



HSI-GPT: A General-Purpose Large Scene-Motion-Language Model for Human Scene Interaction

Yuan Wang^{1,2} Yali Li^{1,2} Xiang Li^{1,2} Shengjin Wang^{1,2†}

¹Department of Electronic Engineering, Tsinghua University, China

²Beijing National Research Center for Information Science and Technology (BNRist), China
wy23@mails.tsinghua.edu.cn {liyali13, wgsgj}@tsinghua.edu.cn [†]Corresponding Author

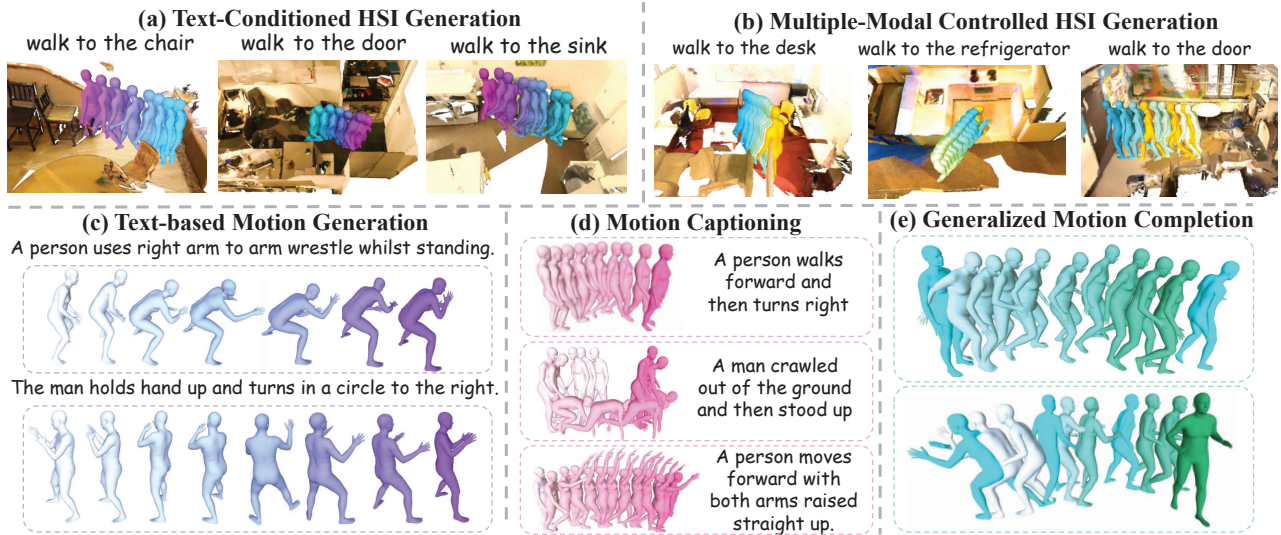


Figure 1. Illustration of our HSI-GPT’s supported tasks. Given different instruction prompts, the proposed HSI-GPT not only accommodates multiple control conditions but handles various HSI-related tasks as well as motion-centric understanding and generation tasks uniformly.

Abstract

While flourishing developments have been witnessed in text-to-motion generation, synthesizing physically realistic, controllable, language-conditioned Human Scene Interactions (HSI) remains a relatively underexplored landscape. Current HSI methods naively rely on conditional Variational AutoEncoder (cVAE) and diffusion models. They are typically associated with **limited modalities of control signals** and **task-specific frameworks design**, leading to inflexible adaptation across various interaction scenarios and descriptive-unfaithful motions in diverse 3D physical environments. In this paper, we propose HSI-GPT, a General-Purpose **Large Scene-Motion-Language Model** that applies “next-token prediction” paradigm of Large Language Models to the HSI domain. HSI-GPT not only exhibits remarkable flexibility to accommodate diverse control signals (3D scenes, textual commands, key-frame poses, as well as

scene affordances), but it seamlessly supports various HSI-related tasks (e.g., multi-modal controlled HSI generation, HSI understanding, and general motion completion in 3D scenes). First, HSI-GPT quantizes textual descriptions and human motions into discrete, LLM-interpretable tokens with multi-modal tokenizers. Inspired by multi-modal learning, we develop a recipe for aligning mixed-modality tokens into the shared embedding space of LLMs. These interaction tokens are then organized into unified instruction following prompts, allowing HSI-GPT to fine-tune on question-and-answer tasks. Extensive experiments and visualizations validate that our general-purpose HSI-GPT model delivers exceptional performance across multiple HSI-related tasks.

1. Introduction

The evolution of Generative Artificial Intelligence [38–40, 59] has prompted extensive research into single-modality conditioned human motion synthesis, where various control

signals—ranging from text descriptions [56, 58, 64, 70] and audio [23, 75, 76] to 3D scene [49, 52, 58]—have been investigated. However, a practical Human Scene Interaction (HSI) conditioned by textual commands remains a longstanding challenge. Indeed, synthesizing motions in a contextual, 3D physical world is a focal point in advancing Embodied AI tasks, *e.g.*, humanoid robots controlling, and scene-level navigation. Intertwined dynamics among motions, textual prompts, and physical constraints of 3D scenes complicate text-conditioned HSI tasks. Thus, achieving a controllable and generalizable HSI holds immense significance.

Existing methods rely on straightforward conditional generative models, *e.g.*, diffusion models [4, 53, 58] and conditional Variational AutoEncoder (cVAE) [34, 48, 52]. While significant strides, they are hindered by several primary limitations: (1) *Non-uniform control conditions with less flexibility*. [52, 53] are constrained to a single control condition, *i.e.*, textual descriptions. While [58] extends multi-frame pose control, it still fails to integrate these multi-modal signals in a unified manner—limiting the flexibility and interactivity of HSI models. (2) *Limited scalability in task-specific frameworks*. Current methods design task-specific models, lacking cross-task generalization capabilities across multiple HSI-related tasks (*e.g.*, captioning, in-betweening, prediction). Moreover, they fail to fully capitalize on the general world knowledge of LLMs, leading to physical inconsistencies and the inability to replicate analogous actions in unseen 3D scenes. (3) *Inadequate exploration of the complex interplay of scene, motion, and language*. Previous studies [4, 52, 53] regard 3D scenes, human motion, and language as isolated modalities, making it challenging to achieve arbitrary conversions between them. Without explicit alignment and relational mapping, the generated human motions are neither descriptive-faithful nor physically plausible in a 3D scene.

Recent studies into the success of Multimodal Large Language Models (MLLMs) [24, 26, 74] have underscored their flexibility and scalability. Our key insight from these studies is to empower the LLM to comprehend and generate human-like motions in 3D scenes. In this work, we propose HSI-GPT, the first **General-Purpose Large Scene-Motion-Language Model** that leverages the powerful reasoning capabilities of LLMs. As illustrated in Fig. 1, our HSI-GPT stands out by unifying diverse control modalities of HSI into the text prompts, and supporting multiple HSI-related tasks, *e.g.*, multi-modal controlled HSI generation, HSI completion, HSI captioning, and motion-centric tasks.

