ANCHORING THE OBSERVER: OBSERVER-CENTERIC POSITIONAL ENCODING FOR SITUATED 3D REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

We present an observer-centric framework for situated question answering (SQA), coined OC-SQA, that reasons about spatial relations from an observer's position and orientation. Existing SQA approaches are largely insensitive to situation descriptions, and observer-specific situation information is insufficiently exploited. To address this limitation, we explicitly incorporate observer poses for the positional encoding of visual tokens via a novel observer-centric rotary position embedding (OC-RoPE). OC-RoPE projects 3D locations of visual tokens onto a shared bird's-eye-view (BEV) plane and then transforms projected positions into the observer's reference frame for observer-centric positional encoding. We further introduce a learnable observer anchor at the head of the visual tokens of each frame to explicitly indicate the observer's origin. Together, these two designs provide direct geometric cues for situated reasoning. Extensive experiments demonstrate that our method achieves state-of-the-art performance on SQA3D and Real3DQA, while also demonstrating strong performance on MSQA. Further in-depth analyses show that our design increases the reliance on visual tokens relevent to the situated questions when predicting answers.

1 INTRODUCTION

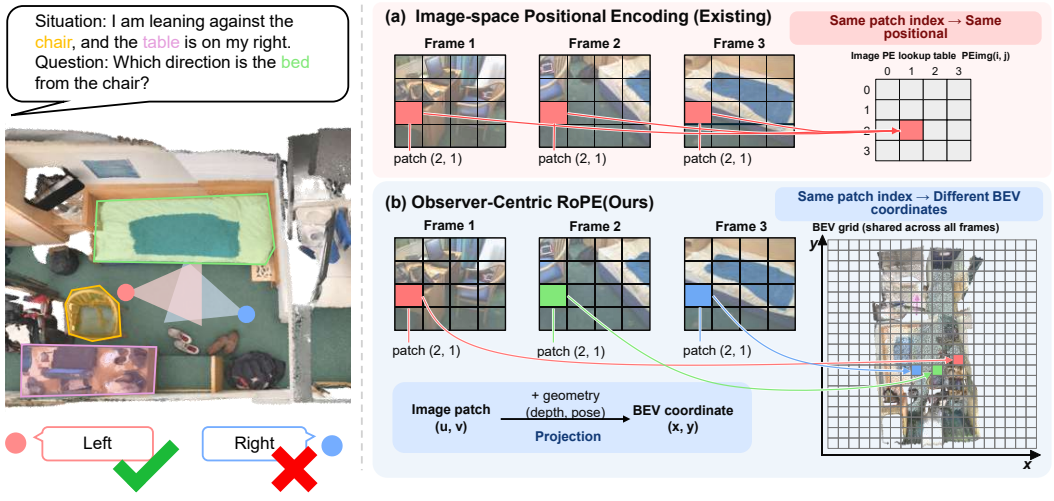

Figure 1: Illustration of SQA and observer-centric positional encoding. **Left:** An incorrect situation pose can directly lead to erroneous answers. For example, when the observer's heading changes, an object on the left may be incorrectly perceived as being on the right. **Right:** Comparison of RoPE-based encoding and the proposed OC-RoPE. Image-wise RoPE encoding relies on view-local patch coordinates and lacks a consistent representation of relative 3D positions across views and the observer's location, whereas OC-RoPE projects patches from different views onto a shared observer-centric BEV grid to preserve their spatial relationships in the observer's reference frame.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Situated question answering (SQA) requires an agent to understand a 3D environment and answer questions from a specific observer's location and orientation. Unlike conventional 3D question answering, where spatial relations can be defined in a global scene coordinate system, SQA requires reasoning in an observer-dependent reference frame. This distinction is fundamentally challenging because spatial relations depend on where and how the observer is situated. For example, an object that is to the observer's left can become to their right after the observer turns around, even though the underlying 3D scene remains unchanged (Fig. 1, left). Accurate situated reasoning therefore requires not only understanding the 3D scene structure but also explicitly grounding the reasoning in the observer's location and orientation.

