm.845}$	$99.76^{\pm.006}$	3.09 ± 1.34	4.18 ± 1.16
Ours @ Enc	$\mathbf{0.156}^{\pm.006}$	$2.597^{\pm.008}$	$95.86^{\pm.323}$	$99.69^{\pm.007}$	3.46 ± 1.15	$\mathbf{4.47 \pm 0.84}$
Ours @ Dec	$\mathbf{0.156}^{\pm.006}$	$2.459^{\pm.009}$	$\mathbf{96.04}^{\pm.298}$	$99.70^{\pm.005}$	$\mathbf{3.55 \pm 1.19}$	4.44 ± 0.85

HUMANISE, we follow the evaluation protocol of Wang et al. [84] and Zhang et al. [97], utilizing metrics of *goal dist.* to determine grounding accuracy, *Average Pairwise Distance (APD)* for the diversity of motions, and physics-based metrics like *contact* and *non-collision* scores. Human perceptual studies further evaluate the *quality* and *action* score of the generated motions. To evaluate **ADM**, we further introduce three grounding metrics, *i.e.*, *min dist.*, *pelvis dist.*, and *all dist.*, to quantify the accuracy of the affordance map in guiding interactions, based on distances from the joints to target objects within the scene. We also employ these metrics on the **novel evaluation set**. Due to unique motion representations, we retrain the motion and text feature extractors as follows Guo et al. [29] for consistent metric calculations. All evaluations are conducted five times to ensure robustness, with a 95% confidence interval indicated by $\pm$. For *quality* and *action* scores, mean and standard deviation are reported.

Baselines For evaluations on **HumanML3D**, we include the following baselines: Language2Pose [3], T2M [29], and MDM [76]. For **HUMANISE**, we utilize the cVAE-based approach by Wang et al. [84], hereafter referred to as *cVAE*. To evaluate **scene affordances** and our **two-stage model** architecture, we implement an *one-stage* diffusion model variant that directly processes the scene point cloud, bypassing the affordance map generation; this variant replicates the AMDM's architecture. Moreover, we examine an encoder adaptation of AMDM that integrates the concatenated features from the three modalities (human motion, affordance map, and language description) directly into self-attention layers, serving additional baselines. The designations @ *Enc* and @ *Dec* refer to encoder and decoder variants, respectively. For **affordance map generation** in the first stage, we explore two additional architectural variations of ADM, MLP and Point Transformer, to further understand their impact on performance. Further details of baseline models' architecture are available in Appendix B.

"†" indicates adjustments following bug fixes in the evaluation code, detailed at https://github.com/GuyTevet/motion-diffusion-model/issues/182.

5.3. Results on HumanML3D

Tab. 1 showcases the quantitative results on HumanML3D, where our method notably excels in the *FID* metric, outperforming all baselines. Specifically, against MDM [76], our method demonstrates enhanced performance in *R-Precision*, *FID*, and *MultiModal Dist.*, while preserving a comparable level of diversity, even in the absence of auxiliary geometric losses. Given that MDM stands as a leading diffusion model in motion generation with a Transformer backbone similar to ours, these findings underscore the benefits of integrating scene affordance into text-to-motion synthesis by enriching movement details such as joint trajectories, evidenced even with a simple floor augmentation to the language-motion dataset. Appendix D provides qualitative results of the predicted affordance maps and generated motions.

5.4. Results on HUMANISE

Quantitative Results Quantitative evaluations presented in Tab. 2 affirm our method's capability in producing high-fidelity human motion sequences that are well-grounded conditioned on scenes and language instructions, outperforming both the *cVAE* and one-stage diffusion model baselines. Notably, our model surpasses these baselines

