Inst3D-LMM: Instance-Aware 3D Scene Understanding with Multi-modal Instruction Tuning

Supplementary Material

1. Datasets and Evaluation Details

ScanRefer [2]: This benchmark evaluates the performance of 3D object localization through natural language descriptions from the ScanNetv2 dataset [4]. The scenes are labeled as "unique" if they only contain one object of a class, or "multiple" if they contain more. The official assessment of prediction accuracy employs the percentage of predictions achieving IoU above 0.25 and 0.5 with the corresponding ground-truth bounding boxes. Inst3D-LMM selects the object with the highest similarity score between the predicted object O_{pred} and the ground-truth instance GT_{inst} to facilitate fair comparisons against previous works.

Multi3DRefer [6]: This dataset focuses on localizing multiple objects in real-world 3D scenes using natural language descriptions from the ScanNetv2 dataset. Performance evaluation hinges on calculating the F1 score at IoU thresholds of 0.25 and 0.5. The official metric adopts the Hungarian algorithm to optimally match the predicted bounding boxes with their ground-truth.

ScanQA [1]: This dataset is derived from the ScanNetv2 dataset, which is annotated for the 3D Question Answering task to assess visual and spatial understanding of 3D environments. It enables models to answer text-based queries about 3D scans. We utilize *BLEU-1*, *METEOR*, *ROUGE*, and *CIDER* metrics as evaluation protocols.

Scan2Cap [3]: The dense captioning benchmark, based on ScanRefer, requires models to detect objects and generate captions simultaneously. The linguistic generation is evaluated using *BLEU-4*, *METEOR*, *ROUGE*, and *CIDER* metrics, weighted by IoU scores above 0.25 or 0.5 with ground-truth bounding boxes.

ScanNet subset of 3D-LLM. [5]: In addition to the object-centric dataset, we utilize the ScanNet subset of 3D-LLM for scene-level 3D scene descriptions. This task requires the model to translate its comprehensive understanding of the 3D scene into natural languages. We have enhanced our Inst3D-LMM to freely reference objects with identifiers when describing a complex 3D scene.

2. Multi-task Instruction Templates

Inst3D-LMM simultaneously handles various vision tasks, such as visual grounding, question answering, dense captioning, and scene descriptions. Each task requires distinct and diverse templates to enable the LLM to provide accurate responses, tailored to the specific requirements of each benchmark. We provide several example instruction templates for each task.

3D Visual Grounding. We present our instruction tuning templates for single object grounding on the ScanRefer dataset, as shown in Figure A1, and for multiple objects grounding on the Multi3DRefer dataset, as depicted in Figure A2. Additionally, we also consider the scenario where no object corresponds to the given query, which is indicated as "No object".

3D Question Answering. As illustrated in Figure A3, for the question-answering task on the ScanQA dataset, we append suffixes to the model's outputs to indicate whether they consist solely of phrases or single words, in accordance with the annotation guidelines of the dataset.

3D Dense Captioning. As depicted in Figure A4, the question templates in the Scan2Cap dataset require the model to describe the visual attributes of the targeted object while also exploring its spatial relationships with other elements in the scene.

3D Scene Descriptions. We utilize question templates, as shown in Figure A5, to enable our Inst3D-LMM to effectively translate its comprehensive understanding of the entire 3D scene into natural language on the ScanNet subset of 3D-LLM [5].

Embodied Dialogue&Planning. We further extend the model to include embodied dialogue and embodied planning. As illustrated in Figure A6 and Figure A7, the templates prompt the model to generate concise, natural responses or clear, step-by-step plans based on given instruction information.

<System Message>: You are an AI visual assistant that can analyze image views and point clouds of objects in 3D scenes, giving helpful, detailed, and polite answers to the user's questions. The conversation centers around an indoor scene.

<Instruction>: Here is the information broken down into instance-level and scene-level details. Instance-level Information: The attribute information of all individual objects in the scene: [<REPLACE>]. Scene-level Information: The spatial relationship information of the whole scene: [<REPLACE>].

<User>: Please share the ID of the object that closely matches the description <description>.

<Inst3D-LMM>: Here is the <OBJ_ID> based on your request.