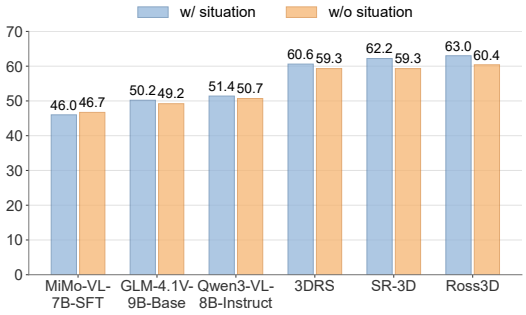

Figure 2: Performance drops of 2D/3D VLMs on SQA3D after completely removing situation descriptions. EM@1 is reported in percentages.

Recent advances have substantially improved 3D vision-language models (3D-VLMs) by incorporating richer geometric features, 3D positional encodings, and spatial supervision (Zheng et al., 2025b; Zhou et al., 2026; Wang et al., 2025; Gu et al., 2026). These developments strengthen a model's ability to represent the global structure of a 3D scene, but they do not necessarily ensure that spatial relations are interpreted with respect to a specified observer. This distinction motivates a fundamental question: *do existing 3D-VLMs genuinely use the observer's reference frame when performing situated reasoning?*

To answer this question, we conduct a diagnostic ablation on SQA3D (Ma et al., 2023) across six representative VLMs and 3D-VLMs. We compare each model with and without the textual situation description that specifies the observer's pose. Surprisingly, removing the situation description causes only a 1.3-percentage-point average decrease in EM@1 (Fig. 2). Despite the central role of the observer pose in defining the semantics of directional relations, these models exhibit only limited sensitivity to this information under the standard evaluation. This result exposes a gap between *representing 3D geometry* and *using observer-conditioned geometry*: a model may encode where objects are in the scene without effectively relating those objects to the observer whose perspective determines the answer.

Several recent approaches attempt to address this gap by introducing explicit pose representations (Ma et al., 2023; Qu et al., 2026) or observer-centric scene representations (Man et al., 2024; Liu et al., 2026). While these approaches establish a stronger connection between the observer and the scene, they typically introduce an additional pose embedding or coordinate frame, requiring the model to learn how it aligns with the directional semantics already encoded in its visual and language representations.

Our key insight for addressing the observer-conditioned gap is to represent observer-relative 3D geometry as a 2D spatial layout aligned with the native organization of visual tokens. In particular, we use an observer-centric bird's-eye-view (BEV) frame as a shared spatial reference. BEV provides a simple yet effective representation in which the relative direction and distance of each visual observation can be expressed directly as its position with respect to the observer. Rather than asking the model to learn a new mapping between an abstract pose representation and directional language, we encode observer-relative geometry directly into the positional encoding of visual tokens.

Based on this insight, we introduce *observer-centric rotary position embedding* (OC-RoPE). Given posed RGB-D observations, OC-RoPE projects visual-token locations onto the BEV plane and transforms them into the observer's coordinate frame. The resulting observer-centric BEV coordinates replace the two spatial indices of M-RoPE (Su et al., 2024; Wang et al., 2024), enabling visual tokens from different views to share a common positional structure centered on the observer. We further introduce a learnable *observer anchor* at the head of the visual tokens of each frame and assign the anchor to the BEV origin. This anchor explicitly represents the observer's location, such that the positional relationship between the anchor and each visual token directly encodes the token's displacement from the observer. Together, OC-RoPE and the observer anchor establish an explicit

path from the observer's physical state to the positional representation used for visual-language reasoning, allowing situation-dependent directional semantics to be grounded in observer-relative geometry.

We evaluate our approach on SQA3D (Ma et al., 2023), MSQA (Linghu et al., 2024), and the debiased Real3DQA benchmark (Ma et al., 2026). Our method consistently improves situated question answering, with particularly pronounced gains on directional questions, improving performance by 9.4 percentage points on SQA3D and 11.7 percentage points on Real3DQA. On SQA3D, it improves EM@1 over Ross3D (Wang et al., 2025) by 1.9 percentage points when using predicted poses. Further analysis shows that OC-RoPE and the observer anchor substantially increase the model's reliance on situation information and on visual tokens relevant to the queried spatial relations.

