HIS-GPT: Towards 3D Human-In-Scene Multimodal Understanding

Jiahe Zhao[1,2], Ruibing Hou[1]*, Zejie Tian[3], Hong Chang[1,2], Shiguang Shan[1,2]
[1]State Key Laboratory of AI Safety, Institute of Computing Technology, CAS, China
[2]University of Chinese Academy of Sciences (CAS), China [3]Communication University of China
zhaojiahe22@mails.ucas.ac.cn, {houruibing, changhong, sgshan}@ict.ac.cn
2021211023003@mails.cuc.edu.cn

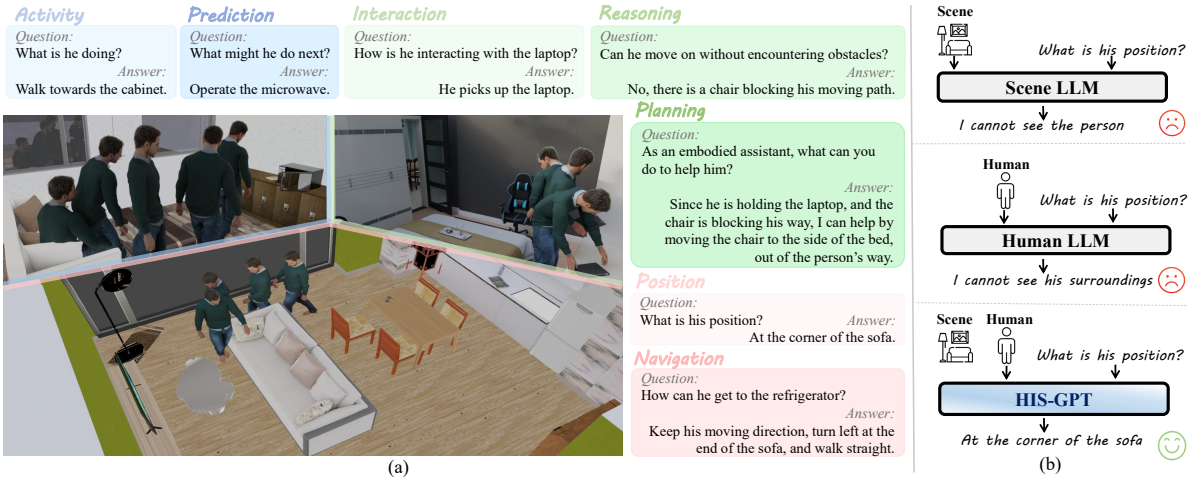


Figure 1. (a) **Illustration of HIS-QA task,** which understands human behaviors in scene context. HIS-QA tasks span from basic perception tasks, such as recognizing human activity, interaction, and position in scene, to higher order functions like prediction, reasoning, planning, and navigation, facilitating embodied intelligence in real world. (b) **Illustration of HIS-GPT**. Unlike previous models that focus solely on either scene or human understanding, HIS-GPT could jointly perceive scene and human modalities to tackle the challenges of HIS-QA.

Abstract

We propose a new task to benchmark human-in-scene understanding for embodied agents: Human-In-Scene Question Answering (HIS-QA). Given a human motion within a 3D scene, HIS-QA requires the agent to comprehend human states and behaviors, reason about its surrounding environment, and answer human-related questions within the scene. To support this new task, we present HIS-Bench, a multimodal benchmark that systematically evaluates HIS understanding across a broad spectrum, from basic perception to commonsense reasoning and planning. Our evaluation of various vision-language models on HIS-Bench reveals significant limitations in their ability to handle HIS-QA tasks. To this end, we propose HIS-GPT, the first foundation model for HIS understanding. HIS-GPT integrates 3D scene context and human motion dynamics into large language models while incorporating specialized mechanisms to capture human-scene interactions. Extensive experiments demonstrate that HIS-GPT sets a new state-of-the-art on HIS-QA tasks. We hope this work inspires future research on human behavior analysis in 3D scenes, advancing embodied AI and world models. Codes and data will be available at https://github.com/ZJHTerry18/HumanInScene.

1. Introduction

In recent years, intelligent systems for 3D vision-language understanding have witnessed remarkable progress [12, 25, 48, 49, 51, 56], largely driven by the advancements in Large Language Models (LLMs) [15–17, 21, 58]. Specifically, 3D scene LLMs [24, 30, 32, 55] excel in tasks such as captioning and grounding within 3D layouts, whereas 3D human LLMs [34, 39, 43, 65] exhibit strong capabilities in open-ended interpretations of human poses and motions. By em-

*Corresponding author.

bracing 3D world, these models significantly promote the developments in robotics and embodied AI.

Despite significant progress in separately perceiving 3D scenes and humans, a critical yet underexplored challenge remains: **human-in-scene (HIS)** understanding. This task requires an agent to jointly comprehend human subjects and their surrounding environments to capture intricate interactions and relationships. Such capability is essential for accurately recognizing fundamental human states (*e.g.*, *positioned in front of the TV)* and actions (*e.g.*, *sit on a chair*) in real-world scenarios. With effective HIS understanding, embodied agents could reason, predict, and react based on their observations of human-scene dynamics, thereby serving as versatile assistants in applications such as household robots. However, the current limitations of 3D LLMs to integrate human and scene perception largely hinder further advancements in embodied intelligence.

To bridge this critical gap, we introduce **HIS-QA**, a novel task for Human-In-Scene Question Answering, where an agent answers questions about human states and behaviors within a 3D scene, as depicted in Fig. 1 (a). To systematically evaluate this task, we propose **HIS-Bench**, the first multimodal benchmark tailored for HIS understanding. As shown in Tab. 1, HIS-Bench differs from previous benchmarks by integrating both human and scene modalities for open-ended, language-guided understanding. A major challenge in constructing HIS-Bench is the lack of detailed textual annotations in existing HIS datasets [4, 18, 28, 29, 35, 63], which primarily provide coarse action labels (*e.g.*, *walking*, *sitting*). Additionally, the intrinsic 3D spatial complexity of human-scene interactions makes it impractical to generate precise annotations using proprietary models like GPT-4o [33]. To overcome this limitation, we develop a specialized data annotation pipeline that combines advanced 3D understanding tools with rule-based algorithms for text annotations. This pipeline enables the generation of rich annotations covering human actions, scene properties, and human-scene interactions. Built upon these detailed annotations, HIS-Bench comprises 800 questions organized hierarchically into 3 general abilities, 7 core tasks, and 16 sub-tasks, spanning a broad spectrum from basic human activity perception to advanced reasoning, prediction, and planning. This comprehensive benchmark establishes a new standard for evaluating HIS understanding.

Utilizing HIS-Bench, we systematically evaluate HIS-QA with existing vision-language models [7, 32, 33, 36]. We observe that existing models fall short in HIS understanding, largely due to their insufficient capacity for jointly modeling human-scene characteristics in 3D space. To address the above limitation, we propose **HIS-GPT**, a multimodal large language model tailored for HIS understanding. As shown in Fig. 1 (b), HIS-GPT fundamentally differs from prior 3D LLMs [11, 30, 32] by jointly interpreting 3D scenes and humans. Specifically, HIS-GPT integrates a scene encoder [64] and a motion encoder [43] to extract structured representations of 3D environments and human motions. These representations are subsequently processed by the core LLM [15], enabling seamless fusion of scene and motion cues to enhance capabilities on HIS tasks.