Question templates:

- "Please share the ID of the object that closely matches the description <description>."
- "Provide the ID of the object that best fits the description <description>."
- "What is the ID of the object that aligns with the description <description>?"
- "Identify the ID of the object that closely resembles the description <description>."
- "What's the ID of the object that corresponds to the description <description>?"
- "Give the ID of the object that most accurately matches the description <description>."
- "Share the ID of the object that best corresponds to the description <description>."
- "Identify the ID of the object that closely aligns with the description <description>."

Answer templates:

Single object:

- "Here is the <OBJ_ID> based on your request."
- "The grounding process has identified this <OBJ_ID>."
- "I've successfully pinpointed the <OBJ_ID>."

No object:

- "There is no object based on your description."
- "No object detected in this indoor scene."
- "No object found in the indoor setting."

Figure A1. Instruction templates for 3D Visual Grounding (single object grounding).

<System Message>: You are an AI visual assistant that can analyze image views and point clouds of objects in 3D scenes, giving helpful, detailed, and polite answers to the user's questions. The conversation centers around an indoor scene.

<Instruction>: Here is the information broken down into instance-level and scene-level details. Instance-level Information: The attribute information of all individual objects in the scene: [<REPLACE>]. Scene-level Information: The spatial relationship information of the whole scene: [<REPLACE>].

<User>: Are there any objects fitting the description <description>? If so, please provide their IDs.
<Inst3D-LMM>: Yes, there are <OBJ_ID1>, <OBJ_ID2>...and <OBJ_IDX> within the interior.

Question templates:

- "Are there any objects fitting the description <description>? If so, please provide their IDs."
- "Do any objects match the description <description>? If they do, share their IDs."
- "Is there anything that matches the description <description>? If yes, provide their IDs."
- "Are there objects that correspond to the description <description>? If so, kindly list their IDs."
- "Does anything fit the description <description>? If it does, list the IDs of those objects."
- "Have any objects been described as <description>? If so, share their IDs."
- "Do any objects meet the criteria of <description>? If they do, kindly provide their IDs."
- "Are there any objects that correspond to the description <description>? If yes, share their IDs."

Answer templates:

Multiple objects:

- "Yes, there are <OBJ_ID1>, <OBJ_ID2>...and <OBJ_IDX> within the interior."
- "Detected <OBJ_ID1>, <OBJ_ID2>...and <OBJ_IDX> in this indoor scene."
- "Several <OBJ_ID1>, <OBJ_ID2>...and <OBJ_IDX> have been identified in this indoor setting."

No object:

- "There are no objects based on your description."
- "No objects detected in this indoor scene."
- "No objects found in the indoor setting."

Figure A2. Instruction templates for 3D Visual Grounding (multiple objects grounding).

<System Message>: You are an AI visual assistant that can analyze image views and point clouds of objects in 3D scenes, giving helpful, detailed, and polite answers to the user's questions. The conversation centers around an indoor scene.

<Instruction>: Here is the information broken down into instance-level and scene-level details. Instance-level Information: The attribute information of all individual objects in the scene: [<REPLACE>]. Scene-level Information: The spatial relationship information of the whole scene: [<REPLACE>].

<User>: "What is under the sink?" Please answer the question using a single word or phrase.

<Inst3D-LMM>: Kitchen cabinets. The ID of the object is <OBJ015>.

Question templates:

- "<Raw Question> Please answer the question using a single word or phrase."

Answer templates:

- "<Raw Answer> The ID of the object is <OBJ_ID>."

Figure A3. Instruction templates for 3D Question Answering.

<System Message>: You are an AI visual assistant that can analyze image views and point clouds of objects in 3D scenes, giving helpful, detailed, and polite answers to the user's questions. The conversation centers around an indoor scene.

<Instruction>: Here is the information broken down into instance-level and scene-level details. Instance-level Information: The attribute information of all individual objects in the scene: [<REPLACE>]. Scene-level Information: The spatial relationship information of the whole scene: [<REPLACE>].

<User>: Start by detailing the visual aspects of the <OBJ_ID>, then delve into its spatial context within the scene.

