

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

Supplementary Material

In Appendix A, we provide a comprehensive overview of the implementation details of HIS-Bench. Appendix C elaborates on the experimental setup, including the HIS-QA baselines and our proposed HIS-GPT model. In Appendix B, we present additional data samples of constructed HIS-Bench. Appendix D and E provide supplementary quantitative and qualitative results, offering deeper insights into the performance and limitations of HIS-GPT.

A. HIS-Bench Details

In this part, we present additional details of HIS-Bench, including a detailed description of the multi-faceted text annotation process (Appendix A.1), the prompts used for question-answer generation (Appendix A.2), the web interface designed for human annotation (Appendix A.3), user studies on benchmark quality (Appendix A.4), and studies on the accuracy of using GPT as automatic evaluator (Appendix A.5).

A.1. Details on Multi-Faceted Text Annotation

In Sec.3.2, we present the multi-faceted text annotation process employed to generate rich linguistic labels for HIS datasets. In this part, we further elaborate on the implementation details, including scene annotation and frame-level contact/position annotation.

Scene Annotation. Following SceneVerse [8], we first utilize a 3D scene segmentor [12] to obtain 6D bounding boxes and semantic labels of each object in the scene. Next, we construct a scene graph, where nodes represent objects, and edges capture the spatial relationships between two objects. From this scene graph, we extract triplets consisting of two objects and their relationship such as (sofa, near, chair). Finally, we apply predefined natural language templates to transform these triplets into referring expressions for each object, e.g., ‘The sofa is near the chair’. For details on the definitions of spatial relationships and the templates, please refer to [8].

Frame-level Contact Annotation. For the annotation of frame-level human contact with the scene, we represent the annotations as tuples (*body joint, anchor object*), indicating that a specific body joint is in contact with an anchor object in the scene. In details, we define 22 body joints based on the 3D human representations following [13]. To determine contact, we compute the closest distance between each body joint and the point cloud of each object, labeling them as in contact if the distance is below a threshold ϵ , which we set to 0.1 in practice. The full list of body joints is provided below:

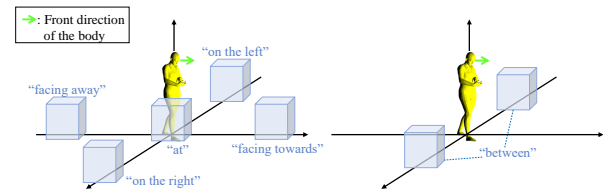


Figure S1. Illustrations of the definition of human-object orientations. We define six types of orientations: ‘facing towards’, ‘on the left’, ‘on the right’, ‘facing away’, ‘at’, and ‘between’.

pelvis, left hip, right hip, lower spine, left knee, right knee, middle spine, left ankle, right ankle, upper spine, left foot, right foot, neck, left collar, right collar, head, left shoulder, right shoulder, left elbow, right elbow, left wrist, right wrist.

Frame-level Position Annotation. For frame-level position annotation within the scene, we represent the annotation as triplets (*orientation, distance, anchor object*). Each triplet captures the relative orientation and distance of an anchor object with respect to the human body, providing a structured representation of the human’s spatial context. Specifically, *orientation* is defined as the relative direction of the anchor object with respect to the human. We categorize orientation into six types: ‘facing towards’, ‘on the left’, ‘on the right’, ‘facing away’, ‘at’ and ‘between’, based on the angle between the object’s direction and the human’s facing direction, as illustrated in Fig. S1. Notably, the ‘between’ category describes a scenario where one object is positioned to the left and another object to the right of the person. *Distance* is defined as the horizontal distance (on xy-plane) between the center point of the anchor object and the pelvis point of human body.

A.2. Prompts for Question-Answer Generation

We formulate a large proportion of HIS-Bench questions by prompting GPT-4 [1]. In Fig. S2 and Fig. S3, we present all prompts designed for the GPT-assisted generation process. Specifically, we create a general template for generating all questions, and then fill it with task-specific instructions when generating questions for each sub-task.

A.3. Interface for Human Annotation

For three of the 16 sub-tasks in HIS-Bench, namely *focus analysis*, *situated analysis* and *navigation*, we employ human annotators to write question-answer pairs based on the raw HIS data. To facilitate this process, we design a web interface for annotation. On the interface, annotators are

Overall template for question generation:

You are an AI assistant that can understand motion of human in a 3D scene. You will be given the following information of a human motion in a 3D scene. The information contains several fields, and is listed as follows:

Scene annotations:

```
{
  "object information": Information of all the objects involved in the motion process of the human. Each object is indexed by a unique id. The information of each object is organized in the form of [<object id>: [{"category": the general category name of the object, "referral": a list of references describing the object}]].
}
```

Motion annotations:

```
{
  "action": an overall description of the activity that the person is engaging with during the whole motion sequence,
  "key moments": a list of multi-perspective descriptions on several key moments evenly sampled from the whole motion, in time order. Descriptions of each key moment includes following aspects:
  {
    "pose": a detailed description on the part-level body pose of a person at this moment.
    "human-scene contacts": a list that specifies all the human body parts that are in contact with the objects in scene at this moment. Each item in the list consists of [<body part name>, <object id>].
    "human-scene spatial relations": a list that specifies all the spatial relations between the person and the objects in scene at this moment. Each item in the list consists of [<distance>, <orientation>, <object id>].
  }
}
```

{TASK-SPECIFIC PROMPT}

Formulate question and answer as if you are directly perceiving a natural scene and human motion, which means that you should not mention that your answer is coming from the scene and motion annotations, such as index of objects, 'key moments', or 'spatial relations'! Try to use diverse sentence patterns in the question and answer.