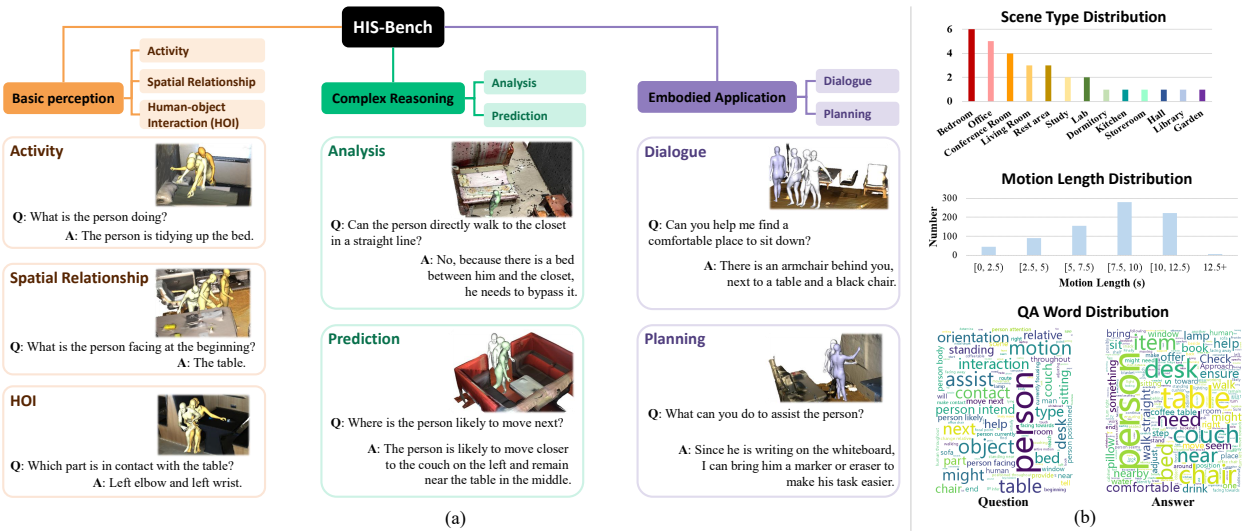
Beyond previous 3D LLMs that focus on perceiving a single modality (either human or scene), a key challenge in HIS understanding lies in accurately modeling human-scene interactions. To this end, HIS-GPT introduces two critical components. On one hand, an **Auxiliary Interaction (AInt)** module enhances interactive cues within each modality, through incorporating multiple training objectives that require a joint understanding of human and their surroundings. By enforcing these constraints, HIS-GPT is guided to learn enriched, contextually aware representations of human-scene interactions. On the other hand, a **Layout-Trajectory Position Encoding (LTP)** module generates position embeddings by encoding the spatial distribution of major objects in the scene layout, along with the temporal trajectories of human motion at each timestamp. By infusing fine-grained spatiotemporal knowledge into latent representations of scene and human, LTP module enhances both modalities, effectively capturing the dynamic interplay between human motions and 3D environments.

To our knowledge, HIS-GPT is the first approach to address the tasks of human-in-scene understanding. Extensive experiments demonstrate that HIS-GPT achieves state-of-the-art performance on HIS-QA task, establishing a strong foundation for future research.

2. Related Work

3D Scene-Language Understanding. 3D scene-language understanding is a critical technique for agents to interact with the real world. It contains a wide range of tasks, including 3D captioning [10, 13], 3D visual grounding [3, 54, 59], and 3D question answering [5, 47]. Recent approaches adopt LLMs to tackle various 3D scene understanding tasks within a unified framework [11, 24, 30–32, 55, 62], benefiting from the synergies of multi-task learning.

Despite their success in interpreting 3D scenes, these models are confined to tasks centered solely on scenes, and cannot handle 3D environments that include human elements. Some efforts [44, 53, 61] explore situated scene understanding by assuming the presence of a subject in 3D scenes. However, these approaches rely on explicit text inputs or first-person views to establish a subject's location, while also lacking full-body pose representation. In contrast, our proposed HIS-QA requires to directly model both 3D scene and humans from vision modalities, while being aware of the human pose. This setting allows for a more comprehensive perception of human states within the scene.



(a) (b)

Figure 2. (a) **Task taxonomy and data samples of HIS-Bench.** HIS-Bench is structured into 3 general abilities and 7 core tasks. (b) **Statistics of HIS-Bench.** HIS-Bench is diverse in term of scene types, motion lengths, and word distributions.

Table 1. Overview of existing benchmarks related to 3D scene and human. 'mo.gen.', 'det.', 'cap.' and 'q.a.' refers to motion generation, detection, caption, and question-answering, respectively.

Benchmark	Task	Modalities			Language task	
		Scene	Human	Language	Open-ended	Text generation
TRUMANS [35]	mo.gen.	✓	✓	✗	-	-
ScanRefer [8]	det.	✓	✗	✓	✗	template
SQA3D [44]	q.a.	✓	✗	✓	✗	template
OpenEQA [46]	q.a.	✓	✗	✓	✓	human
Motion-X [40]	cap.	✗	✓	✓	✓	auto
MoVid-Bench [9]	q.a.	✗	✓	✓	✓	auto
HIS-Bench(Ours)	q.a.	✓	✓	✓	✓	human&auto

3D Human-Language Understanding. 3D human-language understanding primarily focuses on recognizing human poses or motions [19, 20, 27, 37]. Recently, several works introduce LLMs to interpret human pose and motions [9, 23, 34, 39, 43, 57, 65], addressing tasks like motion captioning [26, 40] and question-answering [9, 22]. However, these approaches overlook the environmental context of humans, constraining their ability to comprehensively recognize human status. To overcome this limitation, we present HIS-GPT, which processes human motions alongside scene contexts, enabling a more comprehensive understanding of human behavior in real-world environments.

3. HIS-Bench

To explore the problem of understanding human behaviors in 3D scenarios, we propose HIS-QA, a new task for addressing human-in-scene understanding of AI agents. A problem instance in HIS-QA can be formulated as a quadruplet $\langle \mathcal{S}, \mathcal{M}, \mathcal{Q}, \mathcal{A} \rangle$. $\mathcal{S}$ denotes 3D scene in point cloud. $\mathcal{M}$ denotes 3D human motion sequence, with each frame characterized by a SMPL pose [41]. $\mathcal{Q}$ refers to a natural language question and $\mathcal{A}$ is the ground-truth answer. The agent is tasked with generating an answer $\hat{\mathcal{A}} = \text{Agent}(\mathcal{S}, \mathcal{M}, \mathcal{Q})$ that closely aligns with the true answer $\mathcal{A}$.