2 Related Work

2.1 3D VLMs

Recent years have witnessed significant progress in 3D VLM-based scene understanding. A prominent line of works aligns point-, object-, or superpoint-level 3D features with language embeddings through projection and cross-modal interaction (Hong et al., 2023; Wang et al., 2023; Li et al., 2024; Huang et al., 2024b; Chen et al., 2024; Huang et al., 2024a; Deng et al., 2025), whereas another line of works augments 2D visual tokens with depth priors or explicit 3D positional embeddings to ground image semantics spatially (Zheng et al., 2025b; Zhu et al., 2025; Chen et al., 2025; Cheng et al., 2026). Alternatively, some other methods integrate geometric and semantic information through pretrained geometry encoders or joint reconstruction branches (Zheng et al., 2025a; Fan et al., 2026; Chen et al., 2026; Zhao et al., 2025).

To support spatial reasoning, GPT4Scene, SpaceMind++, and QuatRoPE explicitly model spatial structure using BEV images, allocentric voxel maps, and quaternion-based relational encoding, respectively (Qi et al., 2026; Gu et al., 2026; Zhou et al., 2026). Other methods learn spatial representations through auxiliary reconstruction, localization, or geometric supervision (Wang et al., 2025; Qu et al., 2026; Zhang et al., 2026). However, many of these methods rely on the model to implicitly learn the coordinate transformation required to interpret scene geometry in the observer's reference frame. The weak sensitivity to situation descriptions in Fig. 2 suggests that such implicit learning may not sufficiently ground reasoning in the specified observer's position and orientation.

2.2 Situated Spatial Reasoning

Situated spatial reasoning interprets directional relations relative to the observer's position and orientation. Existing methods mainly follow two categories. The first incorporates explicit pose representations into VLMs. For example, SQA3D (Ma et al., 2023) learns pose representations via auxiliary pose prediction, while Loc3R-VLM (Qu et al., 2026) introduces dedicated pose tokens to condition scene representations on the agent's position and orientation. The second constructs observer-centric spatial representations to establish a shared frame for situation-dependent relations. SIG3D (Man et al., 2024) and Direction-aware (Liu et al., 2026) transform scene geometry into an observer-centric frame and recompute 3D positional encodings, while SPARTUN3D (Zhang et al., 2025) builds observer-conditioned scene graphs and aligns 3D object features with situated text. However, these approaches typically require additional pose or spatial representation learning to encode observer-relative geometry to the VLM. Alternatively, OneCanvas (Baranowski et al., 2026) positions visual tokens using observer-centric panoramic angles in native RoPE. However, angles alone do not convey Euclidean displacements and metric distances. In this work, we instead replace RoPE's two spatial indices with observer-relative BEV coordinates and introduce an observer anchor at the origin for situated 3D reasoning.

3 Method

We aim at situated question answering from the perspective of an observer at a given situation pose $\mathcal{S}$. The inputs to SQA consist of a sequence of posed RGB-D frames $\mathcal{V} = \{(\mathbf{I}_t, \mathbf{D}_t, \mathbf{T}_t)\}_{t=1}^{T}$, a situa-

tion description $\mathbf{s}$, and a question $\mathbf{q}$. Here, $\mathbf{I}_t$, $\mathbf{D}_t$, and $\mathbf{T}_t$ denote the RGB frame, the depth map, and the corresponding camera pose in the world coordinate frame, respectively. The observer's situation pose $\mathcal{S}$ is either provided as a prior or estimated from the input frames and situation description as $\mathcal{S} = g_\phi(\mathcal{V}, \mathbf{s})$. The output answer for SQA is formatted as text y.

Figure 3 summarizes our proposed framework, built on a pretrained VLM (Qwen3-VL-8B-Instruct (Team, 2025a) in our implementation). The VLM consists of a visual encoder, a text tokenizer, and a Large Language Model (LLM) for text prediction. In SQA, the visual encoder maps RGB frames into visual tokens in the LLM's embedding space, while the text tokenizer converts the situation description and question into text tokens. The LLM then processes the visual and text tokens to generate the answer. Our method promotes this VLM-based SQA with two novel designs: observer-centric rotary position embedding (OC-RoPE) and learnable observer anchors. OC-RoPE replaces the image-plane spatial indices in M-RoPE (Wang et al., 2024; Team, 2025a) with observer-centric BEV coordinates for positional encoding of visual tokens (Section 3.1). Furthermore, learnable observer anchor tokens provide an explicit spatial reference of the observer's location and are inserted at the head of each frame's visual tokens (Section 3.2).