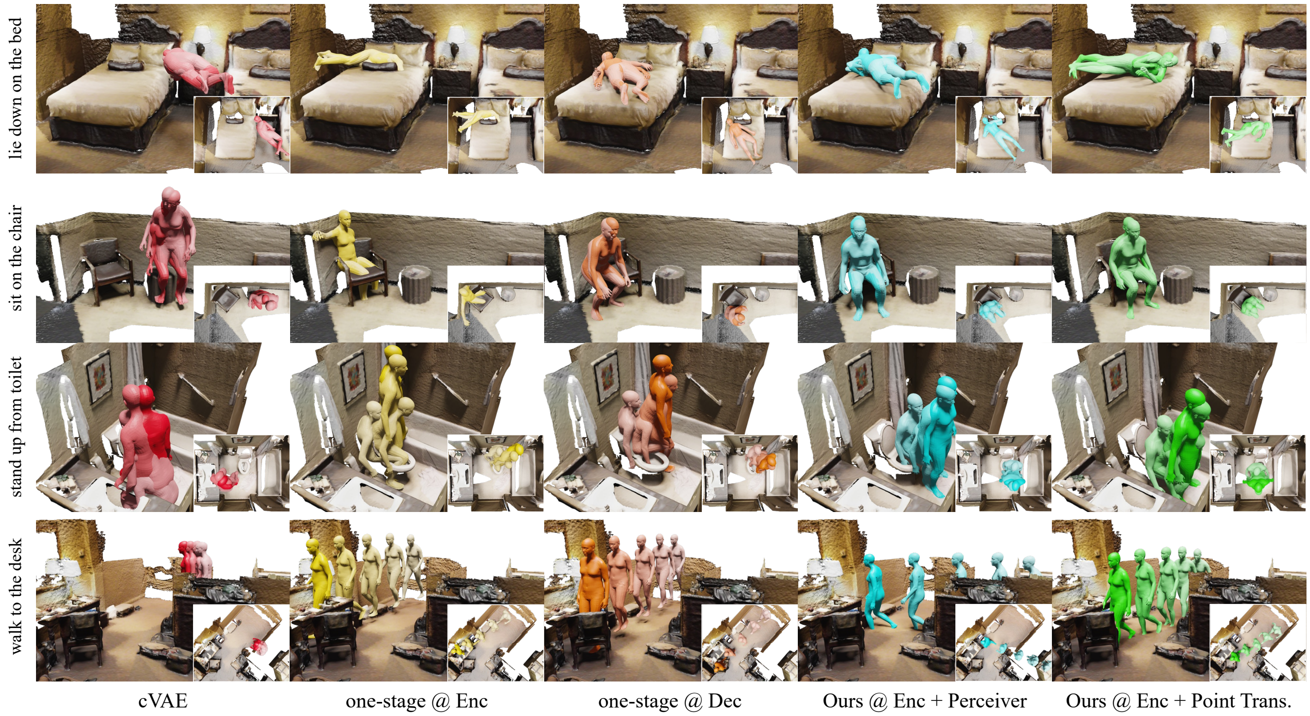


Figure 3. **Qualitative results on HUMANISE dataset.** The bottom-right figure provides a top-down view. Zoom in for better visualization.

across *goal dist.*, *contact*, *quality*, and *action* scores, signaling a pronounced advancement in generating human motion that aligns accurately with scene and language instructions. Note that the diminished diversity in the *APD* metric mainly stems from the enhanced precision in motion generation within our model, which effectively grounds motions in the 3D scene with desired semantics and interactions, as opposed to the motions with potential physical implausibility, incorrect semantics or inadequate grounding observed in the baselines methods.

Qualitative Results Visualizations in Fig. 3 reveal that the cVAE model often fails to accurately ground the target object, indicating limited scene-aware capabilities. Moreover, the one-stage model's lack of scene geometry awareness can result in human-scene collisions and non-contacts.

Affordance Generation Evaluation Grounding distance metrics, detailed in Tab. 3, illustrate that among the three variants, the MLP lags in grounding accuracy when compared to the Perceiver and Point Transformer models. This discrepancy might arise from the MLP's isolated processing of individual scene points, which limits their information exchange. In contrast, the Perceiver consistently excels, presumably due to its effective integration of point and language features through cross-attention mechanisms.

5.5. Results on Novel Evaluation Set

Results The performance on the uniquely curated evaluation set, featuring new scenes and language descriptions from Turkers, is summarized in Tab. 4. Our approach exhibits considerable enhancements in *FID* while maintaining comparable *R-Precision* and *Multimodal-Dist*, suggesting a robust capability to synthesize plausible human motions aligned with the given language instructions. Notably, our method generates a wider variety of human motions, as evidenced by improved scores in metrics *MultiModality* when compared to baseline methods. These results underscore our approach's efficacy in producing physically plausible and semantically consistent human motions conditioned on scenes and language instructions, validated through *contact*, *quality*, and *action* scores. Fig. 4 presents qualitative results generated with unseen language descriptions and 3D scenes.