However, existing 3D scene QA [5, 44] and 3D human QA benchmarks [9, 22] focus solely on scene or human understanding in isolation, overlooking human-scene interactions. To address this gap, we propose HIS-Bench, the first dedicated benchmark for HIS-QA. Next, we introduce the task taxonomy and data generation pipeline for HIS-Bench. More details on constructing HIS-Bench are provided in Appendix A, and additional examples of HIS-Bench are provided in Appendix B.

3.1. Task Taxonomy

As shown in Fig. 2, HIS-Bench defines a structured taxonomy of benchmark tasks, encompassing three fundamental abilities: ***basic perception***, ***complex reasoning*** and ***embodied applications***. These categories comprise 7 core tasks, further divided into 16 sub-tasks:

- **Activity.** (1) *Single Activity:* Recognize the human activity within the scene. (2) *Sequential Activity:* Recognize the human activity before or after a specific action.
- **Spatial Relationship.** (3) *Human Position:* Identify the human's precise location in the scene. (4) *Body Orientation:* Identify the human body's orientation relative to the scene. (5) *Object Orientation:* Identify the object that is at a given orientation relative to the human.
- **Human-object Interaction.** (6) *Interaction Type:* Recognize the type of human-object interaction. (7) *Interacting Object:* Recognize the object the human is interacting with. (8) *Contact Part:* List the human body parts in contact with a given object.
- **Analysis.** (9) *Focus Analysis:* Infer the object or area the human is attending to. (10) *Situated Analysis:* Deduce scene-related knowledge from the human's perspective, such as affordance and approachability.

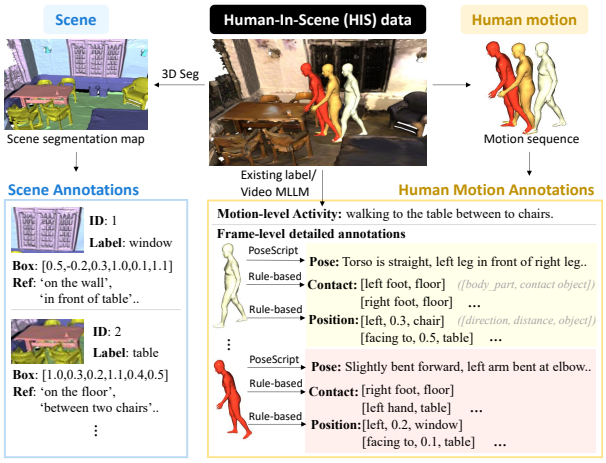


Figure 3. Text annotation pipeline for HIS data. For scene annotations, we segment the 3D scene to derive instance-level labels, bounding boxes, and reference expressions. For motion annotations, we obtain motion-level activities from existing labels or video MLLMs. Additionally, expert models and rules are used to generate frame-level annotations, including pose, human-scene contact, and human position.

- **Prediction.** (11) *Intent Prediction:* Predict the human's next intended activity. (12) *Movement Prediction:* Predict the human's future trajectories and spatial positions.
- **Dialogue.** (13) *Situated Dialogue:* Complete a conversation with the human regarding the scene context.
- **Planning.** (14) *High-level Task:* Provide a general plan to assist the human based on their status. (15) *Low-level Task:* Provide step-by-step instructions for assisting the human. (16) *Navigation:* Provide a route to guide the human towards a specified destination.

3.2. Data Generation Pipeline

Text Annotation. Acquiring multimodal resources for 3D HIS data is challenging, as existing datasets primarily contain 3D scene-language [24, 54] or human-language [40, 45], but lack the necessary human-in-scene descriptions essential for HIS understanding. To bridge this gap, we propose a multi-faceted annotation pipeline that generates rich and comprehensive human-scene descriptions.

As shown in Fig. 3, our multi-faceted annotation pipeline comprises scene annotations and human motion annotations. For **Scene Annotations**, following [24], we utilize 3D scene segmentation tools [50] and visual caption models [38] to generate semantic labels, 6D bounding boxes, and referring expressions for key objects in the scene. For **Human Motion Annotations**, we first generate *motion-level activities*: for scene data with recorded videos, a video captioner [36] is prompted to generate descriptions on human activities. For datasets lacking video recordings, we directly adopt the action labels provided in the original annotations. Additionally, we generate *frame-level detailed annotations* for key frames in the motion sequence, including: (1) *Pose*: PoseScript [19] is used to generate detailed narrations on part-level body postures. (2) *Contact*: Utilizing SMPL fitting model [41], we extract human joint locations and annotate those that establish contact with the 3D mesh of scene objects. (3) *Position*: We design a rule-based approach to compute object orientation and distance relative to the human, categorizing these spatial relationships into predefined classes in natural language format.

Benchmark Construction. First, we collect 3D HIS data from PROX [28] and GIMO [63], two high-quality HIS datasets covering diverse scenarios and human activities. Then, we apply our multi-faceted text annotation pipeline to generate linguistic labels, which are then fed into GPT [2] with self-crafted prompts to create multiplex question-answer (QA) pairs, forming the foundation of HIS-Bench. This process enables the construction of samples for 13 out of 16 sub-tasks. However, for focus analysis, situated analysis, and navigation tasks, existing annotations are insufficient. So we recruit human annotators to manually label these data. To ensure data quality, we manually verify each sample to preclude incorrectness or ambiguity in answers. After these procedures, we finalize HIS-Bench with 800 unique questions (each sub-task has 50 questions) covering 31 scenes and 500 motion segments, possessing diversity across scene types, motion patterns, and linguistic expressions. The statistics of HIS-Bench is presented in Fig. 2 (b).

4. HIS-GPT

Existing vision-language models [9, 11, 32] struggle to jointly model 3D human and scene modalities, limiting their effectiveness in HIS understanding. In this work, we propose HIS-GPT, a multi-modal framework designed to integrate human motion with scene context information, enabling more comprehensive HIS understanding.

4.1. Model Architecture

Overview. As shown in Fig. 4, HIS-GPT takes as input a 3D scene $\mathcal{S}$, a human motion sequence $\mathcal{M}$ and a text instruction $\mathcal{I}$. The scene is represented as a point cloud $\mathcal{S} \in \mathbb{R}^{P\times 6}$, with each point characterized by 3D coordinates and RGB values. The motion $\mathcal{M} = \{M_i\}_{i=1}^{T}$ is a sequence of T SMPL human poses. The 3D scene $\mathcal{S}$ and human motion $\mathcal{M}$ are encoded separately into latent embeddings using dedicated encoders. To enhance human-scene interactions, we introduce two key modules: the **Auxiliary Interaction (AInt)** module, which injects interaction-aware knowledge into the scene and motion embeddings, and the **Layout-Trajectory Position Encoding (LTP)** module, which encodes spatial and temporal relationships between scene and human motions. Finally, the enriched embeddings from both modalities are projected and prefixed to the