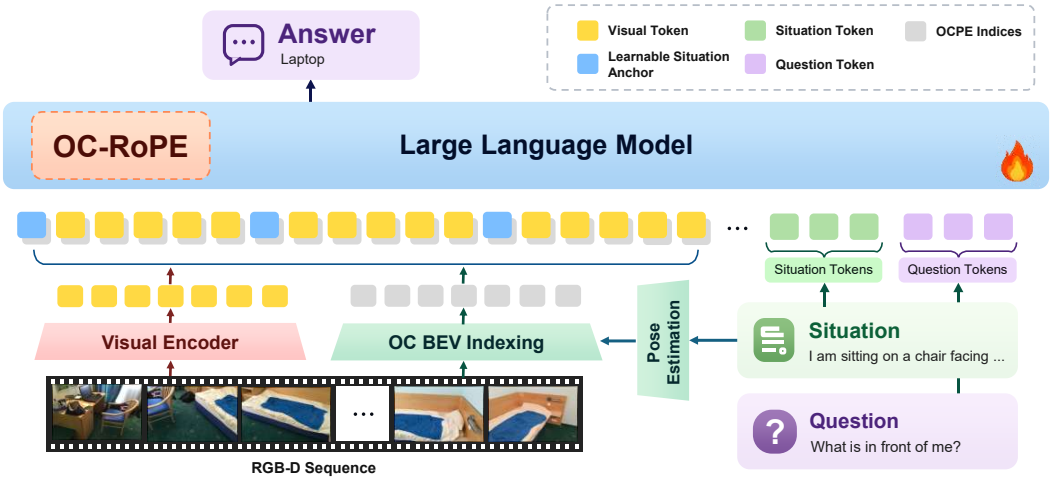

Figure 3: Overview of our framework. RGB frames are encoded into visual tokens (yellow), with a learnable observer anchor (blue) inserted at the head of each frame's tokens. Then OC-RoPE constructs observer-centric BEV position indices (gray) for positional encoding of each visual token and learnable observer anchor. The LLM processes these tokens together with situation tokens (green) and question tokens (purple) to generate the answer.

3.1 OC-RoPE

OC-RoPE computes observer-centric BEV positions of visual tokens. Specifically, for visual token j in frame t, we estimate its 3D location by back-projecting the corresponding pixels using the depth map and camera intrinsics and averaging the resulting 3D coordinates in the camera coordinate system. We denote this 3D token location in homogeneous coordinates by $\mathbf{p}_j^c \in \mathbb{R}^4$. To unify 3D token locations of each frame into a global 3D coordinate frame, we apply the camera-to-world rigid transformation matrix $\mathbf{T}_t \in \mathbb{R}^{4\times4}$:

$$\mathbf{p}_j = \mathbf{T}_t \mathbf{p}_j^c. \tag{1}$$

The mapped global point $\mathbf{p}_j$ is reduced to 2D coordinates in the BEV via $(p_x, p_y)^\mathsf{T} = P_{\text{BEV}}(\mathbf{p}_j)$. As illustrated in Fig. 1(b), these mapped global points represent the visual tokens in a shared world coordinate system, while their pairwise differences encode the relative displacements of visual tokens across frames.

However, the mapped global points do not account for relative perspectives with respect to an observer. To account for the observer's perspective, we transform the global BEV coordinates into an

observer-centric coordinate frame. Let $\mathcal{S} = (\mathbf{a}, \theta)$, where $\mathbf{a} \in \mathbb{R}^2$ is the observer's ground-plane position and θ is its yaw. The transformed coordinates are calculated as:

$$\widetilde{\mathbf{p}}_j = \begin{pmatrix} u_j \\ v_j \end{pmatrix} = \underbrace{\begin{pmatrix} \cos\theta & -\sin\theta \\ -\sin\theta & -\cos\theta \end{pmatrix}}_{\mathbf{R}(\theta)} \left(P_{\mathrm{BEV}}(\mathbf{p}_j) - \mathbf{a}\right). \tag{2}$$

where yaw is measured from the scene $+Y$ axis toward $+X$. Thus, points located in front of the observer have negative v_j.