Table 3. **Quantitative results of affordance map generation.** We report the three distance metrics to evaluate the grounding accuracy.

Arch. of ADM	min dist. ↓	pelvis dist. ↓	all dist. ↓
G.T.	0.736	0.923	1.039
MLP	$0.904^{\pm.003}$	$1.335^{\pm.008}$	$1.513^{\pm.008}$
Point Trans.	$0.878^{\pm.008}$	$1.090^{\pm.008}$	$1.204^{\pm.009}$
Perceiver	$\mathbf{0.756^{\pm.007}}$	$\mathbf{1.005^{\pm.005}}$	$\mathbf{1.086^{\pm.007}}$

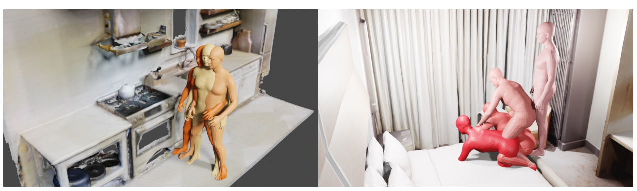
Failure Cases Fig. 5 depicts typical failure cases encountered by our model. For instance, challenges arise with test scenarios of unseen human-scene interactions, resulting in accurately generated motions in the correct space (*e.g.*, hand washing near a tap) but inaccurate interactions (*e.g.*, failure to align the body appropriately facing the sink). The model also fails with language descriptions of high complexity exceeding its current capabilities.

Table 4. **Qualitative results on our novel evaluation set.** “Real” indicates that we compute these metrics as a reference using the language-motion pairs within the test set of HumanML3D. Of note, our novel evaluation set does not contain ground truth motions.

Model	R-Precision (Top 3)↑	FID↓	MultiModal Dist.↓	Diversity→	MultiModality↑	contact↑	non-collision↑	quality score↑	action score↑
Real	$0.875^{\pm.002}$	$0.000^{\pm.000}$	$3.342^{\pm.004}$	$9.442^{\pm.301}$	-	-	-	-	-
one-stage @ Enc	$\mathbf{0.500^{\pm.044}}$	$11.848^{\pm1.634}$	$\mathbf{5.954^{\pm.235}}$	$\mathbf{8.395^{\pm.850}}$	$4.966^{\pm.321}$	$46.64^{\pm4.024}$	$99.88^{\pm.018}$	1.94 ± 1.15	2.61 ± 1.45
one-stage @ Dec	$0.403^{\pm.044}$	$12.268^{\pm.900}$	$6.611^{\pm.227}$	$8.049^{\pm.708}$	$5.031^{\pm.423}$	$26.75^{\pm4.264}$	$\mathbf{99.93^{\pm.023}}$	1.44 ± 0.83	1.96 ± 1.27
Ours @ Enc	$0.478^{\pm.069}$	$\mathbf{7.887^{\pm1.189}}$	$6.226^{\pm.261}$	$7.935^{\pm.857}$	$\mathbf{5.159^{\pm.356}}$	$71.98^{\pm2.542}$	$99.83^{\pm.006}$	$\mathbf{2.06 \pm 1.23}$	$\mathbf{2.63 \pm 1.47}$
Ours @ Dec	$0.428^{\pm.023}$	$12.027^{\pm3.164}$	$6.412^{\pm.204}$	$7.603^{\pm.715}$	$4.966^{\pm.353}$	$\mathbf{88.63^{\pm2.975}}$	$99.82^{\pm.015}$	1.99 ± 1.24	2.49 ± 1.40

“A person wanders in the room around the table.” *“A man dances on the bed happily.”*

Figure 4. **Qualitative comparisons on generalization evaluation set.** The first row is generated by the one-stage diffusion model and the second row is generated by our model. Our method can generate natural and accurately grounded human motions in unseen 3D scenes.



“A person is washing his hands with the tap.” *“A person gets up from one bed and lies down on another bed.”*

Figure 5. **Failure cases.** Our model fails while facing entirely unfamiliar HSIs or too complex descriptions.

Table 5. **Ablation of the architectures of AMDM.** The Perceiver architecture slightly outperforms the Point Transformer in the metrics of *goal dist.* and *contact* score.