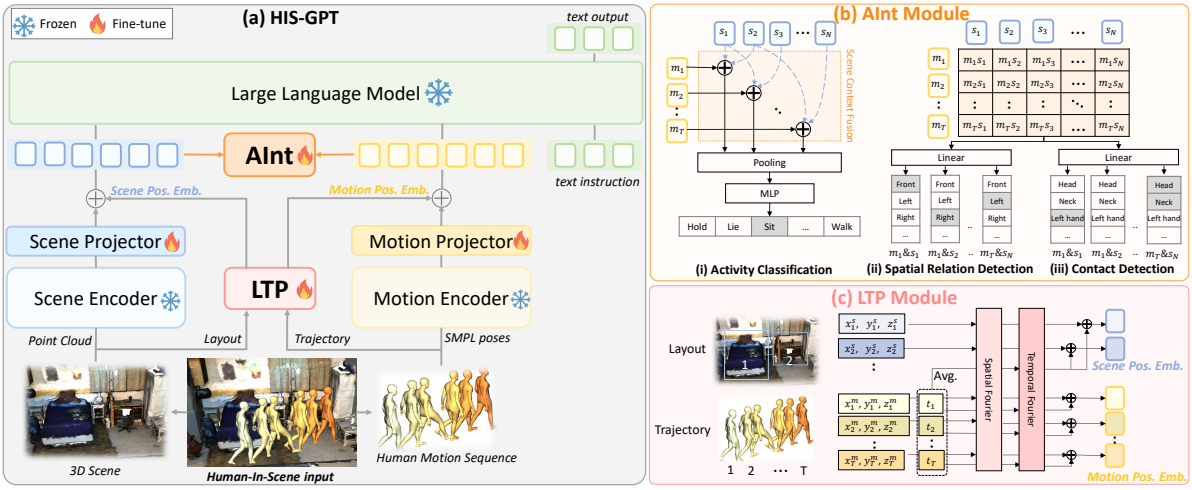


Figure 4. (a) **HIS-GPT overall architecture**. HIS-GPT uses separate pretrained encoders for scene and motion to extract embeddings, which are then combined with instructions and processed by the LLM. (b) **Auxiliary Interaction (AInt) module**: Enhance human-scene interactions through three auxillary sub-tasks. (c) **Layout-Trajectory Position Encoding (LTP) module**: Encode spatial and temporal relationships into position embeddings, injecting contextual knowledge to enhance HIS understanding.

text instruction $\mathcal{I}$, before being fed into the LLM to generate natural language answers.

Scene Encoder. Following [32], we extract object features using a pretrained 3D encoder [64], with object point clouds derived from a 3D scene segmentor [50]. The scene encoder generates a set of scene embeddings $\{s_i \in \mathbb{R}^d\}_{i=1}^N$ for 3D scene $\mathcal{S}$, where N denotes the number of detected objects and d is the latent embedding dimension.

Motion Encoder. Following [43], we adopt a motion VQ-VAE [52] as the motion encoder. The motion encoder maps human motion $\mathcal{M}$ to a set of motion embeddings $\{m_t \in \mathbb{R}^d\}_{t=1}^T$ derived from the learned motion codebook.

Auxiliary Interaction (AInt) Module. However, the scene and human motion embeddings are generated individually, lacking essential human-scene interactive cues. To address this, we propose AInt, which incorporates a set of auxiliary tasks to guide scene and motion embeddings in capturing these interactive cues, as shown in Fig. 4 (b):

(1) *Activity Classification.* As human activities involve interactions with surrounding scenes, we introduce an *activity classification* task to predict human activity within the scene. In detail, we first perform scene context fusion by integrating motion embeddings with the features of objects likely to be involved in the activity. Specifically, for the motion m_t, we identify the k nearest objects based on spatial proximity to m_t, and fuse their latent embeddings with the motion embedding:

$$\tilde{m}_t = m_t + \mathrm{Avg}\left(s_{t_1}, ..., s_{t_k}\right), \tag{1}$$

where $t_1 \sim t_k$ denotes the indices of the k nearest objects for m_t, and $\mathrm{Avg}(\cdot)$ is the averaging operation. The fused motion embedding is then passed through a multi-layer perceptron (MLP) to predict the human activity category, supervised by a cross-entropy loss:

$$\mathcal{L}_{act} = \mathrm{CE}\left(p^a, \mathrm{SM}\left(\mathrm{MLP}(\mathrm{Avg}(\tilde{m}_1, \ldots, \tilde{m}_T))\right)\right), \tag{2}$$

where p^a stands for the ground-truth activity category, SM denotes the softmax operation, and CE denotes the cross-entropy loss function.

(2) *Spatial Relation Detection.* Accurately distinguishing spatial relations between human and scene context is crucial for modeling interactive cues. To enhance this capability, we introduce a *spatial relation detection* task to classify human-scene spatial relations. Specifically, we define 8 categories (*e.g.*, 'facing') to characterize human-object spatial relations. Given the scene embedding s_i and motion embedding m_t, AInt module predicts the spatial relation between the i-th object and human motion at t-th frame, supervised by a cross-entropy loss:

$$\mathcal{L}_{spa} = \sum_{i,t} \mathrm{CE}\left(p^s_{it}, \mathrm{SM}\left(W^{spa}_s(s_i) \cdot W^{spa}_m(m_t)\right)\right), \tag{3}$$

where p^s_{it} stands for the ground-truth spatial relation label between the i-th object and t-th motion frame, W^{spa}_s and W^{spa}_m are linear projection weights.

(3) *Contact Detection.* Another crucial aspect for human-scene interactions is physical contact between human body and surrounding objects. To capture these cues, we introduce a *contact detection* task, which predicts whether an object is in contact with a specific human body part, supervised by a binary cross-entropy loss:

$$\mathcal{L}_{cont} = \sum_{i,t} \mathrm{BCE}(p^c_{it}, \sigma(W^{cont}_s(s_i) \cdot W^{cont}_m(m_t))), \tag{4}$$

where p^c_{it} represents the ground-truth contact label, with $[p^c_{it}]_l = 1$ (or 0) indicating that the i-th object is in contact

(or not) with the l-th body joint at t-th motion frame, W_s^{cont} and W_m^{cont} are projecting weights. σ denotes the sigmoid function and BCE denotes binary cross-entropy function.

Layout-trajectory Position Encoding (LTP) Module. Traditional position encoding in MLLMs primarily model sequential relationships among tokens, overlooking the complex spatiotemporal relationships between human and their surrounding environment. To this end, we propose LTP, which generates position embeddings based on spatial locations and temporal orders of human and scene input. By globally aligning spatial and temporal information across human motion and scene modalities, LTP enhances contextural awareness, enabling each modality to more effectively incorporate relevant information from the other.

As shown in Fig. 4, LTP module consists of a Spatial Fourier-transform (SF) and a Temporal Fourier-transform (TF) layer to encode 3D spatial coordinates and temporal information, respectively. Specifically, given a 3D coordinate $\mu = [x, y, z]$ and a timestamp $t \in [1, T]$, ST and TF layers are implemented as follows:

$$SF(\mu) = \text{sincos}(\phi_{SF} \cdot 2\pi\mu), TF(t) = \text{sincos}(\phi_{TF} \cdot 2\pi t), \quad (5)$$

where ϕ_{SF} and ϕ_{TF} are linear projection weights, and $\text{sincos}(\cdot)$ denotes the concatenation of sine and cosine results along latent dimension.