Figure S2. The overall prompt template for HIS-Bench question-answer generation. The '{TASK-SPECIFIC PROMPT}' placeholder is filled with prompts tailored to specific HIS-QA tasks.

Table S1. User study on the quality of HIS-Bench. We show the average score of all samples in each dimension, graded by three annotators separately.

Benchmark	Answerability	Clarity	Correctness	Difficulty
HIS-Bench	4.74	4.72	4.35	3.14
OpenEQA [11]	4.72	4.80	4.64	3.23

shown the rendered video of the HIS data sequence, and are required to input the question, answer, start timestamp, and end timestamp to create a data sample. The interface also provides detailed instructions on how to annotate, along with examples of good and bad annotations.

A.4. User Study of Benchmark Quality

Considering that the majority of QA pairs in HIS-Bench are generated in an automatic manner, we carry out a user study to verify the quality of these QA pairs. Specifically, we employ three human annotators and let them grade every QA sample in HIS-Bench from the following four aspects: question answerability, question clarity, answer correctness and question difficulty. For each aspect, annotators are instructed to give an integer score from 1 to 5 (1 is lowest, and 5 is highest). Then, for a more objective reference, we additionally ask the annotators to grade 100 QA sam-

ples from another human-annotated 3D scene benchmark, OpenEQA [11], and compare the average grades between OpenEQA and HIS-Bench. As shown in Tab. S1, in all evaluated aspects, HIS-Bench achieves similar scores with the human-annotated and verified OpenEQA benchmark, proving the quality of the automatically synchronized QA pairs in HIS-Bench.

A.5. Accuracy of GPT Evaluation

As HIS-Bench consists of open-ended questions, we deploy GPT-4 [1] as the evaluation tool for the answers. Here we conduct a detailed analysis of the accuracy of GPT evaluations. We derive the model-generated (here we use GPT-4o [7]) answers on HIS-Bench, comparing the consistency of GPT evaluations and human judgements on these answers. Three human evaluators are invited to grade the answers from aspects including accuracy, information completeness, logical soundness and grammar correctness, by giving each dimension a score from 1 to 5. Then, a human-evaluated score for each answer is conducted by averaging the scores of each dimension. Using the scores, following [4], we calculate the Pearson correlation score ($\in [-1, 1]$, > 0 means positive correlation) between GPT and human evaluations. Results show that the Pearson correlation is 0.54, reflecting that GPT evaluations are accurate since they are highly consistent with humans.

Single Activity

Your job is to use all these information of the human and the scene to generate a question-answer pair asking about the human's activity.

Sequential Activity

Your job is to use all these information of the human and the scene to generate a question-answer pair asking about the human's activity before or after another activity.

Human Position

Your job is to use all these information of the human and the scene to generate a question-answer pair asking about the position of the human. Be aware that since the human is moving, the position of the person might change during his motion.

Body Orientation

Your job is to use all these information of the human and the scene to generate a question-answer pair asking about the orientation of a certain object in relation to the human. Be aware that orientation can change in the whole motion process.

Object Orientation

Your job is to use all these information of the human and the scene to generate a question-answer pair asking about the object that is at a certain orientation of the human. Be aware that since the human is moving, there might be multiple objects at the certain orientation during his motion.

Interaction Type

Your job is to use all these information of the human and the scene to generate a question-answer pair asking about the type of interaction between the human and a certain object in the scene.

Interacting Object

Your job is to use all these information of the human and the scene to generate a question-answer pair asking about the object the human (or a specified body part) is in contact with.

Contact Part

Your job is to use all these information of the human and the scene to generate a question-answer pair asking about the body part of the human that is in contact with an object.

Intent Prediction

Your job is to use all these information to generate a question-answer pair asking about the future intent of the human. In this setting, assume that the answerer can only observe the person's earlier activities, and give the answer based on the later activities, which are regarded as future events. Do not include any explanations for your answer.

Movement Prediction

Your job is to use all these information to generate a question-answer pair asking about the future movements of the human. In this setting, assume that the answerer can only observe the person's earlier activities, and give the answer based on the later activities, which are regarded as future events. Focus on the location, and do not answer information other than movements, such as action or pose.