Arch. of ADM	goal dist.↓	contact↑	non-collision↑
G.T.	0.017	90.79	99.84
MLP	$0.394^{\pm.010}$	$73.96^{\pm.434}$	$\mathbf{99.84^{\pm.005}}$
Point Trans.	$0.164^{\pm.010}$	$94.39^{\pm.408}$	$99.82^{\pm.008}$
Perceiver	$\mathbf{0.156^{\pm.006}}$	$\mathbf{95.86^{\pm.323}}$	$99.69^{\pm.007}$

5.6. Ablation Study

We further examine the impact of different ADM architectures on motion generation in the second stage of HUMANISE, utilizing an encoder-based AMDM. As outlined in Tab. 5, both Perceiver and Point Transformer yield superior *goal dist.* outcomes compared to the MLP, echoing findings from Tab. 3. Furthermore, these architectures enhance the physical realism, as indicated by improved *contact* scores, with Perceiver models having higher collision rates relative to Point Transformers, echoing observations in Fig. 3.

6. Conclusion

We introduced a novel two-stage model that leverages scene affordance as an intermediate representation to bridge the 3D scene grounding and subsequent conditional motion generation. The quantitative and qualitative results demonstrate promising improvements in HumanML3D and HUMANISE. The model's adaptability was further validated on a uniquely curated evaluation set featuring unseen scenes and language prompts, showcasing its robustness in novel scenarios.

Limitations (i) The reliance on diffusion models contributes to slower inference times, marking a significant drawback for future work. (ii) Although employing affordance maps mitigates the challenges posed by the scarcity of paired data for training in 3D environments, data limitation remains a critical hurdle. Future initiatives should focus on devising strategies to overcome this persistent challenge.

Acknowledgments The authors would like to thank NVIDIA for their generous support of GPUs and hardware. This work is supported in part by the National Science and Technology Major Project (2022ZD0114900), the National Natural Science Foundation of China (NSFC) (62172043), and the Beijing Nova Program.

References

[1] Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas Guibas. Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes. In *European Conference on Computer Vision (ECCV)*, 2020. 2, 5

[2] Hyemin Ahn, Timothy Ha, Yunho Choi, Hwiyeon Yoo, and Songhwai Oh. Text2action: Generative adversarial synthesis from language to action. In *International Conference on Robotics and Automation (ICRA)*, 2018. 3

[3] Chaitanya Ahuja and Louis-Philippe Morency. Language2pose: Natural language grounded pose forecasting. In *International Conference on 3D Vision (3DV)*, 2019. 2, 3, 6

[4] Joao Pedro Araújo, Jiaman Li, Karthik Vetrivel, Rishi Agarwal, Jiajun Wu, Deepak Gopinath, Alexander William Clegg, and Karen Liu. Circle: Capture in rich contextual environments. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 3

[5] Nikos Athanasiou, Mathis Petrovich, Michael J Black, and Gül Varol. Teach: Temporal action composition for 3d humans. In *International Conference on 3D Vision (3DV)*, 2022. 2, 3

[6] Nikos Athanasiou, Mathis Petrovich, Michael J. Black, and Gül Varol. SINC: Spatial composition of 3D human motions for simultaneous action generation. In *International Conference on Computer Vision (ICCV)*, 2023. 2, 3

[7] Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. Scanqa: 3d question answering for spatial scene understanding. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[8] Emad Barsoum, John Kender, and Zicheng Liu. Hp-gan: Probabilistic 3d human motion prediction via gan. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

[9] Zhe Cao, Hang Gao, Karttikeya Mangalam, Qi-Zhi Cai, Minh Vo, and Jitendra Malik. Long-term human motion prediction with scene context. In *European Conference on Computer Vision (ECCV)*, 2020. 2, 3

[10] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. In *International Conference on 3D Vision (3DV)*, 2017. 5

[11] Yu-Wei Chao, Jimei Yang, Weifeng Chen, and Jia Deng. Learning to sit: Synthesizing human-chair interactions via hierarchical control. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2021. 3

[12] Dave Zhenyu Chen, Qirui Wu, Matthias Nießner, and Angel X Chang. D3net: a speaker-listener architecture for semi-supervised dense captioning and visual grounding in rgb-d scans. In *European Conference on Computer Vision (ECCV)*, 2022. 2