We develop a three-stage training pipeline to enable our HSI-GPT to synthesize physically realistic motions interacting with 3D scenes. (1) *Multi-modal Tokenization*. Our HSI-GPT learns a motion-aware Vector Quantized Variational AutoEncoder (VQ-VAE) [5, 51, 63] to map continuous motions to discrete code indices. The 3D scene and affordance maps are then encoded into visual tokens, using pre-trained

point cloud and affordance encoders. Further, we integrate *diverse conditional tokens into unified instruction templates* to guarantee the scalability and flexible control of our HSI-GPT. (2) *Multi-modal Pre-training Alignment*. Similar to natural language, LLM-interpretable motion tokens convey the interaction intentions between humans and 3D scenes, which can be interpreted as non-verbal “body language”. Motivated by this, we combine the text and motion tokens into a joint motion-language vocabulary and indiscriminately treat tokens from different modalities. For 3D vision tokens, we adopt a Multi-modal Interaction Aggregator (MIA) that aggregates information from affordance map prompts and 3D scenes, resulting in a fixed-length set of learnable querying tokens. These interaction-aware query tokens are mapped into the LLM’s feature space with MLP layers. We optimize the LLM on a mixture of pre-defined tasks with the next-token prediction objective, involving interleaved scene-motion-text triplets. With aligned multi-modal space, HSI-GPT enables us to *explore the comprehensive relationships of scene, motion, and language*. (3) *Instruction Tuning*. HSI-GPT is fine-tuned combined with an elementary markup description of tasks by answering HSI-related questions with efficient LoRA [16]. We present the first effort to extend instruction following to joint scene-motion-language space, *generalizing to various HSI-related tasks*.

Extensive experiments verify that our HSI-GPT delivers competitive performance across multiple HSI-related tasks and supports additional motion-centric tasks. With general world knowledge of LLMs, HSI-GPT showcases powerful zero-shot transfer abilities in unseen 3D environments. In summary, our contributions to the community include:

- We first introduce a general-purpose **Large Scene-Motion-Language Model**, HSI-GPT, capable of unifying diverse control signals and supporting multiple HSI-related tasks.
- We propose a multi-modal pre-training & instruction tuning pipeline to align 3D scene, human motion, and language into LLM’s feature space and enable knowledge transfer across various HSI-related tasks through prompts.
- We establish a general text-conditioned HSI benchmark for multi-task evaluation, highlighting that our HSI-GPT achieves notable performance across diverse tasks.

2. Related Works

2.1. Text-based Human Motion Generation

In the past years, synthesizing human motions conditioned on textual descriptions [1, 5, 9, 11, 22, 35, 43, 63, 65] has been driving and dominating research frontiers. Previous researches [1, 2, 15, 35, 42] learn a joint embedding space of motion and text with the VAE [21] or CLIP [36] models. Recently, applying VQVAE and diffusion models to text-based motion generation tasks has emerged as a burgeoning interest. For instance, T2M-GPT [63], MoMask [11], and UDE [73] quantize the human motions into discrete tokens

with VQVAE and leverage transformers [46] to translate discrete motion indices to text. [5, 43, 64, 65] apply off-the-shelf diffusion models (e.g., DDPM [17] and LDM [39]) for plausible and controllable text-driven motion generation. During the LLM’s carnival [37, 41, 45], MotionGPT [70], MotionLLM [54], and M³-GPT [27] develop unified motion-language model to address multiple motion-related tasks.

2.2. Scene-aware Human Motion Generation

Human motions are goal-oriented and informed by scene layout, as people physically position bodies to interact with the 3D environment. Existing solutions for HSI synthesis follow the common practice to first predict the target position [3, 49] or goal interaction anchor [13, 50, 68], then plan a trajectory and complete the motion along the trajectory [3, 13, 49, 50, 68]. Recent methods [4, 17, 20, 52, 53, 58, 71] have developed *text-conditioned scene interactions* to enhance generation controllability of HSI task. They typically use the task-specific cVAE or diffusion models [14] to learn priors based on human interaction datasets within 3D scenes. HUMANISE [52] stands out as a pioneer by designing a cVAE with separate scenes and text encoders for multi-modality understanding. Afford-Motion [53] employs scene affordance as an intermediate feature to linking 3D scene grounding and diffusion-based motion generation. Cen et al. [4] incorporate ChatGPT [30] and language scene graph to ground referents in 3D scenes, followed by diffusion-based trajectory generation and motion infilling. TeSMo [58] fine-tunes an augmented model on top of a pre-trained text-to-motion diffusion model. In contrast to these methods, our HSI-GPT follows the *autoregressive paradigm of MLLMs* [26, 55, 74] to integrate varied control modalities and generalize well to multiple HSI-related tasks with instruction following.

3. Methods

As demonstrated in Fig. 2, we propose HSI-GPT, a **multi-condition unified, multi-task synergistic** framework for solving HSI-related tasks. Our HSI-GPT utilizes the VQVAE to discretize SMPL-X [33] body meshes $\mathbf{m}$ into LLM-interpretable tokens and embeds 3D scene $\mathcal{S}$ with pre-trained encoders (Sec. 3.1). In Sec. 3.2, we develop a versatile Large Scene-Motion-Language Model to equip the LLM with the HSI understanding and generation abilities. To generalize in HSI-related tasks, we devise a three-stage training scheme (Sec. 3.3) for multi-modal tokenizers training, scene-motion-language alignment, and instruction follow tuning.

3.1. Motion Tokenization and Scene Embedding

Motion Tokenization. In alignment with the LLM’s token-in and token-out nature, we represent articulated motions as discrete semantic-rich tokens with Vector Quantization Variational AutoEncoder (VQVAE). As shown in Fig. 2, the motion-aware VQVAE is composed of a motion en-

coder $\mathcal{E}_m$, a motion decoder $\mathcal{D}_m$ and a codebook $\mathcal{B}_m = \{b_1, b_2, \dots, b_N\} \in \mathbb{R}^{N \times d_c}$, containing N discrete semantic vectors. Given a SMPL-X-based 3D motion sequence $\mathbf{m} = \{m_1, m_2, \dots, m_T\} \in \mathbb{R}^{T \times d_m}$ with T frames, the encoder $\mathcal{E}_m$, built from several 1-D convolutional layers, first maps the input into a latent representation $e = \mathcal{E}_m(\mathbf{m})$. The estimated motion embedding e is then quantized to match discrete entries in the learned motion-aware codebook $\mathcal{B}_m$, a critical step for depicting human motions as a “body language”. The quantization formula is outlined below:

$$\mathbf{z}_m = \arg \min_{b_k \in \mathcal{B}} \|\mathcal{E}_m(\mathbf{m}) - b_k\|_2. \quad (1)$$

The quantized embedding $\mathbf{z}_m$ is mapped back into continuous motion $\hat{\mathbf{m}}$ by the decoder $\mathcal{D}_m$: $\hat{\mathbf{m}} = \mathcal{D}_m(\mathbf{z}_m)$. The total loss $\mathcal{L}_{\text{VQVAE}}$ consists of the reconstruction loss $\mathcal{L}_{\text{re}}$, the embedding loss $\mathcal{L}_{\text{emb}}$, and the commitment loss $\mathcal{L}_{\text{com}}$:

$$\begin{aligned} \mathcal{L}_{\text{VQVAE}} = \mathcal{L}_{\text{re}} + \mathcal{L}_{\text{emb}} + \mathcal{L}_{\text{com}} = & \|\mathcal{D}(\mathcal{E}(\mathbf{m})) - \mathbf{m}\|^2 \\ & + \|\text{sg}[\mathcal{E}(\mathbf{m})] - \mathbf{e}\|_2^2 + \beta \|\mathcal{E}(\mathbf{m}) - \text{sg}[\mathbf{e}]\|_2^2. \end{aligned} \quad (2)$$

Here, $\text{sg}(\cdot)$ is the stop-gradient operator for training stably and β adjusts the weight of the commitment loss.