Leveraging SF and TF layers, for human motion modality, LTP generates a position encoding vector $e_t^m = SF(\mu_t) + TF(t)$ for the t-th motion frame, based on its 3D location $\mu_t = [x_t^m, y_t^m, z_t^m]$ and timestamp t. For 3D scene modality, LTP module yields a position encoding vector $e_i^s = SF(\mu_i) + \frac{1}{T}\sum_t TF(t)$ for the i-th object, based on its 3D location $\mu_i = [x_i^s, y_i^s, z_i^s]$. Note that we apply averaging to the temporal fourier transformations across all motion timestamps, as the object presents throughout the entire motion sequence. Finally, we aggregate the position encodings into the embeddings of each modality as: $f_i^s = s_i + e_i^s$, $f_t^m = m_t + e_t^m$. In this manner, we obtain latent features $F^s = \{f_i^s\}_{i=1}^N$ and $F^m = \{f_t^m\}_{t=1}^T$ for scene and motion modality, respectively.

LLM. After the LTP module, the latent scene feature F^s and motion feature F^m are fed into a decoder-only LLM. Given the test instruction $\mathcal{I}$ and answer $\mathcal{A}$, the LLM predicts the probability distribution of potential next answer token at each step, $P\left(\mathcal{A}_{[n]}|F^s, F^m, \mathcal{I}, \mathcal{A}_{[<n]}\right)$, in an autoregressive manner. The objective is to maximize the log-likelihood of this predicted probability distribution, denoted as $\mathcal{L}_{llm} = -\sum_n \log P\left(\mathcal{A}_{[n]}|F^s, F^m, \mathcal{I}, \mathcal{A}_{[<n]}\right)$.

4.2. Training

To effectively align the 3D scene and human modalities with the LLM, we propose a two-stage training strategy:

Stage1: Modality alignment: In this stage, we use the annotation pipeline described in Sec. 3.2 to craft detailed HIS captions for aligning input modalities with LLM. Additionally, we add scene captions and motion captions to further enhance the alignment. This stage uses the autoregressive loss of LLM along with the auxiliary tasks in AInt module for training: $\mathcal{L} = \mathcal{L}_{llm} + \lambda_{act}\mathcal{L}_{act} + \lambda_{spa}\mathcal{L}_{spa} + \lambda_{cont}\mathcal{L}_{cont}$, where λ_{act}, λ_{spa} and λ_{cont} are hyperparameters.

Stage2: HIS instruction tuning: In this stage, we synthesize a diverse instruction-following HIS data corpus, which covers a wide range of capabilities and formats for tuning. We only fine-tune HIS-GPT with $\mathcal{L}_{llm}$ to ensure the quality of instruction following.

In total, our training data comprises 60k visual captions and 700k instruction tuning samples, covering over 750 diverse scenes. More details about the training data are provided in Appendix C.2.

5. Experiments

5.1. Experimental Setup

HIS-QA Baselines. Inspired by the recent advances in vision-language models, we investigate how well these models could address the proposed HIS-QA task. **(1) 3D scene LLMs.** Current 3D scene LLMs are incapable of processing sequential human motion. To adapt these models for HIS-QA, we convert the human body from a randomly selected frame into a point cloud format, and input it alongside the scene mesh into the 3D scene LLM. We employ LL3DA [11] and Chat-Scene [32] for evaluation. **(2) Vision LLMs.** Since existing vision LLMs cannot directly process 3D input, we render HIS data into video segments and input them into vision LLMs. We select models from the GPT [33], Qwen [6], and LLaVA [36] families. **(3) LLMs w/ Frame Captions.** To leverage strong image captioners, we first derive frame-level captions from rendered HIS videos, and input these captions into a LLM to answer HIS questions. We adopt Qwen-vl-max [6] and LLaVA-OV [36] as captioners, and GPT-4 [2] as LLM. **(4) LLMs w/ Scene&Motion Captions.** To extract linguistic information from HIS data, we separately use captioners for 3D scene and 3D human motions, and feed these scene and motion captions into an LLM to perform HIS tasks. Specifically, we use LL3DA [11], AvatarGPT [65], and GPT-4 [2] as scene captioner, motion captioner, and LLM respectively. The detailed implementation of HIS-QA baselines is provided in Appendix C.3.

Implementation Details for HIS-GPT. We adopt Vicuna-1.5 [15] as LLM backbone, and AdamW [42] optimizer for training. HIS-GPT is trained in two stages: stage 1 runs for 100k steps with a learning rate of 1×10^{-4}, while stage 2 runs for $50k$ steps with a reduced learning rate of 5×10^{-5}. The batch size is set to 16 for both stages. To preserve the original capabilities of the backbones, we keep the scene encoder, motion encoder and LLM frozen throughout train-

Table 2. Quantitative evaluation results on HIS-Bench. We run the evaluation for three times and report the average score for each dimension. The full score for each dimension is 100. **'Avg.'** is the average score across all 16 dimensions. The best and second-best results are **boldfaced** and underlined, respectively.

Methods	Activity		Spatial Relationship			Human-object Interaction			Analysis		Prediction		Dialogue	Planning			Avg.
	AC	SA	HP	BO	OO	IT	IO	CP	FA	SA	IP	MP		HT	LT	NA	
3D Scene MLLMs																	
LL3DA [11]	9.0	4.0	3.5	4.7	19.0	4.0	10.5	11.7	6.5	17.2	4.2	6.3	4.7	1.0	0.3	0.0	6.7
Chat-Scene [32]	1.8	16.5	0.5	6.5	5.2	3.0	24.3	14.7	3.7	18.3	6.3	7.3	3.5	10.0	8.8	1.3	8.2
Vision LLMs																	
GPT-4v [1]	10.5	22.3	7.2	34.7	25.0	24.2	49.2	24.7	5.7	28.3	12.2	16.0	**58.7**	33.5	24.2	10.5	24.2
GPT-4o [33]	24.3	36.0	9.7	36.5	31.3	32.7	46.0	31.2	31.3	39.7	23.3	17.7	36.5	54.3	35.3	15.0	31.3
Qwen-VL-max [6]	25.3	32.0	7.7	31.8	13.2	25.0	54.7	31.7	9.0	17.8	19.3	9.7	33.0	31.5	26.2	8.7	23.5
Qwen2.5-VL [7]	10.2	11.0	5.5	27.3	18.3	16.7	49.0	29.7	2.8	20.3	12.5	16.7	15.5	21.5	20.0	7.7	17.8
LLaVA-OV [36]	15.3	7.7	9.2	16.0	14.3	16.7	41.3	27.7	1.0	14.5	9.5	7.5	16.7	17.8	8.2	4.0	14.2
LLaVA-Video [60]	11.3	16.2	4.0	20.8	9.0	17.8	27.8	29.0	13.8	21.5	12.5	14.0	20.8	19.7	16.0	6.2	16.3
LLMs w/ Frame Captions																	
LLaVA-OV [36]+GPT-4 [2]	9.0	10.3	5.5	22.3	16.0	14.7	29.3	18.0	2.7	21.2	27.5	13.0	53.5	22.7	15.3	5.5	17.9
Qwen-VL-max [6]+GPT-4 [2]	5.3	6.0	3.2	8.3	10.0	3.5	29.7	13.0	0.6	6.0	14.5	5.3	22.0	6.5	1.7	4.8	8.8
LLMs w/ Scene&Motion Captions																	
LL3DA [11]+AvatarGPT [65]+GPT-4 [2]	1.3	0.5	2.5	5.7	0.3	2.5	21.5	12.8	0.0	3.7	6.0	2.7	13.3	3.3	2.7	1.0	5.0
HIS Foundation Models **(Ours)**																	
HIS-GPT	**39.3**	**49.8**	**37.0**	**57.3**	**32.0**	**52.8**	**58.3**	**55.5**	**33.8**	**48.2**	**50.5**	**50.0**	53.2	**55.7**	**58.0**	**48.0**	**48.7**