[13] Sijin Chen, Hongyuan Zhu, Xin Chen, Yinjie Lei, Gang Yu, and Tao Chen. End-to-end 3d dense captioning with vote2cap-detr. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[14] 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 *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 3

[15] Yixin Chen, Siyuan Huang, Tao Yuan, Siyuan Qi, Yixin Zhu, and Song-Chun Zhu. Holistic++ scene understanding: Single-view 3d holistic scene parsing and human pose estimation with human-object interaction and physical commonsense. In *International Conference on Computer Vision (ICCV)*, 2019. 3

[16] Yixin Chen, Qing Li, Deqian Kong, Yik Lun Kei, Tao Gao, Yixin Zhu, and Siyuan Huang. Yourefit: Embodied reference understanding with language and gesture. In *International Conference on Computer Vision (ICCV)*, 2021. 2

[17] Yixin Chen, Sai Kumar Dwivedi, Michael J Black, and Dimitrios Tzionas. Detecting human-object contact in images. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[18] Zhenyu Chen, Ali Gholami, Matthias Nießner, and Angel X Chang. Scan2cap: Context-aware dense captioning in rgb-d scans. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[19] Jieming Cui, Tengyu Liu, Nian Liu, Yaodong Yang, Yixin Zhu, and Siyuan Huang. Anyskill: Learning open-vocabulary physical skill for interactive agents. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

[20] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 5

[21] Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. Embodied question answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[22] Shengheng Deng, Xun Xu, Chaozheng Wu, Ke Chen, and Kui Jia. 3d affordancenet: A benchmark for visual object affordance understanding. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3

[23] Kuan Fang, Te-Lin Wu, Daniel Yang, Silvio Savarese, and Joseph J Lim. Demo2vec: Reasoning object affordances from online videos. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

[24] Anindita Ghosh, Noshaba Cheema, Cennet Oguz, Christian Theobalt, and Philipp Slusallek. Synthesis of compositional animations from textual descriptions. In *International Conference on Computer Vision (ICCV)*, 2021. 2, 3

[25] Anindita Ghosh, Rishabh Dabral, Vladislav Golyanik, Christian Theobalt, and Philipp Slusallek. Imos: Intent-driven full-body motion synthesis for human-object interactions. In *Computer Graphics Forum*, 2023. 3

[26] James J Gibson. The theory of affordances. *Hilldale, USA*, 1(2):67–82, 1977. 3

[27] Helmut Grabner, Juergen Gall, and Luc Van Gool. What makes a chair a chair? In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011. 3

[28] 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 *International Conference on Multimedia*, 2020. 2, 3

[29] 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 *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 3, 4, 5, 6, A3

[30] Abhinav Gupta, Scott Satkin, Alexei A. Efros, and Martial Hebert. From 3d scene geometry to human workspace. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011. 3

[31] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J Black. Resolving 3d human pose ambiguities with 3d scene constraints. In *International Conference on Computer Vision (ICCV)*, 2019. 2, 3, 5

[32] Mohamed Hassan, Duygu Ceylan, Ruben Villegas, Jun Saito, Jimei Yang, Yi Zhou, and Michael J Black. Stochastic scene-aware motion prediction. In *International Conference on Computer Vision (ICCV)*, 2021. 2, 3

[33] Mohamed Hassan, Partha Ghosh, Joachim Tesch, Dimitrios Tzionas, and Michael J Black. Populating 3d scenes by learning human-scene interaction. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3

[34] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 3

[35] Daniel Holden, Jun Saito, and Taku Komura. A deep learning framework for character motion synthesis and editing. *ACM Transactions on Graphics (TOG)*, 35(4):1–11, 2016. 2

[36] Jiangyong Huang, Silong Yong, Xiaojian Ma, Xiongkun Linghu, Puhao Li, Yan Wang, Qing Li, Song-Chun Zhu, Baoxiong Jia, and Siyuan Huang. An embodied generalist agent in 3d world. *arXiv preprint arXiv:2311.12871*, 2023. 2

[37] Siyuan Huang, Zan Wang, Puhao Li, Baoxiong Jia, Tengyu Liu, Yixin Zhu, Wei Liang, and Song-Chun Zhu. Diffusion-based generation, optimization, and planning in 3d scenes. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 3

[38] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International Conference on Machine Learning (ICML)*, 2021. 2, 5