Scene Embedding. The 3D scene is represented by RGB point cloud $\mathcal{S} \in \mathbb{R}^{N_s \times 6}$, where N_s denotes the total number of points. Each point in $\mathcal{S}$ consists of xyz coordinate and RGB color representation. Following Afford-Motion [53], we compute a scene affordance map $\mathbf{d}$ by measuring the l_2 -distance between 3D points in the scene $\mathcal{S}$ and the articulated body motions $\mathbf{m}$, yielding in a per-frame distance field $\mathbf{d} \in \mathbb{R}^{N_s \times J_m}$. Here, $J_m = d_m/3$ denotes the number of skeleton joints. The affordance map provides interaction priors and detailed 3D scene geometry, guiding human motions to conform to physical constraints. The scene and affordance map representations are embedded utilizing the pre-trained encoder as outlined in [53], denoting F_s and F_a .

3.2. Large Scene-Motion-Language Model

In HSI-GPT, a central LLM serves as the pivotal agent for 3D scene perception, HSI generation and understanding. Following the common practice, HSI-GPT takes an *encoder-LLM-decoder* architecture in popular MLLMs [24, 26, 74]. Previous methods [4, 19, 52, 53, 58] regard the 3D scene, human motion, and language as separate modalities. However, they struggle to delve into the complex trimodal relationships in specific scenes, thus hindering the generation of contextually coherent and semantically controlled motions. Although aligning text with one modality is relatively straightforward, integrating multiple modalities (e.g., scene, motion, and language) in a single framework presents ongoing difficulties. Our HSI-GPT tackles this issue through a two-step solution:

Scene-Language Interaction. To align the permutation-invariant 3D scene features with the textual embeddings of LLMs, we introduce a Multi-modal Interaction Aggregator (MIA). Given the computational overhead posed by extensive point clouds in 3D scene, MIA leverages a *fixed*-

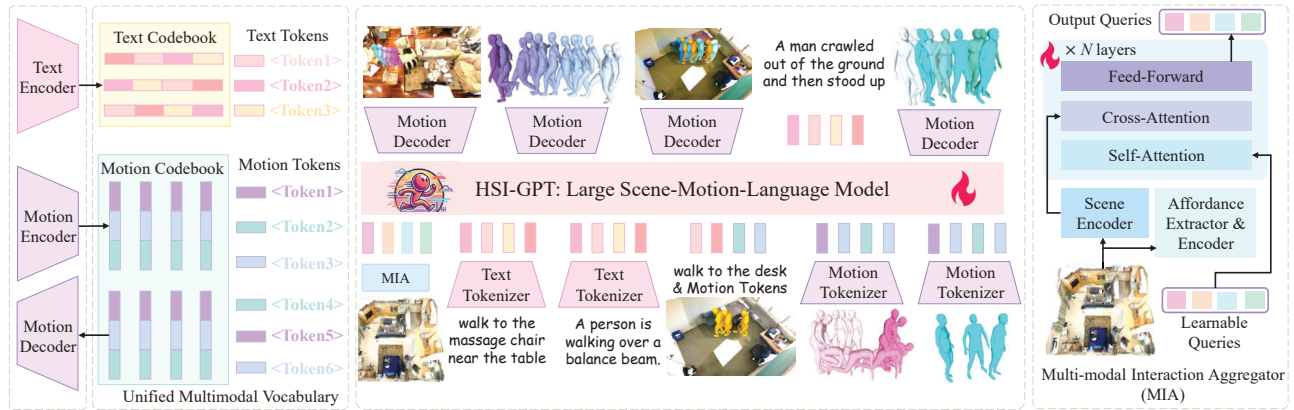


Figure 2. The overall framework of our generalist HSI-GPT in HSI-related tasks. It is composed of a multi-modal tokenizer (Sec. 3.1) and a Large Scene-Motion-Language Model (Sec. 3.2). With an enriched motion-language vocabulary and a Multi-modal Interaction Aggregator (MIA) module, HSI-GPT seamlessly aligns scene, motion, and language modalities utilizing fine-tuned LLMs.

length set of learnable query tokens $Q \in \mathbb{R}^{N_q \times d_q}$ to extract interaction-aware 3D scene features in a Q-Former [24] way. As shown in Fig. 2, the learnable tokens first interact with the affordance maps F_a through weight-shared self-attention layers, and cross-attend to task-agnostic scene embeddings F_s for enhancing scene understanding and task-specific interaction. These 3D visual tokens with compact representations, preserves 3D scene structure and affordance cues. Following LLaVA [26], two MLP layers project the output query tokens of the MIA module into the LLM’s space and perform dimensional alignment with the textual embeddings. During training, we use $\langle \text{scene_token} \rangle$ in the instruction templates as the placeholder for interaction-aware 3D scene features.

Extended Motion-Language Vocabulary. Just as natural language conveys intentions through vocabulary and syntax, human interacting with environments expresses intentions via a motion vocabulary. With structured patterns, complex actions can be articulated, forming a non-verbal “body language”. Therefore, our HSI-GPT develops an enriched multi-modal vocabulary (Fig. 2), **enabling seamless “any-to-any” translation between motion and language.**

Equipped with the motion tokenizer in Sec. 3.1, we categorize human motions $\mathbf{m}$ into discrete tokens. To accommodate the expanded motion vocabulary $\mathcal{B}_m$ to the original text vocabulary $\mathcal{B}_t$, we extend the LLM’s text embedding layer, initializing the newly incorporated parameters randomly. We then insert special tokens $\langle \text{motion} \rangle$ and $\langle \text{/motion} \rangle$ to indicate the start and end of the motion sequence. The scene features are not discretized, avoiding information loss caused by vector quantization. With the unified motion-language vocabulary $\mathcal{B} = \{\mathcal{B}_m, \mathcal{B}_t\}$ and harmonized scene-language representations, HSI-GPT **establishes a well-aligned scene-motion-language feature space**, allowing human motions to be both scene-aware and text-informed. Thus, we can formulate various HSI-related tasks with a general textual template (*Refer to Supp.*). The output of the LLM can flexibly denote

descriptive texts, human motion, or their combinations.

3.3. Training Strategy

The training process of HSI-GPT is partitioned into three stages. (1) *Multimodal Tokenizers Training* stage to represent human motion as discrete semantic tokens. (2) *Progressive Scene-Motion-Language Alignment* stage for aligning 3D scene, motion, and language modalities, and fostering synergy across various HSI-related tasks. (3) *Unified Instruction Tuning* stage, which improves the HSI-GPT’s instruction-following capability with task-aware prompts.

Multimodal Tokenizers Training. In Sec. 3.1, we first learn the discrete motion representations to align with the LLM’s token-in-token-out nature, guided by the objective in Eq. 2. Each human motion $\mathbf{m}$ in the 3D scene is represented as highly informative tokens. To further prevent the motion-aware codebook collapse during training, we employ Exponential Moving Average (EMA) and codebook reset techniques [63, 70]. We freeze the weights of the codebook in the following stages to ensure the stability of LLM training. Once optimized, the motion tokenizer is kept fixed.