Situated Dialogue

Your job is to use all these information to generate a meaningful conversation about the human motion in the scene. The conversation is between the human which is presented in the scene, and an intelligent agent which has the full knowledge of the scene. The dialogue should be the human inquiring for certain assistance related to his/her current activities, and the agent providing helpful and accurate answers. You should try to make the dialogue relevant to the human's position, orientation and pose status, and avoid generating questions that could be answered without perceiving the human's current status in the scene.

High-level Planning

Your job is to imagine yourself as an intelligent agent, use all these information as observations of a human activity, and formulate a question-answer pair asking about what you can do to help the human. Your answer should generally state a task you could perform to help the person. You should try to make the task relevant to the human's position, orientation and pose status, and avoid generating plans that can be conducted without knowing the status of the human in the scene.

Low-level Planning

Your job is to imagine yourself as an intelligent agent, use all these information as observations of a human activity, and formulate a question-answer pair asking about what you can do to help the human. Your answer should give a step-wise decomposition on the action steps to perform a plan to help the human. You should try to make the task relevant to the human's position, orientation and pose status, and avoid generating plans that can be conducted without knowing the status of the human in the scene.

Figure S3. The task-specific prompts for each HIS-QA task.

Table S2. The consistencies of HIS-Bench average scores on multiple baseline models and HIS-GPT, by using GPT-4 [10] and Qwen2.5-7B [3] as evaluators, respectively.

Evaluator	Chat-Scene	GPT-4o	LLaVAOV+GPT4	LL3DA+AvatarGPT+GPT4	HIS-GPT
GPT-4	8.2	31.3	17.9	5.0	48.7
Qwen2.5	7.4	29.6	16.9	4.4	46.1

Moreover, we verify the reproducibility of using LLM as evaluators by adopting an open-sourced LLM, Qwen2.5-7B [3] as the evaluator. We find that the Pearson correlation score between Qwen2.5-7B and GPT-4 is 0.75, and their judgement scores on the answers of multiple models are consistent, as shown in Tab. S2, the evaluated performance scores between Qwen2.5-7B and GPT-4 are very close on multiple models, demonstrating the general adaptability of our evaluation methods on multiple LLMs. This makes the evaluation process of HIS-Bench more easily accessible and reproducible.

B. More HIS-Bench Data Samples

In Fig. S8-S10, we present more data samples of HIS-Bench. Fig. S8 showcases samples under the core ability of basic perception, covering a total of 8 sub-tasks. Fig. S9 shows examples under the core ability of complex reasoning, covering 4 sub-tasks. Fig. S10 shows examples under the core ability of embodied applications, which contains 4 sub-tasks.

C. Experimental Details

In this part, we elaborate on the experimental details, including the instruction templates (Appendix C.1) and training data (Appendix C.2) for HIS-GPT, the implementation details for HIS-QA baseline models (Appendix C.3), and the prompts designed for the GPT-assisted evaluation of HIS-Bench (Appendix C.4).

Table S3. Statistics of training data. We generate captions and QA pairs with both scene and motion input, serving as training corpus for HIS tasks. We also collect existing scene and motion caption data to facilitate modality alignment in stage 1.

Stage	Datasets	scene	motion	type	#pairs
Stage 1	HUMANISE [13]	✓	✓	Caption	24k
	TRUMANS [9]	✓	✓	Caption	9.4k
	SceneVerse [8]	✓	✗	Caption	1.5k
	HumanML3D [5]	✗	✓	Caption	21k
Stage 2	HUMANISE [13]	✓	✓	QA	491k
	TRUMANS [9]	✓	✓	QA	209k

C.1. Instruction Templates

Our training tasks include various input visual modalities: scene only, human motion only, and human-scene inputs. For each set of input modalities, we design a specific instruction template as follows:

- Scene only: “Examine the indoor scene. Object information in the scene: [REPLACE].”
- Motion only: “Examine the human motion sequence. Motion sequence: [REPLACE].”
- Scene and Motion: Stage1: “Examine the indoor scene and a human motion sequence in the scene. Object information in scene: [REPLACE]. Motion sequence in scene:[REPLACE].” Stage2: “The conversation centers around an indoor scene and a human motion sequence. Object information in scene: [REPLACE]. Motion sequence: [REPLACE]. Based on the provided information, give an accurate answer to the following question raised by the user:” .

Here [REPLACE] is replaced by scene/motion embeddings before feeding into LLM.

C.2. Training Data

In Tab. S3, we provide a comprehensive list of the data used during the training of HIS-GPT. Specifically, for stage 1 of Modality Alignment, we incorporate a total of 33.4k HIS caption data from HUMANISE [13] and TRUMANS [9], which are generated using our text annotation pipeline. Additionally, to facilitate the alignment of each modality in HIS data, we use 1.5k scene captions from SceneVerse [8] and 21k human motion captions from HumanML3D [5], respectively. For stage 2 of Instruction Tuning, we synthesize 700k diverse QA data using HUMANISE and TRUMANS datasets.