[39] Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, et al. Perceiver io: A general architecture for structured inputs & outputs. In *International Conference on Learning Representations (ICLR)*, 2022. 2, 5

[40] Krishna Murthy Jatavallabhula, Alihusein Kuwajerwala, Qiao Gu, Mohd Omama, Tao Chen, Shuang Li, Ganesh Iyer, Soroush Saryazdi, Nikhil Keetha, Ayush Tewari, et al. Conceptfusion: Open-set multimodal 3d mapping. *arXiv preprint arXiv:2302.07241*, 2023. 2

[41] Baoxiong Jia, Yixin Chen, Huangyue Yu, Yan Wang, Xuesong Niu, Tengyu Liu, Qing Li, and Siyuan Huang. Sceneverse: Scaling 3d vision-language learning for grounded scene understanding. *arXiv preprint arXiv:2401.09340*, 2024. 2

[42] Hanwen Jiang, Shaowei Liu, Jiashun Wang, and Xiaolong Wang. Hand-object contact consistency reasoning for human grasps generation. In *International Conference on Computer Vision (ICCV)*, 2021. 3

[43] Nan Jiang, Tengyu Liu, Zhexuan Cao, Jieming Cui, Zhiyuan Zhang, Yixin Chen, He Wang, Yixin Zhu, and Siyuan Huang. Full-body articulated human-object interaction. In *International Conference on Computer Vision (ICCV)*, 2023. 3

[44] Nan Jiang, Zhiyuan Zhang, Hongjie Li, Xiaoxuan Ma, Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, and Siyuan Huang. Scaling up dynamic human-scene interaction modeling. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

[45] Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. Lerf: Language embedded radiance fields. In *International Conference on Computer Vision (ICCV)*, 2023. 2

[46] Mia Kokic, Danica Kragic, and Jeannette Bohg. Learning task-oriented grasping from human activity datasets. *IEEE Robotics and Automation Letters (RA-L)*, 5(2):3352–3359, 2020. 3

[47] Hema S Koppula and Ashutosh Saxena. Physically grounded spatio-temporal object affordances. In *European Conference on Computer Vision (ECCV)*, 2014. 3

[48] Sumith Kulal, Tim Brooks, Alex Aiken, Jiajun Wu, Jimei Yang, Jingwan Lu, Alexei A Efros, and Krishna Kumar Singh. Putting people in their place: Affordance-aware human insertion into scenes. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[49] Hsin-Ying Lee, Xiaodong Yang, Ming-Yu Liu, Ting-Chun Wang, Yu-Ding Lu, Ming-Hsuan Yang, and Jan Kautz. Dancing to music. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 2, 3

[50] Chen Li, Zhen Zhang, Wee Sun Lee, and Gim Hee Lee. Convolutional sequence to sequence model for human dynamics. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

[51] Jing Li, Di Kang, Wenjie Pei, Xuefei Zhe, Ying Zhang, Zhenyu He, and Linchao Bao. Audio2gestures: Generating diverse gestures from speech audio with conditional variational autoencoders. In *International Conference on Computer Vision (ICCV)*, 2021. 2, 3

[52] Jiaman Li, Jiajun Wu, and C Karen Liu. Object motion guided human motion synthesis. *ACM Transactions on Graphics (TOG)*, 42(6):1–11, 2023. 3

[53] Puhao Li, Tengyu Liu, Yuyang Li, Yiran Geng, Yixin Zhu, Yaodong Yang, and Siyuan Huang. Gendexgrasp: Generalizable dexterous grasping. In *International Conference on Robotics and Automation (ICRA)*, 2023. 3

[54] Xueting Li, Sifei Liu, Kihwan Kim, Xiaolong Wang, Ming-Hsuan Yang, and Jan Kautz. Putting humans in a scene: Learning affordance in 3d indoor environments. In *Confer-*

ence on Computer Vision and Pattern Recognition (CVPR), 2019. 3

[55] Yuyang Li, Bo Liu, Puhao Li, Yaodong Yang, Yixin Zhu, Tengyu Liu, and Siyuan Huang. Grasp multiple objects with one hand. *IEEE Robotics and Automation Letters (RA-L)*, 9 (5):4027–4034, 2024. 3

[56] Hung Yu Ling, Fabio Zinno, George Cheng, and Michiel Van De Panne. Character controllers using motion vaes. *ACM Transactions on Graphics (TOG)*, 39(4):40–1, 2020. 3