To retain the integer position indexing used by the pretrained VLM (i.e., M-RoPE), we discretize the continuous observer-centric coordinates into a BEV grid with cell size $c = 0.1\,\mathrm{m}$:

$$\boldsymbol{\gamma}_j = Q(\widetilde{\mathbf{p}}_j) = \left\lfloor \frac{\widetilde{\mathbf{p}}}{c} \right\rfloor, \tag{3}$$

where $\boldsymbol{\gamma}_j = (\gamma_j^u, \gamma_j^v)^\mathsf{T}$ denotes the discrete observer-centric spatial index. Then the position indices for M-RoPE-based positional encoding for visual token j in frame t can be formatted as

$$\boldsymbol{\pi}_{t,j} = \left(\tau_t,\, \tau_1 + \gamma_{t,j}^v,\, \tau_1 + \gamma_{t,j}^u\right), \tag{4}$$

where τ_t denotes the temporal position assigned by the pretrained VLM to frame t and shared by all visual tokens within that frame. Following Qwen's native positional construction, the temporal and spatial dimensions of the first frame start from the same position offset. We therefore use τ_1 as the common spatial base for all frames. While τ_t varies across frames to preserve their native temporal organization, τ_1 remains fixed in the two spatial dimensions, ensuring that the same observer-centric grid cell receives the same spatial indices across different frames.

In this way, OC-RoPE modifies only the spatial positional indices of visual tokens. The visual features produced by the pretrained visual encoder, their temporal indices, and their sequence order remain unchanged. Text tokens likewise retain the original sequential positional encoding. OC-RoPE can therefore be integrated into the pretrained VLM without introducing an additional spatial encoder or altering the visual backbone.

3.2 Learnable Observer Anchors

OC-RoPE establishes an observer-centric coordinate system, but the observer itself is not explicitly represented as a token in the VLM's input tokens. We therefore introduce *learnable observer anchors* to explicitly instantiate the observer at the origin of the observer-centric spatial coordinate system. Specifically, we insert one learnable observer anchor immediately at the head of the visual tokens of each frame. These anchors are assigned to the same quantized observer origin:

$$\boldsymbol{\gamma}_0 = Q(\mathbf{0}). \tag{5}$$

For the anchor associated with frame t, its M-RoPE indices are defined as

$$\boldsymbol{\pi}_{t,\mathrm{anchor}} = \left(\tau_t,\, \tau_1 + \gamma_0^v,\, \tau_1 + \gamma_0^u\right), \tag{6}$$

where the anchor shares the temporal index τ_t with the visual tokens of frame t. This inserted observer token facilitates that visual tokens in each frame can directly access the spatial relationships relative to an observer via relative positional encoding-based token interactions. Importantly, standard inter-visual token interactions capture only pairwise spatial relationships and do not explicitly encode positional offsets with respect to the observer.

4 Experiments

4.1 Datasets and Metrics

Datasets. We report results on three benchmark, including SQA3D (Ma et al., 2023), Real3DQA (Ma et al., 2026), and MSQA (Linghu et al., 2024). SQA3D is a situated question-answering benchmark built from ScanNet (Dai et al., 2017). Real3DQA extends SQA3D in two

ways: it filters out shortcut-solvable questions and constructs four viewpoint variants for each situation pose to evaluate viewpoint robustness. MSQA expands the scene diversity by incorporating indoor scenes from ScanNet (Dai et al., 2017), 3RScan (Wald et al., 2019), and ARKitScenes (Dehghan et al., 2021). For MSQA, we replace the point-cloud input with RGB-D videos and camera poses from the source datasets while retaining the original QA annotations and observer situation poses. We follow the official splits for training and testing for SQA3D and MSQA. For Real3DQA, we directly used the trained model on SQA3D for evaluation.