Progressive Scene-Motion-Language Alignment. In this stage, we aim to align scene-motion-language modality and learn multi-modal world knowledge with the proposed HSI-GPT. As “text” serves as a highly semantic descriptor and is the most easily interpretable modality for LLMs, we propose leveraging texts as a bridge. Specifically, to learn the motion-language vocabulary, we thus utilize a mixture of language and motions corpus to train the LLM [41, 45, 57] in both unsupervised and supervised manners. With the LoRA adaption [16], HSI-GPT aligns “body language” semantics with textual embeddings, while preserving the LLM’s text generation capability. Inspired by current MLLMs [12, 26, 61, 62], we keep both the 3D visual encoder and LLM weights frozen. On the whole HSI dataset, we train the MIA model as well as the projection layer to

Model	Top 1	R-Precision $\uparrow$ Top 2	Top 3	FID $\downarrow$	MultiModal Dist. $\downarrow$	Diversity $\rightarrow$	MultiModality $\uparrow$
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	—
TM2T [10]	0.424 $\pm$.003	0.618 $\pm$.003	0.729 $\pm$.002	1.501 $\pm$.046	3.467 $\pm$.008	8.589 $\pm$.058	2.424 $\pm$.093
T2M [9]	0.457 $\pm$.002	0.639 $\pm$.003	0.740 $\pm$.003	1.067 $\pm$.002	3.340 $\pm$.008	9.188 $\pm$.002	2.090 $\pm$.083
MDM [44]	0.319 $\pm$.005	0.498 $\pm$.004	0.611 $\pm$.007	0.544 $\pm$.001	5.566 $\pm$.027	9.559 $\pm$.086	2.799 $\pm$.072
MotionDiffuse [64]	0.491 $\pm$.001	0.681 $\pm$.001	0.782 $\pm$.001	0.630 $\pm$.001	3.113 $\pm$.001	9.410 $\pm$.049	1.553 $\pm$.042
MLD [5]	0.481 $\pm$.003	0.673 $\pm$.003	0.772 $\pm$.002	0.473 $\pm$.013	3.169 $\pm$.010	9.724 $\pm$.082	2.413 $\pm$.079
T2M-GPT [63]	0.491 $\pm$.003	0.680 $\pm$.003	0.775 $\pm$.002	0.116 $\pm$.004	3.118 $\pm$.011	9.761 $\pm$.081	1.856 $\pm$.011
AttT2M [72]	0.499 $\pm$.003	0.690 $\pm$.002	0.786 $\pm$.002	0.112 $\pm$.006	3.038 $\pm$.007	9.700 $\pm$.090	2.452 $\pm$.051
MotionGPT $\ddagger$ [70]	0.364 $\pm$.005	0.533 $\pm$.003	0.629 $\pm$.004	0.805 $\pm$.002	3.914 $\pm$.013	9.972 $\pm$.026	2.473 $\pm$.041
MotionGPT $\ddagger$ [18]	0.492 $\pm$.003	0.681 $\pm$.003	0.733 $\pm$.006	0.232 $\pm$.008	3.096 $\pm$.008	9.528 $\pm$.071	2.008 $\pm$.084
MotionLLM $\ddagger$ [54]	0.482 $\pm$.004	0.672 $\pm$.003	0.770 $\pm$.002	0.491 $\pm$.019	3.138 $\pm$.010	9.838 $\pm$.244	—
Afford-Motion [53]	0.432 $\pm$.007	0.629 $\pm$.007	0.733 $\pm$.006	0.352 $\pm$.109	3.430 $\pm$.061	9.825 $\pm$.159	2.835 $\pm$.075
HSI-GPT (Ours)	0.495 $\pm$.002	0.694 $\pm$.003	0.788 $\pm$.004	0.187 $\pm$.004	3.058 $\pm$.013	9.845 $\pm$.026	2.234 $\pm$.022

Table 1. Quantitative results on the HumanML3D test set. $\ddagger$ denotes the text-based motion generation method based on LLM. $\pm$ represents a 95% confidence interval following [9]. $\rightarrow$ indicates that metrics improve as they get closer to “Real”. **Bold face** indicates the best result.

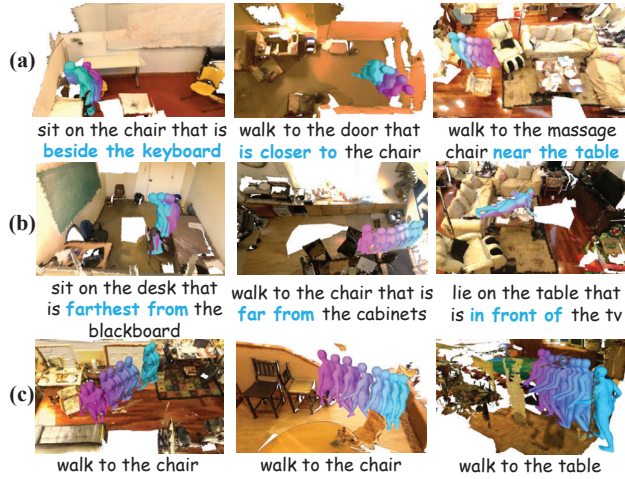


Figure 3. Visualization results for the HSI generation task on HUMANISE dataset. (a), (b) include spatially referring descriptions.

connect the 3D scene encoder and affordance map encoder with pre-trained textual embedding. The complete unification in LLM’s space enables HSI-GPT to bridge multi-modal gaps and combine semantic tokens into text representations, which are used interchangeably as inputs and outputs. The alignment makes the relationship of trimodal features apparent, *improving our HSI-GPT’s flexibility and scalability*.

Unified Instruction Tuning. After progressive trimodal alignment, the relationships among scene, motion, and language are clearly revealed. Similar to InstructGPT [31], the instruction tuning enhances LLMs’ instruction-following capabilities for handling diverse HSI-related tasks. During instruction tuning, we freeze the weight of the 3D scene and affordance encoder while updating the MIA weights and fine-tuning the LLM parameters using LoRA [16]. Leveraging datasets, *e.g.*, HUMANISE [52], HumanML3D [9], and PROX [13], we construct various task-specific instruction

prompts, which makes our proposed model **more versatile and user-friendly for multiple HSI-related tasks**. In contrast to [4, 52, 53] focus solely on text modality, HSI-GPT can integrate key-frame poses, texts as well as affordance maps with a unified instruction template. For instance, given a LLM $\mathcal{F}$, an instruction $\mathcal{I}$, the answer of LLM $\hat{\mathcal{P}} = \mathcal{F}(\mathcal{I})$ for multi-modal controlled HSI generation task is defined as:

% Multi-modal Controlled HSI Generation

Instruction ($\mathcal{I}$): {**Task Prompts:** “Given a 3D scene <scene>, please generate a sequence of motion to tokens matching the following human motion description given the init/last/key pose tokens.”} {**Control Conditions:** <Scene> 3D Scene conditions $\mathcal{S}$ </Scene> <Pos> Pose conditions $\mathcal{P}$ </Pos> Text condition $\mathcal{T}$ }
Answer ($\hat{\mathcal{P}}$):{Sequences of Human Motion Tokens $\mathcal{A}$ }

In the training phase, the general language modeling objective $\mathcal{L}_{\text{LoRA}}$ is applied to maximize the log-likelihood of motion distribution, which is mathematically formalized as:

$$\mathcal{L}_{\text{LoRA}} = - \sum \log p_{\theta}(x_t | x_{<t}, \mathcal{T}, \mathcal{S}, \mathcal{P}, \mathcal{A}). \quad (3)$$

4. Experiments

4.1. Datasets and Metrics

Datasets. The scene-agnostic *HumanML3D* [9] dataset collects 14,616 motion sequences from the AMASS [29] and HumanAct12 [8] benchmarks, along with 44,970 descriptions. To obtain affordance maps, we follow [53] to augment it by adding a floor. The *HUMANISE* [52] dataset pioneers as the first extensive and semantic-rich HSI dataset that aligns 19.6K high-quality motion sequences from AMASS with 643 indoor scenes from the ScanNet [6]. *PROX* [13] comprises 60 RGB-D videos documenting human interactions with real-world objects, captured by attached markers.