Table 3. Ablations on the key components of HIS-GPT. 'act', 'spa' and 'cont' denotes the activity classification, spatial relation detection and human-scene contact detection task in AInt module. 'PE' denotes position encoding methods.

Methods	AInt			PE	HIS-Bench			
	act	spa	cont		Act.	Spa.	HoI.	Avg.
1				sine	41.8	34.7	45.8	43.0
2	✓	✓	✓	sine	43.5	35.3	51.0	44.1
3				LTP	43.5	38.8	50.3	46.0
4	✓			LTP	**44.8**	36.5	47.5	45.3
5		✓		LTP	42.4	39.7	48.8	47.3
6			✓	LTP	43.3	38.5	52.0	46.9
7(Ours)	✓	✓	✓	LTP	44.6	**42.1**	**55.5**	**48.7**

Table 4. Ablations on the training strategy of HIS-GPT. 'HIS', 'Scene' and 'Motion' denotes the usage of HIS, scene and motion data in stage 1 training.

Stage 1			Stage 2	HIS-Bench							
HIS	Scene	Motion		Act.	Spa.	HoI.	Ana.	Pre.	Dia.	Pla.	Avg.
✓	✓	✓		39.3	30.2	41.0	32.8	40.5	35.5	41.8	37.5
			✓	39.0	31.3	47.8	37.0	46.0	47.5	50.8	42.6
✓			✓	42.2	36.5	51.8	37.8	47.0	41.5	52.3	45.8
✓	✓		✓	39.0	39.7	46.9	38.8	46.0	42.5	51.2	44.0
✓		✓	✓	39.5	**42.8**	49.7	**41.0**	46.0	48.5	52.5	46.0
✓	✓	✓	✓	**44.6**	42.1	**55.5**	**41.0**	**50.3**	**53.2**	**53.9**	**48.7**

ing, fine-tuning only the projection layers, AInt and LTP modules. The loss weights λ_{act}, λ_{spa} and λ_{cont} are set to 0.5, 0.5 and 0.1, determined by grid search.

Evaluation Metrics of HIS-Bench. Considering that HIS-Bench consists of open-ended questions, we use GPT-4 as an automatic evaluator to assess answer correctness. Following [14], we prompt GPT-4 to assign a score between 0 and 2 for each answer. Since each task in HIS-Bench consists of 50 questions, the full score for each task is 100.

5.2. Quantitative Results

Tab. 2 provides the quantitative results on HIS-Bench. Based on the results, we summarize our findings as follows:

Question types. From Tab. 2, we observe that almost all models perform relatively well on dialogue (SD) and task-planning (HT, LT), likely because these dimensions are closely aligned with the conversation and reasoning abilities inherent in original LLMs. In contrast, tasks requiring a strong understanding of spatial characteristics, such as Human Position (HP) and Navigation (NA), present significant challenges, highlighting the need for the development of more advanced spatial interaction modeling capabilities.

HIS-QA baselines results. Among all the baselines, vision LLMs demonstrate significantly superior performance compared to other model families. This can be attributed to their outstanding instruction-following abilities which enable them to effectively follow HIS-QA questions, as well as their robust visual generalization to rendered videos. While 3D scene LLMs excel in 3D understanding, their ability to generalize to 3D human input is limited due to the absence of human-centric data in their training corpus. Similarly, LLMs w/ captions are restricted by the lack of detailed spatial information and human-scene interactions within their generated captions, leading to a weaker understanding of complex human-in-scene behaviors.

HIS-GPT results. As evident in Tab. 2, HIS-GPT significantly outperforms all HIS-QA baselines, achieving an average score exceeding the highest-performing baseline, GPT-4o, by 17.4 points. Compared to other vision-language models, HIS-GPT demonstrates particular strength in tasks requiring a nuanced understanding of spatial relations between humans and their 3D surroundings, such as Human

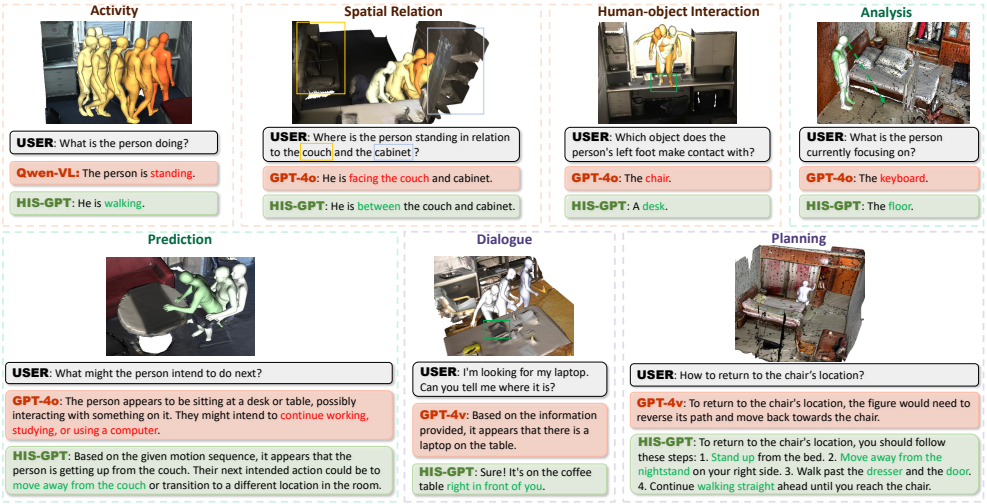


Figure 5. Qualitative comparisons of HIS-GPT and other baselines on HIS-QA. Red/green color denotes wrong/correct outputs.

Position (HP) and Contact Part (CP). Also, HIS-GPT performs well in prediction tasks, showcasing its ability to accurately infer human states and perform complex reasoning.

5.3. Ablation Studies

We conduct ablation studies to validate the effectiveness of HIS-GPT. Additional ablations, including loss weight and LLM tuning strategy, are provided in Appendix D.