C.3. Detailed Implementation of HIS-QA Baselines

We have developed a set of baseline models for HIS-QA, as shown in Fig. S4. In this part, we provide the implementation details of these baselines.

3D Scene LLMs. Current 3D scene LLMs are not capable of receiving sequential human motion input in SMPL for-

Table S4. The average scores of zero-shot and fine-tuned baseline methods on HIS-Bench. The fine-tuning data is consistent with HIS-GPT training data, as listed in Tab. S3. HIS-GPT does not have zero-shot version since it is trained with HIS data from scratch.

Eval Model	Chat-Scene	LLaVAOV	LLaVAOV+GPT4	HIS-GPT
Zero-shot	8.2	14.2	17.9	-
Fine-tuned	13.4	19.7	20.8	48.7

mat. Therefore, as shown in Fig. S4(a), we select one frame m_i from the motion sequence \mathcal{M} , densely sample a set of vertices \mathcal{P}_i from the SMPL-fitted 3D human mesh, and feed these vertices into the 3D scene LLM together with the scene mesh to obtain an answer, *i.e.* $\hat{A} = f_{3D}([\mathcal{S}, \mathcal{P}_i], \mathcal{Q})$.

Vision LLMs. Existing vision LLMs excel at understanding image and video input, but cannot directly process 3D input. Therefore, as shown in Fig. S4(b), we render the 3D HIS data into video segments $\mathcal{V} = \{v_i\}_{i=1}^T$, and leverage image or video LLMs to perform HIS-QA task, *i.e.* $\hat{A} = f_{VLM}(\mathcal{V}, \mathcal{Q})$.

LLMs w/ Frame Captions. As shown in Fig. S4(c), by leveraging powerful image captioners, we first generate frame-level captions $\mathcal{C} = \{c_i\}_{i=1}^T$ from T frames evenly sampled from \mathcal{V} . Then we use an LLM to answer the HIS-QA question, conditioned on the frame captions, *i.e.* $\hat{A} = \text{LLM}(\mathcal{C}, \mathcal{Q})$.

LLMs w/ Scene&Motion Captions. As shown in Fig. S4(d), we also try to derive linguistic information of HIS data by separately using scene and motion captioners to generate captions for 3D scene and 3D human motion sequence. Then we feed the scene caption c_s and motion caption c_m into an LLM, obtaining the answer $\hat{A} = \text{LLM}(c_s, c_m, \mathcal{Q})$.

C.4. Prompt for GPT-assisted Evaluation

In Fig. S6, we show the prompts used for the GPT-assisted evaluation of HIS-Bench. The prompt template follows [4].

D. More Quantitative Results

D.1. Results of Fine-tuned Baselines

Since the majority of our compared baselines could not directly receive the input format of HIS-GPT training data, we test their results on HIS-Bench in a zero-shot manner. To make a fairer comparison, we convert the training corpus of HIS-GPT into compatible formats for the baseline methods, and fine-tune these baselines with the training data. We report the results of these fine-tuned models on HIS-Bench. As shown in Tab. S4, although being fine-tuned, the performance of baseline methods still largely lag behind HIS-GPT, showing the necessity of raising the HIS-GPT framework to process 3D human and scene modalities in a unified

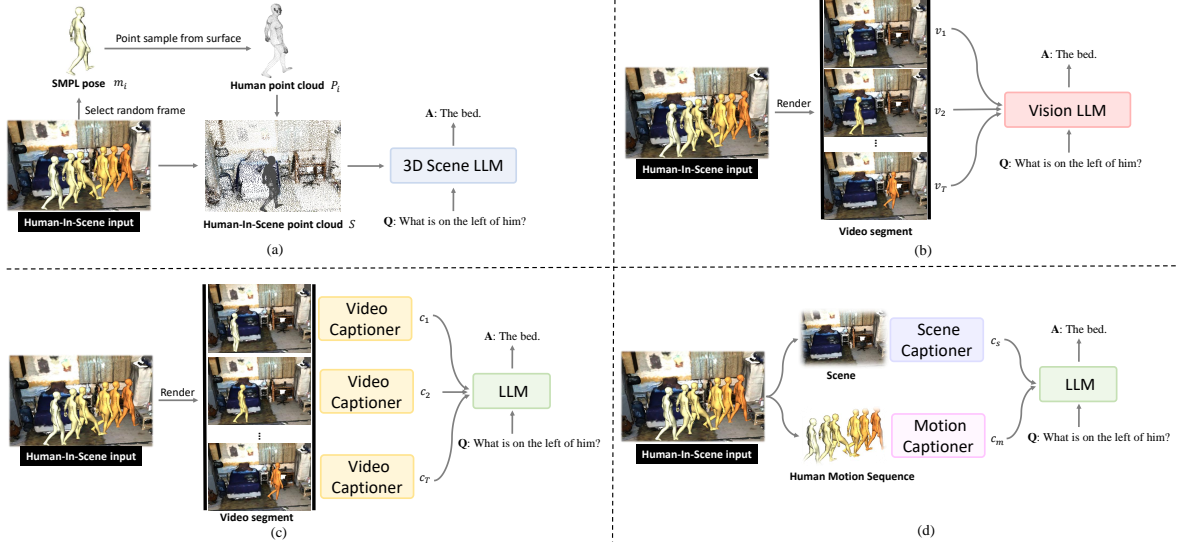


Figure S4. Illustrations of the HIS-QA baselines. (a) 3D Scene LLMs. (b) Vision LLMs. (c) LLMs w/ Frame Captions. (d) LLMs w/ Scene&Motion Captions.