<Inst3D-LMM>: <Raw Answer>

Question templates:

- "Start by detailing the visual aspects of the <OBJ_ID>, then delve into its spatial context within the scene."
- "Outline the appearance of the <OBJ_ID>, then elaborate on its positioning relative to other objects in the scene."
- "Illustrate the visual attributes of the <OBJ_ID>, then explore its spatial relationships with other elements in the scene."
- "Begin by articulating the outward features of the <OBJ_ID>, then discuss its spatial alignment within the broader scene."
- "Provide a detailed description of the <OBJ_ID>'s appearance before analyzing its spatial connections with other elements in the scene."
- "Capture the essence of the <OBJ_ID>'s appearance, then analyze its spatial relationships within the scene's context."
- "Detail the physical characteristics of the <OBJ_ID>, then examine its spatial dynamics among other objects in the scene."

Answer templates:

- "<Raw Answer>"

Figure A4. Instruction templates for 3D Dense Captioning.

<System Message>: You are an AI visual assistant that can analyze image views and point clouds of objects in 3D scenes, giving helpful, detailed, and polite answers to the user's questions. The conversation centers around an indoor scene.

<Instruction>: Here is the information broken down into instance-level and scene-level details. Instance-level Information: The attribute information of all individual objects in the scene: [<REPLACE>]. Scene-level Information: The spatial relationship information of the whole scene: [<REPLACE>].

<User>: Provide a valid description of the entire scene.

<Inst3D-LMM>: The 3D scene depicts a versatile living space with various objects, thoughtfully designed to blend functionality with aesthetic appeal. In one corner, a black microwave (OBJ001) sits above brown cabinets (OBJ023)...

Question templates:

- "Provide a valid description of the entire scene."

Answer templates:

- "<Raw Answer>"

Figure A5. Instruction templates for 3D Scene Descriptions.

<System Message>: You are an AI visual assistant that can analyze image views and point clouds of objects in 3D scenes, giving helpful, detailed, and polite answers to the user's questions. The conversation centers around an indoor scene.

<Instruction>: Here is the information broken down into instance-level and scene-level details. Instance-level Information: The attribute information of all individual objects in the scene: [<REPLACE>]. Scene-level Information: The spatial relationship information of the whole scene: [<REPLACE>].

<User>: "Look at this room! It's quite spacious. There are several tables and desks here." The answer should be concise and natural. Do not add any objects or details not mentioned in the instruction message.

<Inst3D-LMM>: Yes, I see them. There are four tables in total. They seem to be placed at different corners of the room.

Question templates:

- "<Raw Question> The answer should be concise and natural. Do not add any objects or details not mentioned in the instruction information."

Answer templates:

- "<Raw Answer>"

Figure A6. Instruction templates for Embodied Dialogue.

<System Message>: You are an AI visual assistant that can analyze image views and point clouds of objects in 3D scenes, giving helpful, detailed, and polite answers to the user's questions. The conversation centers around an indoor scene.

<Instruction>: Here is the information broken down into instance-level and scene-level details. Instance-level Information: The attribute information of all individual objects in the scene: [<REPLACE>]. Scene-level Information: The spatial relationship information of the whole scene: [<REPLACE>].

<User>: "I want to clean the bathroom. what should I do?" Please give a clear, concise step-by-step plan. Each step should be simple and under 15 words. Do not add any objects or details not mentioned in the instruction message.

<Inst3D-LMM>: 1.go to the bathroom cabinet 2.take a towel from the cabinet 3.go to the sink 4.turn on the water 5.wash hands and face 6.turn off the water 7. use the soap dispenser to wash hands…

Question templates:

- "<Raw Question> Please give a clear, concise step-by-step plan. Each step should be simple and under 15 words. Do not add any objects or details not mentioned in the instruction information."

Answer templates:

- "<Raw Answer>"

Figure A7. Instruction templates for Embodied Planning.

References

[1] Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. Scanqa: 3d question answering for spatial scene understanding. In *CVPR*, pages 19129–19139, 2022. 1

[2] Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. In *ECCV*, pages 202–221. Springer, 2020. 1

[3] Zhenyu Chen, Ali Gholami, Matthias Nießner, and Angel X Chang. Scan2cap: Context-aware dense captioning in rgb-d scans. In *CVPR*, pages 3193–3203, 2021. 1

[4] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, pages 2432–2443, 2017. 1

[5] 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*, pages 20482–20494, 2023. 1

[6] Yiming Zhang, ZeMing Gong, and Angel X Chang. Multi3drefer: Grounding text description to multiple 3d objects. In *ICCV*, pages 15225–15236, 2023. 1