CVP: Central-Peripheral Vision-Inspired Multimodal Model for Spatial Reasoning

Supplementary Material

Benchmarks	Scan2Cap				ScanQA					SQA3D
	CIDEr	BLEU-4	METEOR	ROUGE	CIDEr	BLEU-4	METEOR	ROUGE	EM	EM
GPT-4o [4]	7.0	5.5	17.1	38.5	43.3	16.6	16.3	31.6	11.1	41.7
CVP	90.5	41.7	28.9	62.2	107.1	17.8	20.8	50.9	31.2	62.3

Table 1. Comparison between GPT-4o and CVP across Scan2Cap, ScanQA, and SQA3D benchmarks.

1. Implementation Details

1.1. Evaluation with GPT Model

To compare with strong closed-source systems, we evaluate GPT-4o [4] on the benchmarks used in the main paper. Following the same strategy as CVP and Video-3D-LLM [7], we uniformly extract frames from the scene videos as input to generate visual tokens. Since GPT-4o produces only text output, it is largely unable to predict the 3D bounding-box coordinates required by the referring benchmarks (ScanRefer [2] and Multi3DRefer [6]). Therefore, we report its performance only on Scan2Cap [3], ScanQA [1], and SQA3D [5]. As shown in Table 1, without task-specific fine-tuning, even the powerful GPT-4o model falls short of the spatial understanding demonstrated by specialized models such as our proposed CVP. This gap is particularly evident on the Scan2Cap [3] benchmark, where precise object localization based on coordinate inputs is essential.

1.2. Task Relevant Objects

We describe how target objects are identified and task-relevant object sets $\mathcal{E}_+$ are constructed from the five training datasets.

In ScanRefer [2], each question is linked to one specific object from the given 3D scene. We directly use this object as the target object.

For Multi3DRefer [6], each question may refer to zero, one, or multiple objects in the 3D scene. Consequently, the number of target objects for a sample from this dataset ranges from zero to several instances.

In Scan2Cap [3], the model is tasked to give a precise description for a single referring object, with its 3D bounding box provided as input. Since the dataset provides the metadata of the object to be described for each data sample, we assign that object as the target.

ScanQA [1] provides a large collection of QA pairs, each accompanied by an object ID list and a corresponding name list of all objects relevant to the pair. Since some samples include multiple objects from different categories and distant locations, which may impede the learning of our task-affinity token, we apply a filtering strategy. Specifically, we retain only cases where (1) the object list contains a single instance, or (2) all object names are identical and matches the answer, which commonly occurs in questions such as "What is to the left of the chair?".

SQA3D [5] benchmarks the situated Question Answering task in 3D Scenes. It does not provide corresponding objects to their QA pairs. Therefore, we do not compute the InfoNCE loss for the task-affinity token when training on its samples.

1.3. Allocentric Grid Prompt Template

Given a set of 3D object bounding boxes, we retain their x- and y-axis coordinates and discretize them to place all objects into our allocentric grid. The grid is first represented as a dictionary, where each key corresponds to a 2D grid cell coordinate (x, y) and the value is the list of object names located in that cell. This dictionary is then used to populate the following text prompt:

```
This is a top-down view of a scene divided
into a {grid_H} by {grid_W} grid. Each cell
may contain multiple objects, and the
objects are separated by commas. This is an
abstraction of the scene and might be
incomplete.
At (row={x}, col={y}), there is: {obj_str},
At (row={x}, col={y}), there is: {obj_str},
...
```

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] 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. 1

[5] 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*, 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

[7] Duo Zheng, Shijia Huang, and Liwei Wang. Video-3d llm: Learning position-aware video representation for 3d scene understanding. In *CVPR*, pages 8995–9006, 2025. 1