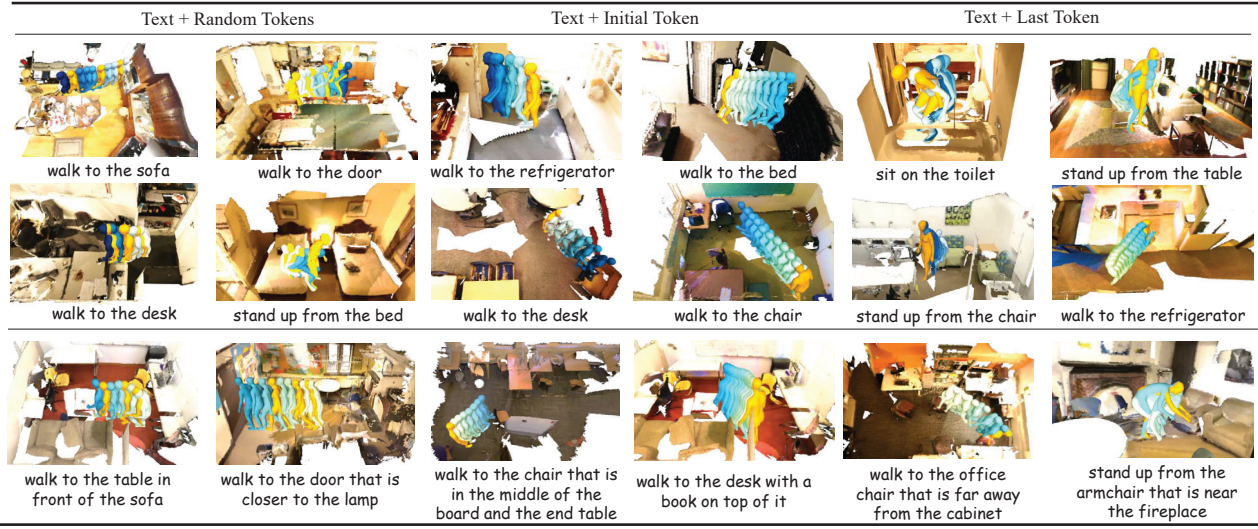


Figure 4. Gallery presenting the results of text-conditioned HSI generation under various control conditions, namely, *Text+Initial Motion Token*, *Text+Last Motion Token*, and *Text+Random Motion Tokens* (yellow poses) on the HUMANISE [9] benchmark. Best viewed in color.

Model	Goal Dist↓	APD↑	Contact↑	N-collision↑
cVAE [52]	0.422±.011	4.094±.013	84.06±.716	99.77±.004
1-stage@En [53]	0.326±.013	5.510±.019	76.11±.684	99.71±.014
1-stage@De [53]	0.185±.014	4.063±.020	86.43±.845	99.76±.014
2-stage@En [53]	0.156±.006	2.597±.008	95.86±.323	99.69±.007
2-stage@De [53]	0.156±.006	2.459±.009	96.04±.298	99.70±.005
MotionGPT† [18]	0.186±.008	3.461±.009	87.92±.526	99.70±.004
MotionGPT† [70]	0.358±.016	3.553±.027	84.41±.469	99.76±.018
HSI-GPT (Ours)	0.182±.008	3.492±.011	92.31±.478	99.82±.005

Table 2. Quantitative performance of text-conditioned HSI generation. † marks our reproduced results on HUMANISE [52] dataset.

Metrics. For the HumanML3D [9] dataset, we utilize metrics, *e.g.*, FID, Multi-modal Distance (MM Dist), and R-Precision (Top-1/2/3 motion & text retrieval accuracy), along with Diversity and Multi-Modality scores. In the *canonical HSI generation* task, we follow [52, 53], where the Average Pairwise Distance (APD) is used to assess motion diversity, alongside contact and non-collision scores as physics-based metrics. For the *HSI captioning* task, we include BLEU [32], ROUGE [25], CIDEr [47], and BertScore [67] to evaluate the quality of generated motion captions. For *HSI Completion* task, we utilize the Average Displacement Error (ADE) and Final Displacement Error (FDE) [18, 28, 60] as metrics¹.

4.2. Text-conditioned HSI Generation Task

As shown in Tab. 2, we conduct a comparative analysis of our HSI-GPT with cVAE [52] and Afford-Motion [53] on the HUMANISE. Our HSI-GPT is pre-trained on multiple HSI-related tasks **without additional task-specific fine-tuning**, highlighting versatility and scalability across diverse tasks. Despite being a generalist model, HSI-GPT effectively tunes

¹More details about the **Dataset**, **Metrics**, and **Implementation Details** are available in the *Supplementary Materials*.

only 1% of LLM parameters, yet still achieves competitive performance against specialist models [52, 53]. We also reproduce two LLM-based text-to-motion methods [18, 70] on HUMANISE dataset. Admirable performance on *contact* and *non-collision* metrics highlights that our model excels in avoiding implausible penetrations or collisions. To explain, HSI-GPT aligns 3D scenes, motion sequences, and instructions into the LLM’s feature space. Leveraging cross-task knowledge transfer abilities of LLMs, the HSI-GPT learns generalized interaction patterns between humans and 3D scenes. Via the LLM’s interpretation of physical realism and scene affordances, HSI-GPT generates motions that exhibit a high degree of diversity while maintaining naturalness. In Fig. 3, our results vividly visualize HSI-GPT’s ability for grounding motions in the 3D scene with desired semantics and interactions, under spatially-referenced descriptions and basic textual commands. It is attributed to multi-modal tokenization and pre-training alignment, granting LLM a comprehensive understanding of geometric relationships.

Tab. 1 showcases the quantitative results we obtained on HumanML3D dataset with augmented floors. In terms of *R-Precision* (**0.495**, **0.788** at Top1, Top3), *FID* (**0.187**), and *MultiModal Dist* (**3.058**) metrics, our HSI-GPT outperforms existing LLM-based text-to-motion methods [18, 54, 70], as well as several diffusion-based specialized models [44, 56, 66]. Moreover, HSI-GPT also achieves comparable diversity compared to scene-agnostic methods. Compared with [53], we underscore the advantage of LLM’s knowledge in interpreting 3D scene contexts and achieving motion-language alignment with a unified multi-modal vocabulary. In Fig. 6, the proposed method with fine-tuned LLM can also synthesize visually realistic motions, based on abstract descriptions, *e.g.*, “walks in an arc” and “walks in an s shape”.