Figure S5. A comparison between the RGB third-person video frame (left) and the rendered video frame from raw 3D HIS data (right).

approach, which is aware to 3D natures and human-scene interactions.

D.2. Ablations on Loss Weight

In Fig. S7, we conduct ablations on the weights of different loss components in the training loss of HIS-GPT stage 1. From the results, we observe that the optimal weights for the three losses are $\lambda_{act} = 0.5$, $\lambda_{spa} = 0.5$ and $\lambda_{cont} = 0.1$, respectively. Moreover, when the loss weights are too large, the performance on HIS-Bench declines. We argue that an excessively large weight of auxiliary tasks can interfere with the supervision of autoregressive objectives during LLM training.

D.3. Ablations on LLM Tuning Strategy

For HIS-GPT training, to preserve the instruction-following and generalization ability of the LLM backbone, we keep the entire LLM frozen during all training stages. Tab. S5 compares the performance of freezing the entire LLM with using LoRA [6] for LLM tuning. The results show that

Table S5. Ablations on the LLM tuning strategy of HIS-GPT.

LoRA		HIS-Bench							
Stage 1	Stage 2	Act.	Spa.	HoI.	Ana.	Pre.	Dia.	Pla.	Avg.
✗	✓	40.3	43.8	54.5	32.0	51.5	55.0	49.2	46.6
✓	✓	40.5	39.7	42.7	27.8	49.5	46.5	26.9	38.1
✗	✗	44.6	42.1	55.5	41.0	50.3	53.2	53.9	48.7

Table S6. Ablations on the input video type for Vision LLM evaluation. ‘Render’ denotes using the rendered videos from original 3D scene point cloud and 3D human SMPL data. ‘RGB’ denotes using the recorded video data in these HIS datasets, which is filmed by RGB cameras in third-person view.

Model	Video type	HIS-Bench							
		Act.	Spa.	HoI.	Ana.	Pre.	Dia.	Pla.	Avg.
Qwen-vl-max	RGB	25.6	15.7	36.0	12.0	15.7	29.5	17.8	21.5
	Render	28.7	17.6	37.1	13.4	14.5	33.0	22.1	23.5
GPT-4o	RGB	29.5	17.1	35.8	33.5	13.9	33.0	35.9	28.3
	Render	30.2	25.8	36.6	35.5	20.5	36.5	34.8	31.3

keeping the LLM frozen achieves better performance than adopting LoRA. This suggests that maintaining LLM frozen is crucial for achieving satisfactory performance in HIS understanding, likely because the inherent reasoning and generalization capabilities of the pre-trained LLMs are well preserved. Notably, performance significantly declines when using LoRA for modality alignment (Stage 1), as it may cause the LLM to overfit to the caption data, thereby losing instruction following and complex reasoning abilities to some extent.


```

[Instruction]
Please act as an impartial judge and evaluate the quality of the response provided by an AI assistant to the user question displayed below. Your evaluation should consider correctness and helpfulness. You will be given a reference answer and the assistant's answer. Begin your evaluation by comparing the assistant's answer with the reference answer. Be as objective as possible. Also, equally treat short and long answers and focus on the correctness of answers. Rate the response with either 0, 0.5, 1, 1.5 or 2:
0: The answer is responding off topic.
0.5: The answer addressed the question as requested.
1: The answer does not match well with the reference answer but not completely incorrect.
1.5: The answer almost matches with the reference answer, but some of the content is incorrect.
2: The information in the answer matches the reference answer.

Here is the input:
[Question]
{question}

[The Start of Reference Answer]
{gt_ans}
[The End of Reference Answer]

[The Start of Assistant's Answer]
{ans}
[The End of Assistant's Answer]

Use the following format:
Evaluation: your evaluations.
Rating: the rating you give, answer with only a float number, either 0, 0.5, 1, 1.5 or 2.

```

Figure S6. The prompt used for GPT-4 evaluation of HIS-Bench.

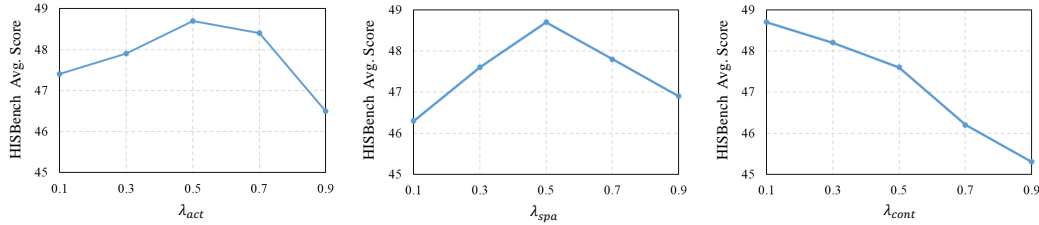


Figure S7. Ablations on the loss weight selection.