[57] Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. Sqa3d: Situated question answering in 3d scenes. In *International Conference on Learning Representations (ICLR)*, 2023. 2

[58] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *International Conference on Computer Vision (ICCV)*, 2019. 5

[59] Wei Mao, Richard I Hartley, Mathieu Salzmann, et al. Contact-aware human motion forecasting. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 3, 4

[60] Tushar Nagarajan and Kristen Grauman. Learning affordance landscapes for interaction exploration in 3d environments. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 3

[61] Tushar Nagarajan, Christoph Feichtenhofer, and Kristen Grauman. Grounded human-object interaction hotspots from video. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[62] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 4, A1

[63] Songyou Peng, Kyle Genova, Chiyu Jiang, Andrea Tagliasacchi, Marc Pollefeys, Thomas Funkhouser, et al. Openscene: 3d scene understanding with open vocabularies. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[64] Xue Bin Peng, Ze Ma, Pieter Abbeel, Sergey Levine, and Angjoo Kanazawa. Amp: Adversarial motion priors for stylized physics-based character control. *ACM Transactions on Graphics (TOG)*, 40(4):1–20, 2021. 3

[65] Xue Bin Peng, Yunrong Guo, Lina Halper, Sergey Levine, and Sanja Fidler. Ase: Large-scale reusable adversarial skill embeddings for physically simulated characters. *ACM Transactions on Graphics (TOG)*, 41(4):1–17, 2022. 3

[66] Mathis Petrovich, Michael J Black, and Gül Varol. Action-conditioned 3d human motion synthesis with transformer vae. In *International Conference on Computer Vision (ICCV)*, 2021. 2, 3

[67] Mathis Petrovich, Michael J Black, and Gül Varol. Temos: Generating diverse human motions from textual descriptions. In *European Conference on Computer Vision (ECCV)*, 2022. 2, 3

[68] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021. 2

[69] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 5

[70] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning (ICML)*, 2015. 3

[71] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 3

[72] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, et al. The replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019. 5

[73] Omid Taheri, Vasileios Choutas, Michael J Black, and Dimitrios Tzionas. Goal: Generating 4d whole-body motion for hand-object grasping. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[74] Ayça Takmaz, Elisabetta Fedele, Robert W Sumner, Marc Pollefeys, Federico Tombari, and Francis Engelmann. OpenMask3D: Open-Vocabulary 3D Instance Segmentation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 2

[75] Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. Motionclip: Exposing human motion generation to clip space. In *European Conference on Computer Vision (ECCV)*, 2022. 2, 3

[76] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *International Conference on Learning Representations (ICLR)*, 2023. 2, 3, 5, 6, A1

[77] Jesse Thomason, Mohit Shridhar, Yonatan Bisk, Chris Paxton, and Luke Zettlemoyer. Language grounding with 3d objects. In *Conference on Robot Learning (CoRL)*, 2022. 2

[78] Matthew Thorne, David Burke, and Michiel Van De Panne. Motion doodles: an interface for sketching character motion. *ACM Transactions on Graphics (TOG)*, 23(3):424–431, 2004. 2

[79] Mikaela Angelina Uy, Quang-Hieu Pham, Binh-Son Hua, Thanh Nguyen, and Sai-Kit Yeung. Revisiting point cloud classification: A new benchmark dataset and classification model on real-world data. In *International Conference on Computer Vision (ICCV)*, 2019. 5, A2

[80] Jiashun Wang, Huazhe Xu, Jingwei Xu, Sifei Liu, and Xiaolong Wang. Synthesizing long-term 3d human motion and interaction in 3d scenes. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2, 3

[81] Jingbo Wang, Sijie Yan, Bo Dai, and Dahua Lin. Scene-aware generative network for human motion synthesis. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3

[82] Jingbo Wang, Yu Rong, Jingyuan Liu, Sijie Yan, Dahua Lin, and Bo Dai. Towards diverse and natural scene-aware 3d human motion synthesis. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 3

[83] Xiaolong Wang, Rohit Girdhar, and Abhinav Gupta. Binge watching: Scaling affordance learning from sitcoms. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 3

[84] Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. Humanise: Language-conditioned human motion generation in 3d scenes. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 2, 3, 5, 6, A1