Method	Affordance	R-Precision Top3 $\uparrow$	FID $\downarrow$	MultiModal Dist $\downarrow$	Diversity $\uparrow$	MModality $\uparrow$	Contact $\uparrow$	Non-Collision $\uparrow$
Real	—	$0.875 \pm .002$	$0.000 \pm .000$	$3.342 \pm .004$	$9.442 \pm .301$	—	—	—
Afford Motion @ Enc	$\times$	$0.500 \pm .044$	11.848 ± 1.634	$5.954 \pm .235$	$8.395 \pm .850$	$4.966 \pm .321$	46.64 ± 4.024	$99.88 \pm .018$
Afford Motion @ Dec	$\times$	$0.403 \pm .044$	$12.268 \pm .900$	$6.611 \pm .227$	$8.049 \pm .708$	$5.031 \pm .423$	26.75 ± 4.264	$99.93 \pm .023$
HSI-GPT (Ours)	$\times$	$0.569 \pm .020$	$10.567 \pm .082$	$5.054 \pm .184$	$8.542 \pm .798$	$5.148 \pm .458$	48.18 ± 3.754	$99.93 \pm .019$
Afford Motion @ Enc	$\checkmark$	$0.478 \pm .069$	7.887 ± 1.189	$6.226 \pm .261$	$7.935 \pm .857$	$5.195 \pm .356$	71.98 ± 2.542	$99.83 \pm .006$
Afford Motion @ Dec	$\checkmark$	$0.428 \pm .023$	12.027 ± 3.164	$6.412 \pm .204$	$7.603 \pm .715$	$4.966 \pm .353$	88.63 ± 2.975	$99.82 \pm .015$
HSI-GPT (Ours)	$\checkmark$	$0.587 \pm .016$	6.452 ± 1.095	$5.853 \pm .237$	$7.882 \pm .754$	$4.932 \pm .361$	85.47 ± 2.920	$99.88 \pm .025$

Table 3. Qualitative Results on Novel Evaluation Set [53]. “Real” denotes the metrics computed within the test set of HumanML3D dataset.

Method	HumanML3D [9]		HUMANISE [52]		
	FID $\downarrow$	MM Dist $\downarrow$	N-Collision $\uparrow$	APD $\uparrow$	Contact $\uparrow$
Text-Only	0.187	3.058	99.77	3.853	92.31
Text+Initial	0.185	3.194	99.80	3.802	92.67
Text+Last	0.234	3.261	99.80	3.847	92.45
Text+Random	0.180	3.015	99.82	3.709	92.86

Table 4. Investigation of HSI generation task on the HumanML3D and HUMANISE datasets with various control signal modalities.

LLM	Goal Dist $\downarrow$	APD $\uparrow$	Contact $\uparrow$	N-Collision $\uparrow$
Qwen2-1.5B	$0.332 \pm .019$	$4.025 \pm .009$	$84.79 \pm .573$	$99.76 \pm .010$
	$0.262 \pm .008$	$4.172 \pm .010$	$87.62 \pm .605$	$99.80 \pm .009$
Gemma-2B	$0.305 \pm .012$	$4.011 \pm .007$	$86.08 \pm .688$	$99.85 \pm .006$
	$0.245 \pm .008$	$4.199 \pm .008$	$88.04 \pm .561$	$99.88 \pm .002$
Gemma-7B	$0.285 \pm .017$	$3.746 \pm .008$	$87.25 \pm .604$	$99.76 \pm .012$
	$0.237 \pm .013$	$3.952 \pm .006$	$89.23 \pm .427$	$99.80 \pm .009$
LLaMA 3-8B	$0.214 \pm .009$	$3.855 \pm .009$	$88.73 \pm .465$	$99.82 \pm .008$
	$0.182 \pm .008$	$3.492 \pm .011$	$92.31 \pm .478$	$99.82 \pm .005$

Table 5. Ablations on the effects of foundational LLM on text-conditioned HSI generation, evaluated on the HUMANISE dataset.

4.3. Experiments on Other HSI-related Tasks

Multi-modal Controlled HSI Generation. Beyond text inputs, HSI-GPT also incorporates key-frame human poses as an additional control modality. Fig. 4 visually exemplifies the remarkable naturalness of generated motions in 3D scenes. Incorporating key-frame poses—*e.g.*, initial, last, or random poses) as additional control modalities—provides precise motion guidance. In Tab. 4, the physical plausibility and coherence of generated motions in 3D scenes are enhanced, as evidenced by the improved *Non-collision* and *Contact* accuracy on HUMANISE [52] dataset. Notably, the reduced diversity in *APD* is primarily due to constrained motion trajectories and strict pre-defined pose adherence. On HumanML3D [9], the lower *FID* and *MM Dist* scores for Text+Initial and Text+Random settings suggest that pose controls produce motions that are physically realistic and close-aligned with textual description. Instead, the Text+Last Pose results in an excessive goal-oriented focus, leading HSI-GPT to overlook the smoothness and diversity of motions.

HSI Captioning Task. We first propose the HSI caption-

ing task, involving generating a motion description from a given motion sequence and 3D scene representation. Tab. 6 presents results on the augmented HumanML3D, demonstrate the superiority of our HSI-GPT, compared to existing motion captioning approaches, *e.g.*, TM2T [10] and MotionGPT [18]. HSI-GPT achieves the highest *R-Precision* scores, with Top1 at **0.551** and Top3 at **0.838**, highlighting the seamless alignment of scene, motion, and textual modalities as well as precision in captions retrieval. Our HSI-GPT also demonstrates exceptional performance in ROUGE (38.4) and CIDEr (30.8), indicating a high degree of similarity to human-annotated captions. Approaching human motion as a form of “body language”, HSI-GPT learns a unified motion-language vocabulary to generate more contextually coherent motion descriptions while preserving diversity.

Generalized HSI Completion Task. We investigate the motion in-between and motion prediction task on the HumanML3D [9] with augmented 3D floors, a setup we refer to as *HSI completion*. To tailor the HSI-GPT for the HSI completion task, we design motion-only instruction prompts during the LLM’s fine-tuning. Similar to [18, 28, 60, 69], we input around the first 20% of motion sequences as conditions for the prediction task, while intentionally masking 50% of the motions for the in-between task. As detailed in Sec. 4.1, we employ metrics, *e.g.*, *FID*, *ADE*, and *FDE* to evaluate the motion completion performance. As shown in Tab. 7, compared with MDM [44] and MotionGPT [18], the proposed HSI-GPT incorporates affordance maps as joint trajectories, facilitate contextually appropriate motion completions.

4.4. Comparisons on Novel Evaluation Set

We herein investigate the generalizability of our HSI-GPT in handling new scenes and language descriptions on the Novel Evaluation Set [53]. As shown in Tab. 3, our method obtains impressive *R-Precision* and *FID* scores while maintaining high *Diversity*. By aligning scene-motion-language feature space, HSI-GPT taps into the generalization of conceptualized text to generate semantically consistent human motions in various 3D scenes. Further, with affordance maps as interaction priors and scene geometry, HSI-GPT can synthesize physically plausible motions, validated through the *contact*, and *Non-Collision* scores. Fig. 5 provides remarkable visual-

Method	R-Precision $\uparrow$			MultiModal Dist $\downarrow$	BLEU-1 $\uparrow$	BLEU-4 $\uparrow$	ROUGE $\uparrow$	CIDEr $\uparrow$	BertScore $\uparrow$
	Top1	Top2	Top3						
TM2T* [10]	0.516	0.720	0.823	2.935	48.9	7.00	<u>38.1</u>	16.8	<u>32.2</u>
MotionGPT* [18]	<u>0.543</u>	—	<u>0.827</u>	<u>2.821</u>	48.2	<u>12.5</u>	37.4	<u>29.2</u>	32.4
HSI-GPT (LLaMA 3-8B)	0.551	0.735	0.838	2.746	<u>48.6</u>	14.2	38.4	30.8	31.8

Table 6. Experiments of HSI captioning task on the augmented HumanML3D [9]. Results marked with * are from MotionGPT [18].