D.4. Video Format for Vision LLM Evaluation

For the evaluation of vision LLM baselines, we choose to render the HIS data into third-person videos and input them into the vision LLMs. This raises a concern that the domain gap between the rendered video and the common RGB input of vision LLMs could affect their performance. To explore this issue, we experiment with using the RGB video provided in the original data source of HIS-Bench for evaluating vision LLMs. A comparison between the RGB video and our rendered video is shown in Fig. S5. As shown in Tab. S6, using rendered video does not weaken the model's performance on HIS-Bench. In fact, on both Qwen-vl-max [2] and GPT-4o [7], rendered videos even exhibit higher performance than RGB videos. We hypothesize that the reason is that rendered videos from 3D data can more clearly present the spatial relationships between human and objects in the HIS data.

E. More Qualitative Results

In Fig. S11-S13, we present more qualitative results of HIS-GPT on HIS-Bench. Compared with existing vision-language baselines, HIS-GPT demonstrates a significant advantage in basic perception, complex reasoning, and embodied application abilities.

References

- [1] 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. 1, 2
- [2] 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
- [3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun

- Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 3
- [4] 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. 2, 4
- [5] 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. 4
- [6] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *ICLR*, 2022. 5
- [7] 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
- [8] Baoxiong Jia, Yixin Chen, Huangyue Yu, Yan Wang, Xuesong Niu, Tengyu Liu, Qing Li, and Siyuan Huang. Sceneverse: Scaling 3d vision-language learning for grounded scene understanding. In *ECCV*, 2024. 1, 4
- [9] 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. 4
- [10] 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. 3
- [11] 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. 2
- [12] Jonas Schult, Francis Engelmann, Alexander Hermans, Or Litany, Siyu Tang, and Bastian Leibe. Mask3d: Mask transformer for 3d semantic instance segmentation. In *ICRA*, 2023. 1
- [13] Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. Humanise: Language-conditioned human motion generation in 3d scenes. In *NeurIPS*, 2022. 1, 4

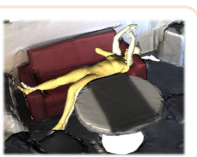

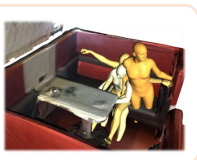
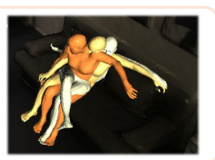


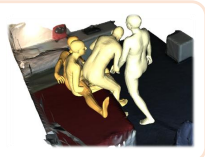
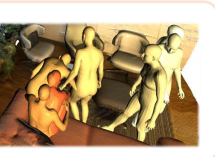
<p>Single Activity</p> <p>Q: What is the person doing?</p> <p>A: The person is lying on a red sofa.</p>		<p>Human Position</p> <p>Q: Where is the person positioned during the motion?</p> <p>A: Initially, the person is lying on the bed near the blanket, then they sit up on the bed facing the blanket. Later, they move away from the bed and the pillow, standing near the ottoman and facing the window. Finally, they are between the window and the bed, facing away from the ottoman.</p>	
<p>Sequential Activity</p> <p>Q: What does the person do after sitting?</p> <p>A: He stands up.</p>		<p>Interaction Type</p> <p>Q: What is the type of interaction between the person and the couch?</p> <p>A: He is sitting on the couch with his arms resting on the armrests.</p>	
<p>Body Orientation</p> <p>Q: What is the orientation of the table relative to the human when they stand up?</p> <p>A: The table is beside the human.</p>		<p>Object Orientation</p> <p>Q: What object is the person facing towards at the end?</p> <p>A: A table and a chair.</p>	
<p>Interacting Object</p> <p>Q: What object is the person's pelvis in contact with?</p> <p>A: A bed.</p>		<p>Contact Part</p> <p>Q: What body parts are touching the couch after the person has moved to the right?</p> <p>A: Pelvis, left hip, and right hip.</p>	

Figure S8. Examples of HIS-Bench, under the ‘**basic perception**’ core ability, with 8 sub-tasks in total.