Ablations on AInt module. Tab. 3 reports the ablation studies on the AInt module. The results indicate that integrating AInt increases the average score on HIS-Bench by 1.1, demonstrating its effectiveness for human-in-scene understanding. To further analyze its impact, we break down the contributions of individual sub-tasks within the AInt module. As shown in Tab. 3, activity classification (act), spatial relation detection (spa), and contact detection (cont) tasks improve their corresponding HIS-Bench core tasks (Activity, Spatial Relationship, and HoI) by 1.3, 0.9, and 1.7, respectively. These results indicate that explicitly modeling fine-grained human-scene interactions through AInt substantially benefits the overall capabilities of HIS-GPT.

Ablations on LTP module. As shown in Tab. 3, integrating the LTP module leads to a 3.0 average score gain on HIS-Bench, demonstrating its effectiveness. Furthermore, when AInt and LTP are used jointly, they achieve a significant 5.7 performance gain over the baseline. This result highlights the complementary nature of these modules, suggesting that combining fine-grained human-scene interaction modeling with structured spatial-temporal encoding can further enhance the model's ability to comprehensively understand human activities in 3D environments.

Ablations on Training Strategy. Tab. 4 presents ablation study on the two-stage training strategy. The results indicate that both modality alignment (Stage 1) and instruction tuning (Stage 2) are essential for effectively training HIS-GPT. Additionally, incorporating scene and motion caption data in Stage 1 leads to a rise of 2.9 in average score, validating their effectiveness in facilitating modality alignment.

5.4. Qualitative Results

Fig. 5 presents qualitative examples of HIS-GPT across various HIS-QA tasks. Compared to baseline models, HIS-GPT gives more accurate answers in basic perceptions about human activities, spatial relation to scene, and interaction with objects. Moreover, HIS-GPT generates highly plausible responses in reasoning and prediction tasks, showcasing a strong understanding of human behavior within the scene. Additionally, HIS-GPT excels in dialogue and planning tasks, which are crucial for embodied AI applications. Notably, while GPT-4v frequently produces generic or uninformative responses that are not helpful enough for users to address their problems, HIS-GPT provides constructive replies with situated knowledge (*e.g.*, 'right in front of you') and detailed guidance (*e.g.*, 'stand up', 'walk straight') that can effectively assist users in real-world scenarios. More qualitative results are provided in Appendix E.

6. Conclusion

In this paper, we introduce HIS-QA, a new task formulation for Human-In-Scene understanding. To evaluate this task, we raise HIS-Bench, the first multimodal benchmark tailored for HIS-QA, featuring diverse questions that span basic perception, complex reasoning and embodied applications. Additionally, we propose HIS-GPT, a foundation model that jointly perceives 3D human and scene inputs, effectively addressing HIS-QA tasks in a unified framework. We believe the benchmark and model could benefit future research and applications in human-centric understanding.

Acknowledgements. This work is partially supported by National Natural Science Foundation of China U2336213, 62376259, 62306301.

References

[1] Gpt-4v(ision) system card. 2023. 7

[2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 4, 6, 7

[3] 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 *ECCV*, 2020. 2

[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 *CVPR*, 2023. 2

[5] Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. Scanqa: 3d question answering for spatial scene understanding. In *CVPR*, 2022. 2, 3

[6] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023. 6, 7

[7] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 2, 7

[8] Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. In *ECCV*, 2020. 3

[9] Ling-Hao Chen, Shunlin Lu, Ailing Zeng, Hao Zhang, Benyou Wang, Ruimao Zhang, and Lei Zhang. Motionllm: Understanding human behaviors from human motions and videos. *arXiv preprint arXiv:2405.20340*, 2024. 3, 4

[10] Sijin Chen, Hongyuan Zhu, Xin Chen, Yinjie Lei, Gang Yu, and Tao Chen. End-to-end 3d dense captioning with vote2cap-detr. In *CVPR*, 2023. 2

[11] Sijin Chen, Xin Chen, Chi Zhang, Mingsheng Li, Gang Yu, Hao Fei, Hongyuan Zhu, Jiayuan Fan, and Tao Chen. Ll3da: Visual interactive instruction tuning for omni-3d understanding reasoning and planning. In *CVPR*, 2024. 2, 4, 6, 7

[12] Yilun Chen, Shuai Yang, Haifeng Huang, Tai Wang, Runsen Xu, Ruiyuan Lyu, Dahua Lin, and Jiangmiao Pang. Grounded 3d-llm with referent tokens. *arXiv preprint arXiv:2405.10370*, 2024. 1

[13] Zhenyu Chen, Ali Gholami, Matthias Nießner, and Angel X Chang. Scan2cap: Context-aware dense captioning in rgb-d scans. In *CVPR*, 2021. 2

[14] Sijie Cheng, Zhicheng Guo, Jingwen Wu, Kechen Fang, Peng Li, Huaping Liu, and Yang Liu. Egothink: Evaluating first-person perspective thinking capability of vision-language models. In *CVPR*, 2024. 7

[15] Wei-Lin Chiang, Zhuohan Li, Ziqing Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. 2023. 1, 2, 6

[16] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *JMLR*, 2023.

[17] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *JMLR*, 2024. 1

[18] Peishan Cong, Ziyi Wang, Zhiyang Dou, Yiming Ren, Wei Yin, Kai Cheng, Yujing Sun, Xiaoxiao Long, Xinge Zhu, and Yuexin Ma. Laserhuman: language-guided scene-aware human motion generation in free environment. *arXiv preprint arXiv:2403.13307*, 2024. 2

[19] Ginger Delmas, Philippe Weinzaepfel, Thomas Lucas, Francesc Moreno-Noguer, and Grégory Rogez. Posescript: 3d human poses from natural language. In *ECCV*, 2022. 3, 4

[20] Ginger Delmas, Philippe Weinzaepfel, Francesc Moreno-Noguer, and Grégory Rogez. Posefix: correcting 3d human poses with natural language. In *ICCV*, 2023. 3

[21] 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. 1

[22] Mark Endo, Joy Hsu, Jiaman Li, and Jiajun Wu. Motion question answering via modular motion programs. In *ICML*, 2023. 3

[23] Yao Feng, Jing Lin, Sai Kumar Dwivedi, Yu Sun, Priyanka Patel, and Michael J Black. Chatpose: Chatting about 3d human pose. In *CVPR*, 2024. 3

[24] Rao Fu, Jingyu Liu, Xilun Chen, Yixin Nie, and Wenhan Xiong. Scene-llm: Extending language model for 3d visual understanding and reasoning. *arXiv preprint arXiv:2403.11401*, 2024. 1, 2, 4

[25] Qiao Gu, Ali Kuwajerwala, Sacha Morin, Krishna Murthy Jatavallabhula, Bipasha Sen, Aditya Agarwal, Corban Rivera, William Paul, Kirsty Ellis, Rama Chellappa, et al. Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning. In *(ICRA)*, 2024. 1

[26] 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 *CVPR*, 2022. 3

[27] 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 *ECCV*, 2022. 3

[28] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J Black. Resolving 3d human pose ambiguities with 3d scene constraints. In *ICCV*, 2019. 2, 4

[29] Mohamed Hassan, Duygu Ceylan, Ruben Villegas, Jun Saito, Jimei Yang, Yi Zhou, and Michael J Black. Stochastic scene-aware motion prediction. In *ICCV*, 2021. 2

[30] Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 3d-llm: Injecting the 3d world into large language models. In *NeurIPS*, 2023. 1, 2

[31] Yining Hong, Zishuo Zheng, Peihao Chen, Yian Wang, Junyan Li, and Chuang Gan. Multiply: A multisensory object-centric embodied large language model in 3d world. In *CVPR*, 2024.