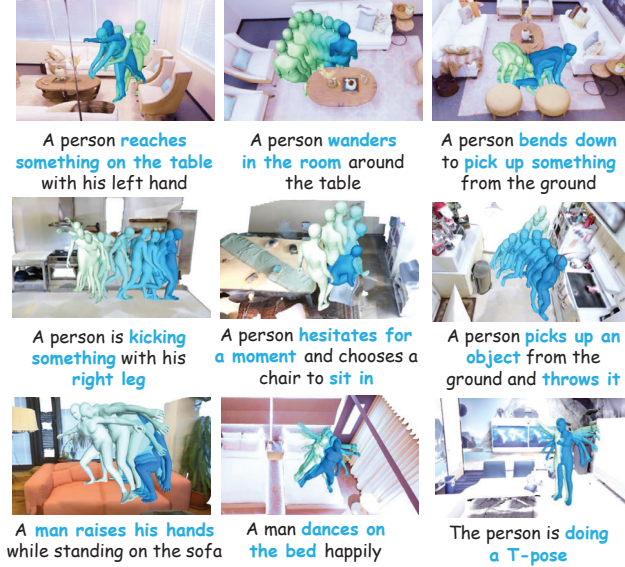


Figure 5. Qualitative results on generalization evaluation set [53]. HSI-GPT can generate high-fidelity motions in unseen 3D scenes.

Methods	Motion Prediction				Motion In-Between		
	FID	Diversity	ADE	FDE	FID	Diversity	ADE
Real	0.002	9.503	—	—	0.002	9.503	—
MDM	6.031	7.813	5.446	8.561	2.698	8.420	3.787
MotionGPT	<u>0.905</u>	<u>8.972</u>	<u>4.745</u>	<u>6.040</u>	0.214	9.560	<u>3.762</u>
HSI-GPT	0.572	9.491	4.586	5.893	<u>0.468</u>	<u>9.357</u>	3.745

Table 7. Evaluation of generalized HSI completion using our HSI-GPT (LLaMA 3-8B) on the HumanML3D augmented with floors. The underlined results indicate the second-best performance.

izations generated with unseen descriptive texts and scenes. HSI-GPT learns universal interaction principles with multi-task training, rather than specific motions tied to particular scenes, which effectively transfers in novel 3D scenes.

4.5. Ablation Study

As shown in Tab. 5, we conduct ablations to examine how different types and model scales of foundational LLM affect the text-conditioned HSI generation task on the HUMANISE [52] dataset. The **colored cells** indicate that the scene encoder uses embedded features from the Point-Transformer’s decoder layers. Among these, LLaMA 3-8B [7] with 89M LoRA parameters, emerges as the best performer on all metrics, as evidenced by **0.182** goal distance score and a **92.31%** contact score. We conclude that larger LLMs, with enhanced

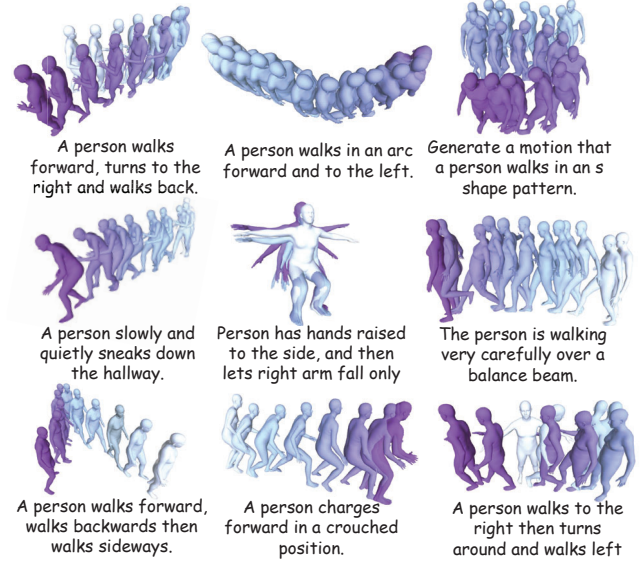


Figure 6. Visualization samples for the text-conditioned HSI generation task using the augmented HumanML3D [9] dataset.

representational capacity, take into account physical feasibility of motions, 3D scene’s affordance, and linguistic cues. Such a capability guarantees physically plausible and semantically consistent motions. Further, LoRA-based fine-tuning preserves the extensive cross-task knowledge of LLM, with its plug-and-play nature providing increased adaptability.

5. Conclusion

For the first time, we propose HSI-GPT, a general-purpose Large Scene-Motion-Language Model for text-conditioned human scene interaction task. HSI-GPT, *with remarkable flexibility and scalability*, integrates diverse control modalities (3D scenes, textual commands, human motions, and scene affordances) to unified instruction templates. Building on top of LLMs, HSI-GPT acts as a HSI-generalist, addresses multi-modal controlled HSI generation, HSI understanding, and general HSI completion tasks. With a three-stage training pipeline, HSI-GPT is capable of aligning the scene-motion-language modalities and generalizing to various HSI-related tasks. Extensive experiments demonstrate that our HSI-GPT follows human instructions and performs competitively across multiple HSI-related tasks, compared to diffusion-based methods. We envision that our HSI-GPT will contribute positively to the community and pave the way for versatile 3D HSI foundational models and Embodied AI.

6. Acknowledgment

This work is supported by the research fund under Grant No. 20242910035 from the Tsinghua University-Jiangsu CRRC Digital Technology Co.,Ltd. Joint Research Center for Data Driven Intelligence of Industry.

References

- [1] Chaitanya Ahuja and Louis-Philippe Morency. Language2pose: Natural language grounded pose forecasting. In *2019 International Conference on 3D Vision (3DV)*, pages 719–728. IEEE, 2019. 2
- [2] Nikos Athanasiou, Mathis Petrovich, Michael J Black, and Gül Varol. Teach: Temporal action composition for 3d humans. In *2022 International Conference on 3D Vision (3DV)*, pages 414–423. IEEE, 2022. 2
- [3] 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*, pages 387–404. Springer, 2020. 3
- [4] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1855–1866, 2024. 2, 3, 5
- [5] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 18000–18010, 2023. 2, 3, 5
- [6] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5828–5839, 2017. 5
- [7] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 8
- [8] 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 *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2021–2029, 2020. 5
- [9] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5152–5161, 2022. 2, 5, 6, 7, 8
- [10] 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 *European Conference on Computer Vision*, pages 580–597. Springer, 2022. 5, 7, 8
- [11] Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, and Li Cheng. Momask: Generative masked modeling of 3d human motions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1900–1910, 2024. 2
- [12] Jiaming Han, Kaixiong Gong, Yiyuan Zhang, Jiaqi Wang, Kaipeng Zhang, Dahua Lin, Yu Qiao, Peng Gao, and Xiangyu Yue. Onellm: One framework to align all modalities with language. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 26584–26595, 2024. 4
- [13] Mohamed Hassan, Duygu Ceylan, Ruben Villegas, Jun Saito, Jimei Yang, Yi Zhou, and Michael J Black. Stochastic scene-aware motion prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 11374–11384, 2021. 3, 5
- [14] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 3
- [15] Fangzhou Hong, Mingyuan Zhang, Liang Pan, Zhongang Cai, Lei Yang, and Ziwei Liu. Avatarclip: Zero-shot text-driven generation and animation of 3d avatars. *arXiv preprint arXiv:2205.08535*, 2022. 2
- [16] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. 2, 4, 5
- [17] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 16750–16761, 2023. 3
- [18] Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. *Advances in Neural Information Processing Systems*, 36:20067–20079, 2023. 5, 6, 7, 8
- [19] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1737–1747, 2024. 3
- [20] Korrawe Karunratanakul, Konpat Preechakul, Supasorn Suwajanakorn, and Siyu Tang. Guided motion diffusion for controllable human motion synthesis. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2151–2162, 2023. 3
- [21] Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2
- [22] Hanyang Kong, Kehong Gong, Dongze Lian, Michael Bi Mi, and Xinchao Wang. Priority-centric human motion generation in discrete latent space. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 14806–14816, 2023. 2
- [23] 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 *Proceedings of the IEEE International Conference on Computer Vision*, pages 11293–11302, 2021. 2
- [24] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023. 2, 3, 4