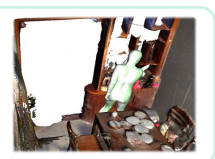
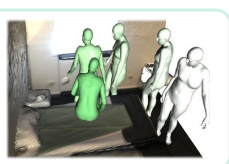
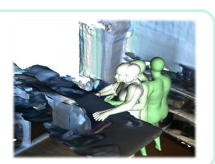
<p>Focus Analysis</p> <p>Q: What is the person currently focusing on?</p> <p>A: Curtain.</p>		<p>Situated Analysis</p> <p>Q: Can the person touch the shelf with his hands?</p> <p>A: Yes, because he stands in front of the shelf.</p>	
<p>Intent Prediction</p> <p>Q: What might the person intend to do next?</p> <p>A: He might start making the bed.</p>		<p>Movement Prediction</p> <p>Q: Where is the person likely to move next after standing up from the desk?</p> <p>A: The person is likely to move towards the shelf behind the desk.</p>	

Figure S9. Examples of HIS-Bench, under the ‘**complex reasoning**’ core ability, with 4 sub-tasks in total.

Situated Dialogue

Q: I'm feeling a bit tired, where can I sit down?



A: You can sit on the couch, which is right in front of you. There are also a few chairs around the room if you prefer.

High-level Task

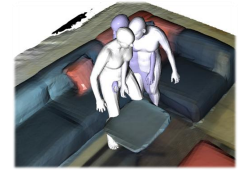
Q: What can you do to assist the person who is squatting near a table and then leaning back on one leg?



A: The person seems to be in an awkward position, possibly looking for something or needing support. I can bring them a chair to sit on or offer to help them find whatever they might be searching for under the table. If they need any items from the table, I can also pass those to them to make their task easier.

Low-level Task

Q: What can you do to assist the person standing near the sofa?



A: It seems like the person is standing near the sofa and might be looking for a place to sit down or relax. I can help by bringing them a pillow for added comfort. Here's how I can do it:

1. Locate the Pillow: Identify the location of the pillow. From the scene, it appears that there is one inside the couch.
2. Retrieve the Pillow: Approach the couch and gently remove the pillow from it.
3. Approach the Person: Walk towards the person who is standing near the sofa.
4. Offer the Pillow: Extend the pillow to the person and ask if they would like to sit down or use the pillow for support.
5. Assist with Seating: If they accept, help them find a comfortable spot on the sofa and place the pillow where they need it, whether it's behind their back or under their legs.

Navigation

Q: How does this person walk to the area near the computer?



A: Walk straight, then turn slightly right and continue walking straight to approach the computer.

Figure S10. Examples of HIS-Bench, under the 'embodied applications' core ability, with 4 sub-tasks in total.

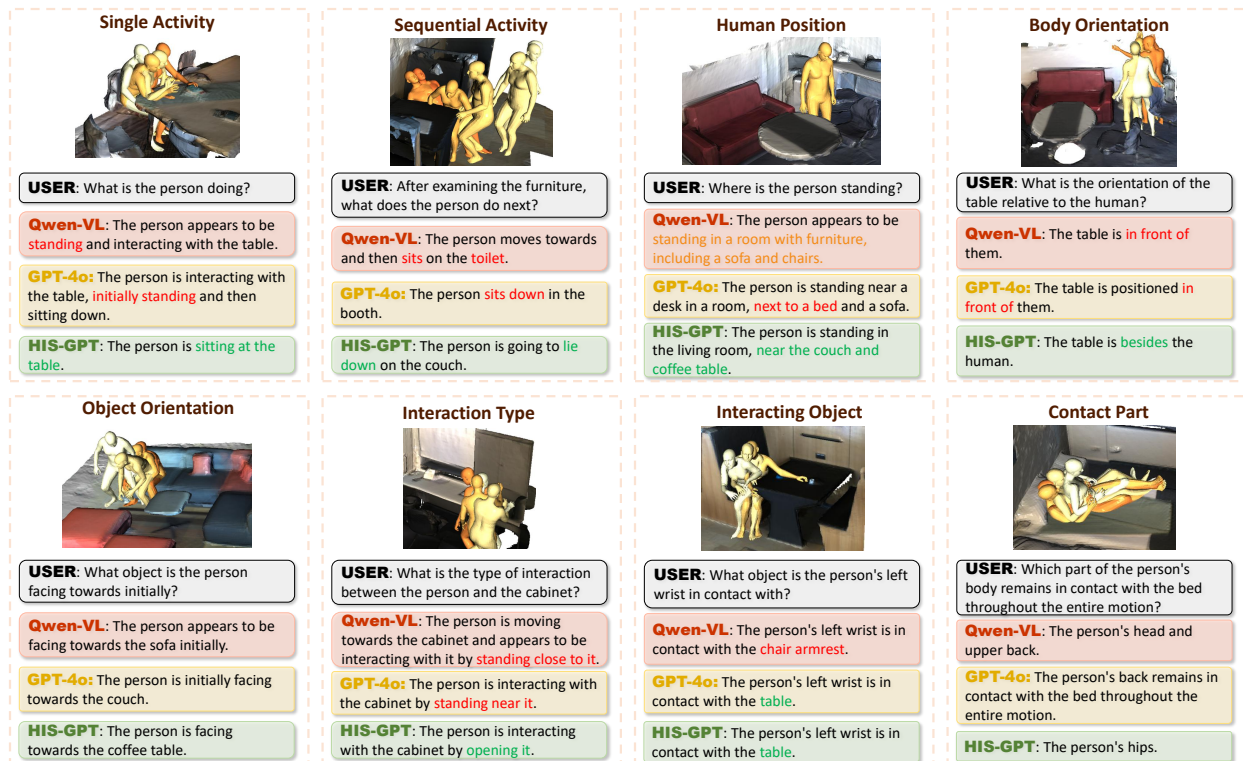


Figure S11. Qualitative examples of HIS-GPT on 'basic perception' sub-tasks in HIS-Bench, compared with Qwen-vl-max and GPT-4o.