[32] Haifeng Huang, Yilun Chen, Zehan Wang, Rongjie Huang, Runsen Xu, Tai Wang, Luping Liu, Xize Cheng, Yang Zhao, Jiangmiao Pang, et al. Chat-scene: Bridging 3d scene and large language models with object identifiers. In *NeurIPS*, 2024. 1, 2, 4, 5, 6, 7

[33] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 2, 6, 7

[34] Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. In *NeurIPS*, 2023. 1, 3

[35] 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 *CVPR*, 2024. 2, 3

[36] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *TMLR*, 2024. 2, 4, 6, 7

[37] Chuqiao Li, Julian Chibane, Yannan He, Naama Pearl, Andreas Geiger, and Gerard Pons-Moll. Unimotion: Unifying 3d human motion synthesis and understanding. *arXiv preprint arXiv:2409.15904*, 2024. 3

[38] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023. 4

[39] Yiheng Li, Ruibing Hou, Hong Chang, Shiguang Shan, and Xilin Chen. Unipose: A unified multimodal framework for human pose comprehension, generation and editing. *arXiv preprint arXiv:2411.16781*, 2024. 1, 3

[40] Jing Lin, Ailing Zeng, Shunlin Lu, Yuanhao Cai, Ruimao Zhang, Haoqian Wang, and Lei Zhang. Motion-x: A large-scale 3d expressive whole-body human motion dataset. In *NeurIPS*, 2023. 3, 4

[41] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: a skinned multi-person linear model. *TOG*, 2015. 3, 4

[42] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 6

[43] Mingshuang Luo, RuiBing Hou, Zhuo Li, Hong Chang, Zimo Liu, Yaowei Wang, and Shiguang Shan. M3gpt: An advanced multimodal, multitask framework for motion comprehension and generation. In *NeurIPS*, 2024. 1, 2, 3, 5

[44] Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. Sqa3d: Situated question answering in 3d scenes. In *ICLR*, 2022. 2, 3

[45] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *ICCV*, 2019. 4

[46] Arjun Majumdar, Anurag Ajay, Xiaohan Zhang, Pranav Putta, Sriram Yenamandra, Mikael Henaff, Sneha Silwal, Paul Mcvay, Oleksandr Maksymets, Sergio Arnaud, et al. Openeqa: Embodied question answering in the era of foundation models. In *CVPR*, 2024. 3

[47] Maria Parelli, Alexandros Delitzas, Nikolas Hars, Georgios Vlassis, Sotirios Anagnostidis, Gregor Bachmann, and Thomas Hofmann. Clip-guided vision-language pre-training for question answering in 3d scenes. In *CVPR*, 2023. 2

[48] Zekun Qi, Runpei Dong, Shaochen Zhang, Haoran Geng, Chunrui Han, Zheng Ge, Li Yi, and Kaisheng Ma. Shapellm: Universal 3d object understanding for embodied interaction. In *ECCV*, pages 214–238, 2024. 1

[49] Zhangyang Qi, Ye Fang, Zeyi Sun, Xiaoyang Wu, Tong Wu, Jiaqi Wang, Dahua Lin, and Hengshuang Zhao. Gpt4point: A unified framework for point-language understanding and generation. In *CVPR*, 2024. 1

[50] Jonas Schult, Francis Engelmann, Alexander Hermans, Or Litany, Siyu Tang, and Bastian Leibe. Mask3d: Mask transformer for 3d semantic instance segmentation. In *ICRA*, 2023. 4, 5

[51] Chunyi Sun, Junlin Han, Weijian Deng, Xinlong Wang, Zishan Qin, and Stephen Gould. 3d-gpt: Procedural 3d modeling with large language models. *arXiv preprint arXiv:2310.12945*, 2023. 1

[52] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. In *NeurIPS*, 2017. 5

[53] Tai Wang, Xiaohan Mao, Chenming Zhu, Runsen Xu, Ruiyuan Lyu, Peisen Li, Xiao Chen, Wenwei Zhang, Kai Chen, Tianfan Xue, et al. Embodiedscan: A holistic multi-modal 3d perception suite towards embodied ai. In *CVPR*, 2024. 2

[54] Jianing Yang, Xuweiyi Chen, Nikhil Madaan, Madhavan Iyengar, Shengyi Qian, David F Fouhey, and Joyce Chai. 3d-grand: A million-scale dataset for 3d-llms with better grounding and less hallucination. *arXiv preprint arXiv:2406.05132*, 2024. 2, 4

[55] Jihan Yang, Runyu Ding, Weipeng Deng, Zhe Wang, and Xiaojuan Qi. Regionplc: Regional point-language contrastive learning for open-world 3d scene understanding. In *CVPR*, 2024. 1, 2

[56] Fukun Yin, Xin Chen, Chi Zhang, Biao Jiang, Zibo Zhao, Wen Liu, Gang Yu, and Tao Chen. Shapegpt: 3d shape generation with a unified multi-modal language model. *IEEE Transactions on Multimedia*, 2025. 1

[57] Mingyuan Zhang, Daisheng Jin, Chenyang Gu, Fangzhou Hong, Zhongang Cai, Jingfang Huang, Chongzhi Zhang, Xinying Guo, Lei Yang, Ying He, et al. Large motion model for unified multi-modal motion generation. In *ECCV*, 2024. 3

[58] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022. 1

[59] Yiming Zhang, ZeMing Gong, and Angel X Chang. Multi3drefer: Grounding text description to multiple 3d objects. In *ICCV*, 2023. 2

[60] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024. 7

[61] Yue Zhang, Zhiyang Xu, Ying Shen, Parisa Kordjamshidi, and Lifu Huang. Spartun3d: Situated spatial understanding of 3d world in large language models. *arXiv preprint arXiv:2410.03878*, 2024. 2

[62] Haoyu Zhen, Xiaowen Qiu, Peihao Chen, Jincheng Yang, Xin Yan, Yilun Du, Yining Hong, and Chuang Gan. 3d-vla: A 3d vision-language-action generative world model. 2024. 2

[63] Yang Zheng, Yanchao Yang, Kaichun Mo, Jiaman Li, Tao Yu, Yebin Liu, C Karen Liu, and Leonidas J Guibas. Gimo: Gaze-informed human motion prediction in context. In *ECCV*, 2022. 2, 4

[64] Junsheng Zhou, Jinsheng Wang, Baorui Ma, Yu-Shen Liu, Tiejun Huang, and Xinlong Wang. Uni3d: Exploring unified 3d representation at scale. In *ICLR*, 2024. 2, 5

[65] Zixiang Zhou, Yu Wan, and Baoyuan Wang. Avatargpt: All-in-one framework for motion understanding planning generation and beyond. In *CVPR*, 2024. 1, 3, 6, 7