Evaluation Metrics. For SQA3D, we report per-category and overall exact-match accuracy at rank 1 (EM@1) and its refined variant (EM@R1). For Real3DQA, we report EM@1 and EM@R1 on the debiased benchmark. For viewpoint robustness evaluation, we report per-category EM@R1 together with the Viewpoint Robustness Score (VRS), defined as the average fraction of question groups with at least k correct predictions across the four viewpoint variants, over $k \in 1, 2, 3, 4$. Correctness for each prediction is determined by EM@R1. For MSQA, we report per-category and overall GPT-Score following its official evaluation protocol (Linghu et al., 2024). Given the question, reference answer, and model response, a GPT-based evaluator assigns a correctness score $s_i \in \{1, 2, 3, 4, 5\}$ to each response. GPT-Score is computed as $\frac{1}{N}\sum_{i=1}^{N}\frac{s_i-1}{4} \times 100\%$, where N is the number of evaluated samples. All metrics are reported as percentages.

4.2 Training and Evaluation Protocol

Training Details. We instantiate the reasoning model with Qwen3-VL-8B-Instruct and jointly train LoRA adapters and the shared observer anchor embedding. The LoRA adapters take a rank of 32 and an alpha value of 64. The training consists of two stages. In stage 1, we use GT situation poses for training. In stage 2, to improve robustness to imperfect poses, we fine-tune the model with a mixture of GT and perturbed situation poses on the SQA3D training split. More details are provided in Appendix A.

Testing Details. We report our results on three settings based on how the situation poses are obtained. The *GT pose* setting directly uses the given situation pose for metric evaluation. The *Top-1* setting uses the situation pose assigned the highest score by an auxiliary pose estimator (see Appendix B). The *Top-3 oracle* setting selects the situation pose closest to the GT pose from the top three predictions. For MSQA, we use the GT setting for evaluation following the general setting.

4.3 Main Results

Results on SQA3D. Our method improves situated QA performance under all settings, as shown in Table 1. OC-SQA achieves the best performance of 64.9 EM@1 and 68.1 EM@R1 using the predicted Top-1 pose, an average gain of 1.9 on EM@1 and 2.4 on EM@R1. With the Top-3 oracle pose selection, the performance further improves to 66.7 EM@1 and 69.9 EM@R1. When GT poses are provided, our method reaches 71.4 EM@1 and 74.6 EM@R1, surpassing OneCanvas under the same setting by 6.1 and 6.2 points, respectively, with consistent improvements across all six question types. These results highlight the substantial impact of pose accuracy on situated QA performance; further analysis is provided in Appendix C.

OC-SQA achieves particularly pronounced improvements for *the directional reasoning of 'Which" questions*. Our Top-1 results outperform Ross3D by 9.4 points. Notably, over 75% of these questions begin with 'which direction" or "which way" and explicitly require directional reasoning. This substantial improvement supports our hypothesis that representing scene geometry in a unified observer-centric coordinate frame facilitates directional reasoning.

Results on Debiased Real3DQA. Our method maintains its advantage on the debiased Real3DQA, where textual shortcuts are reduced, as shown in Table 2 shows that OC-SQA achieves 42.6 EM@1 and 48.6 EM@R1 using Top-1 predicted poses, outperforming Ross3D by 4.9 points and SR-3D by 5.0 points. With GT poses, performance further improves to 51.4 EM@1 and 57.5 EM@R1. These results highlight the effectiveness of observer-conditioned geometry for viewpoint-robust situated QA.

As shown in Table 3, our method consistently improves performance across rotated viewpoints. The baseline results in the first five rows are taken from the original Real3DQA paper (Ma et al., 2026). When utilizing Top-1 estimated poses, our approach achieves a VRS of 20.7, outperforming

Table 1: Comparison with prior work on SQA3D. EM@1 and EM@R1 are reported in percentages.