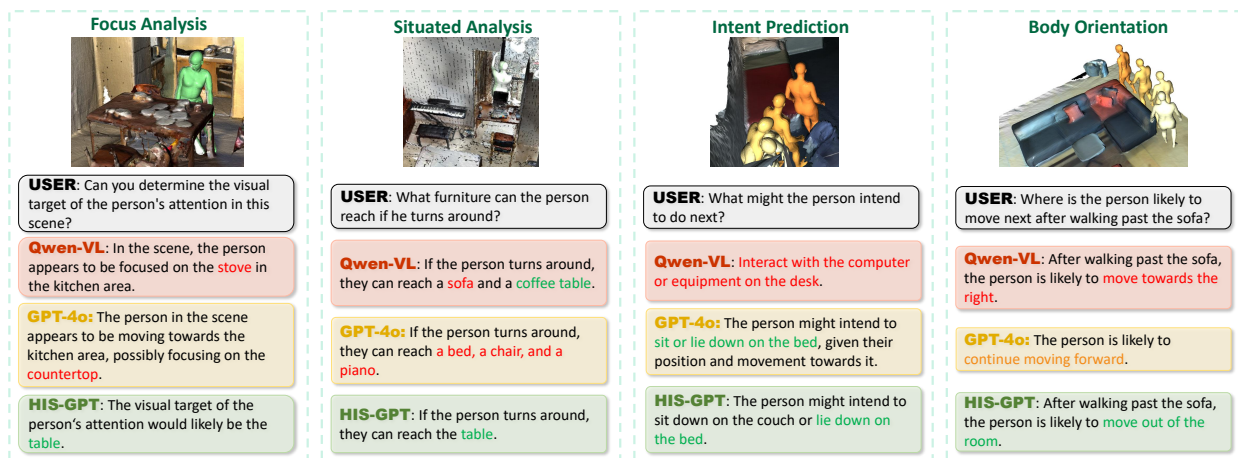


Figure S12. Qualitative examples of HIS-GPT on 'complex reasoning' sub-tasks in HIS-Bench, compared with Qwen-vl-max and GPT-4o.

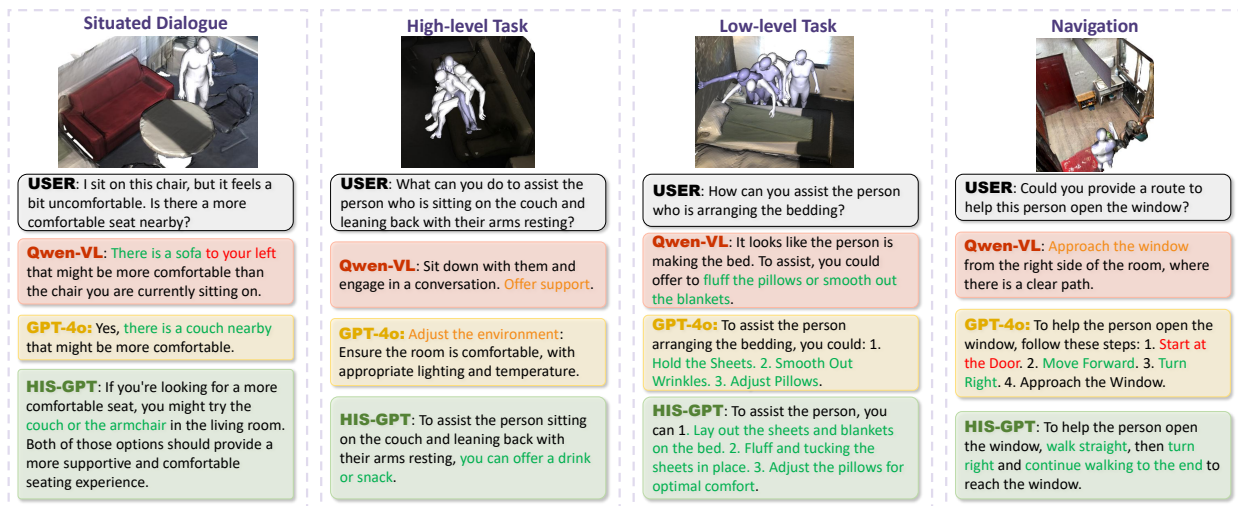


Figure S13. Qualitative examples of HIS-GPT on 'embodied application' sub-tasks in HIS-Bench, compared with Qwen-vl-max and GPT-4o.