[85] Yueh-Hua Wu, Jiashun Wang, and Xiaolong Wang. Learning generalizable dexterous manipulation from human grasp affordance. In *Conference on Robot Learning (CoRL)*, 2023. 3

[86] Zeqi Xiao, Tai Wang, Jingbo Wang, Jinkun Cao, Wenwei Zhang, Bo Dai, Dahua Lin, and Jiangmiao Pang. Unified human-scene interaction via prompted chain-of-contacts. In *International Conference on Learning Representations (ICLR)*, 2024. 3

[87] Kevin Xie, Tingwu Wang, Umar Iqbal, Yunrong Guo, Sanja Fidler, and Florian Shkurti. Physics-based human motion estimation and synthesis from videos. In *International Conference on Computer Vision (ICCV)*, 2021. 3

[88] Shuquan Ye, Dongdong Chen, Songfang Han, and Jing Liao. 3d question answering. *IEEE Transactions on Visualization and Computer Graph (TVCG)*, pages 1–16, 2022. 2

[89] Hongwei Yi, Hualin Liang, Yifei Liu, Qiong Cao, Yandong Wen, Timo Bolkart, Dacheng Tao, and Michael J Black. Generating holistic 3d human motion from speech. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 3

[90] Ye Yuan and Kris Kitani. Dlow: Diversifying latent flows for diverse human motion prediction. In *European Conference on Computer Vision (ECCV)*, 2020. 3

[91] Zhihao Yuan, Xu Yan, Yinghong Liao, Yao Guo, Guanbin Li, Shuguang Cui, and Zhen Li. X-trans2cap: Cross-modal knowledge transfer using transformer for 3d dense captioning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[92] 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. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 3

[93] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *arXiv preprint arXiv:2208.15001*, 2022. 2, 3

[94] Siwei Zhang, Yan Zhang, Qianli Ma, Michael J Black, and Siyu Tang. Place: Proximity learning of articulation and contact in 3d environments. In *International Conference on 3D Vision (3DV)*, 2020. 3

[95] Xiaohan Zhang, Bharat Lal Bhatnagar, Sebastian Starke, Vladimir Guzov, and Gerard Pons-Moll. Couch: Towards controllable human-chair interactions. In *European Conference on Computer Vision (ECCV)*, 2022. 3

[96] Yan Zhang and Siyu Tang. The wanderings of odysseus in 3d scenes. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[97] Yan Zhang, Mohamed Hassan, Heiko Neumann, Michael J Black, and Siyu Tang. Generating 3d people in scenes without people. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2, 3, 6

[98] Yiming Zhang, ZeMing Gong, and Angel X Chang. Multi3drefer: Grounding text description to multiple 3d objects. In *International Conference on Computer Vision (ICCV)*, 2023. 2

[99] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip HS Torr, and Vladlen Koltun. Point transformer. In *International Conference on Computer Vision (ICCV)*, 2021. 5, A1

[100] Kaifeng Zhao, Shaofei Wang, Yan Zhang, Thabo Beeler, and Siyu Tang. Compositional human-scene interaction synthesis with semantic control. In *European Conference on Computer Vision (ECCV)*, 2022. 2, 3

[101] Kaifeng Zhao, Yan Zhang, Shaofei Wang, Thabo Beeler, and Siyu Tang. Synthesizing diverse human motions in 3d indoor scenes. In *International Conference on Computer Vision (ICCV)*, 2023. 3

[102] Wentao Zhu, Xiaoxuan Ma, Dongwoo Ro, Hai Ci, Jinlu Zhang, Jiaxin Shi, Feng Gao, Qi Tian, and Yizhou Wang. Human motion generation: A survey. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, pages 1–20, 2023. 3

[103] Yuke Zhu, Alireza Fathi, and Li Fei-Fei. Reasoning about object affordances in a knowledge base representation. In *European Conference on Computer Vision (ECCV)*, 2014. 3

[104] Yixin Zhu, Chenfanfu Jiang, Yibiao Zhao, Demetri Terzopoulos, and Song-Chun Zhu. Inferring forces and learning human utilities from videos. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3

[105] Ziyu Zhu, Xiaojian Ma, Yixin Chen, Zhidong Deng, Siyuan Huang, and Qing Li. 3d-vista: Pre-trained transformer for 3d vision and text alignment. In *International Conference on Computer Vision (ICCV)*, 2023. 2