- [25] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004. 6
- [26] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 2, 3, 4
- [27] Mingshuang Luo, Ruibing Hou, Hong Chang, Zimo Liu, Yaowei Wang, and Shiguang Shan. M3gpt: An advanced multimodal, multitask framework for motion comprehension and generation. *arXiv preprint arXiv:2405.16273*, 2024. 3
- [28] Hengbo Ma, Jiachen Li, Ramtin Hosseini, Masayoshi Tomizuka, and Chiho Choi. Multi-objective diverse human motion prediction with knowledge distillation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8161–8171, 2022. 6, 7
- [29] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5442–5451, 2019. 5
- [30] OpenAI. Chatgpt (mar 14 version) [large language model]. <https://chat.openai.com/chat/>, 2023. 3
- [31] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022. 5
- [32] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, 2002. 6
- [33] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10975–10985, 2019. 3
- [34] Mathis Petrovich, Michael J Black, and Gül Varol. Action-conditioned 3d human motion synthesis with transformer vae. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10985–10995, 2021. 2
- [35] Mathis Petrovich, Michael J Black, and Gül Varol. Temos: Generating diverse human motions from textual descriptions. In *European Conference on Computer Vision*, pages 480–497. Springer, 2022. 2
- [36] 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*, pages 8748–8763. PMLR, 2021. 2
- [37] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020. 3
- [38] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022. 1
- [39] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 3
- [40] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 1
- [41] Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivi re, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024. 3, 4
- [42] 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*, pages 358–374. Springer, 2022. 2
- [43] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *The Eleventh International Conference on Learning Representations*, 2023. 2, 3
- [44] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *The Eleventh International Conference on Learning Representations*, 2023. 5, 6, 7
- [45] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timoth e Lacroix, Baptiste Rozi re, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models, 2023. 3, 4
- [46] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 3
- [47] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4566–4575, 2015. 6
- [48] Jiashun Wang, Huazhe Xu, Jingwei Xu, Sifei Liu, and Xiao-long Wang. Synthesizing long-term 3d human motion and interaction in 3d scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9401–9411, 2021. 2
- [49] Jingbo Wang, Sijie Yan, Bo Dai, and Dahua Lin. Scene-aware generative network for human motion synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12206–12215, 2021. 2, 3
- [50] Jingbo Wang, Yu Rong, Jingyuan Liu, Sijie Yan, Dahua Lin, and Bo Dai. Towards diverse and natural scene-aware 3d human motion synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 20460–20469, 2022. 3
- [51] Yuan Wang, Zhao Wang, Junhao Gong, Di Huang, Tong He, Wanli Ouyang, Jile Jiao, Xuetao Feng, Qi Dou, Shixiang Tang,

- et al. Holistic-motion2d: Scalable whole-body human motion generation in 2d space. *arXiv preprint arXiv:2406.11253*, 2024. [2](#)
- [52] Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. Humanise: Language-conditioned human motion generation in 3d scenes. *Advances in Neural Information Processing Systems*, 35:14959–14971, 2022. [2](#), [3](#), [5](#), [6](#), [7](#), [8](#)
- [53] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 433–444, 2024. [2](#), [3](#), [5](#), [6](#), [7](#), [8](#)
- [54] Qi Wu, Yubo Zhao, Yifan Wang, Yu-Wing Tai, and Chi-Keung Tang. Motionllm: Multimodal motion-language learning with large language models. *arXiv preprint arXiv:2405.17013*, 2024. [3](#), [5](#), [6](#)
- [55] Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *arXiv preprint arXiv:2309.05519*, 2023. [3](#)
- [56] Chen Xin, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, Jingyi Yu, and Gang Yu. Executing your commands via motion diffusion in latent space. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023. [2](#), [6](#)
- [57] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024. [4](#)
- [58] Hongwei Yi, Justus Thies, Michael J Black, Xue Bin Peng, and Davis Rempe. Generating human interaction motions in scenes with text control. In *European Conference on Computer Vision*, pages 246–263. Springer, 2025. [2](#), [3](#)
- [59] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation, 2022. [1](#)
- [60] Ye Yuan and Kris Kitani. Dlow: Diversifying latent flows for diverse human motion prediction. In *European Conference on Computer Vision*, pages 346–364. Springer, 2020. [6](#), [7](#)
- [61] Jun Zhan, Junqi Dai, Jiasheng Ye, Yunhua Zhou, Dong Zhang, Zhigeng Liu, Xin Zhang, Ruibin Yuan, Ge Zhang, Linyang Li, et al. Anygpt: Unified multimodal llm with discrete sequence modeling. *arXiv preprint arXiv:2402.12226*, 2024. [4](#)
- [62] Hao Zhang, Hongyang Li, Feng Li, Tianhe Ren, Xueyan Zou, Shilong Liu, Shijia Huang, Jianfeng Gao, Chunyuan Li, Jainwei Yang, et al. Llava-grounding: Grounded visual chat with large multimodal models. In *European Conference on Computer Vision*, pages 19–35. Springer, 2025. [4](#)
- [63] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023. [2](#), [4](#), [5](#)
- [64] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model, 2022. [2](#), [3](#), [5](#)
- [65] Mingyuan Zhang, Xinying Guo, Liang Pan, Zhongang Cai, Fangzhou Hong, Huirong Li, Lei Yang, and Ziwei Liu. Remodiffuse: Retrieval-augmented motion diffusion model. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 364–373, 2023. [2](#), [3](#)
- [66] 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 Transactions on Pattern Analysis and Machine Intelligence*, 2024. [6](#)
- [67] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019. [6](#)
- [68] 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*, pages 518–535. Springer, 2022. [3](#)
- [69] Yan Zhang, Michael J Black, and Siyu Tang. We are more than our joints: Predicting how 3d bodies move. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3372–3382, 2021. [7](#)
- [70] Yaqi Zhang, Di Huang, Bin Liu, Shixiang Tang, Yan Lu, Lu Chen, Lei Bai, Qi Chu, Nenghai Yu, and Wanli Ouyang. Motiongpt: Finetuned llms are general-purpose motion generators, 2023. [2](#), [3](#), [4](#), [5](#), [6](#)
- [71] Kaifeng Zhao, Yan Zhang, Shaoqi Wang, Thabo Beeler, and Siyu Tang. Synthesizing diverse human motions in 3d indoor scenes. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 14738–14749, 2023. [3](#)
- [72] Chongyang Zhong, Lei Hu, Zihao Zhang, and Shihong Xia. Att2m: Text-driven human motion generation with multi-perspective attention mechanism. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 509–519, 2023. [5](#)
- [73] Zixiang Zhou and Baoyuan Wang. Ude: A unified driving engine for human motion generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5632–5641, 2023. [2](#)
- [74] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigt-4: Enhancing vision-language understanding with advanced large language models, 2023. [2](#), [3](#)
- [75] Lingting Zhu, Xian Liu, Xuanyu Liu, Rui Qian, Ziwei Liu, and Lequan Yu. Taming diffusion models for audio-driven co-speech gesture generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10544–10553, 2023. [2](#)
- [76] Wenlin Zhuang, Congyi Wang, Jinxiang Chai, Yangang Wang, Ming Shao, and Siyu Xia. Music2dance: Dancenet for music-driven dance generation. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 18(2):1–21, 2022. [2](#)