Method	Per-question-type EM@1						EM@1	EM@R1
	What	Is	How	Can	Which	Others		
PQ3D (Zhu et al., 2024)	37.1	61.3	44.5	60.9	47.0	45.1	47.1	49.3
3D-VisTA (Zhu et al., 2023)	34.8	63.3	45.4	69.8	47.2	48.1	48.5	50.9
LEO (Huang et al., 2024b)	39.0	63.9	44.9	66.2	47.7	51.1	50.0	52.4
SIG3D (Man et al., 2024)	35.6	67.2	48.5	71.4	49.1	45.8	52.6	54.4
View2Cap (Yuan et al., 2025)	–	–	–	–	–	–	54.0	56.0
Scene-LLM (Fu et al., 2025)	40.9	69.1	45.0	70.8	47.2	52.3	54.2	56.2
Chat-Scene (Huang et al., 2024a)	45.4	67.0	52.0	69.5	49.9	55.0	54.6	57.5
SPARTUN3D (Zhang et al., 2025)	49.4	67.3	47.1	63.4	45.4	56.6	54.8	–
LLaVA-3D (Zhu et al., 2025)	–	–	–	–	–	–	60.1	-
Video-3D LLM (Zheng et al., 2025b)	51.1	72.4	55.5	69.8	51.3	56.0	58.6	60.8
GPT4Scene-HDM (Qi et al., 2026)	55.9	69.9	50.8	68.7	53.3	60.4	59.4	62.4
VLM-3R (Fan et al., 2026)	–	–	–	–	–	–	60.7	63.4
SpaceMind (Zhao et al., 2025)	54.1	74.8	61.7	71.0	51.9	53.6	61.1	63.8
3DRS (Huang et al., 2025)	53.9	76.5	60.4	70.1	49.9	56.9	60.6	63.2
SR-3D (Cheng et al., 2026)	55.0	76.4	59.8	71.6	54.7	61.1	62.2	–
Loc3R-VLM (Qu et al., 2026)	–	–	–	–	–	–	62.8	65.0
Ross3D (Wang et al., 2025)	56.0	79.8	60.6	70.4	55.3	60.1	63.0	65.7
Ours (Top-1 pose)	57.2	79.3	62.8	71.9	64.7	61.8	64.9	68.1
Ours (Top-3 oracle)	58.2	80.7	59.6	74.6	71.5	66.1	66.7	69.9
OneCanvas (Baranowski et al., 2026) (GT pose)	56.6	74.7	60.2	75.4	72.6	65.7	65.3	68.4
Ours (GT pose)	**63.1**	**81.6**	**66.2**	**78.4**	**78.6**	**72.3**	**71.4**	**74.6**

Table 2: Performance on debiased Real3DQA, where shortcut-solvable questions are filtered out. We report per-category EM@1 and overall EM@1 and EM@R1 (%) and evaluate our method using all three settings.

Method	What	Is	How	Can	Which	Others	All	
							EM@1	EM@R1
3DRS	35.0	43.2	38.2	28.6	28.2	27.4	34.3	38.9
SR-3D	35.9	45.0	42.2	26.0	38.9	35.6	37.6	43.6
Ross3D	36.5	47.3	40.0	31.2	37.6	33.8	37.7	43.0
Ours (Top-1 Pose)	38.5	53.3	43.6	33.8	54.4	40.2	42.6	48.6
Ours (Top-3 oracle)	39.8	57.4	44.0	40.3	59.7	45.2	45.3	51.4
Ours (GT Pose)	**45.8**	**56.8**	**46.7**	**49.4**	**72.5**	**55.3**	**51.4**	**57.5**

Table 3: Viewpoint robustness on Real3DQA across four rotation variants. "One" through "Four" report the percentages of question groups answered correctly under at least one to four rotations, respectively. EM@R1 (%) is reported for each question type.

Method	Correct rotations				VRS (%)	Question type			
	One	Two	Three	Four		Dist.	Dir.	Count.	Exist.
3D-LLM	33.2	4.1	1.1	0.1	9.6	5.5	24.6	24.2	44.2
Chat-3D v2	23.2	2.7	0.5	0.0	6.6	4.3	14.6	16.7	21.2
LEO	46.9	8.1	1.6	0.4	14.3	11.8	23.2	20.0	36.5
Chat-Scene	43.3	7.1	1.2	0.1	12.9	10.1	23.7	16.7	40.4
GPT4Scene	55.5	14.3	2.5	0.5	18.2	15.8	25.2	31.7	36.5
Ours (Top-1 Pose)	58.0	20.3	3.9	0.8	20.7	16.5	36.9	31.7	48.1
Ours (Top-3 oracle)	60.8	24.1	6.1	0.4	22.9	18.5	41.2	29.2	48.1
Ours (GT Pose)	**76.5**	**42.7**	**16.0**	**3.3**	**34.6**	**29.3**	**55.5**	**44.2**	**76.9**

the strongest baseline, GPT4Scene, by 2.5 percentage points. Under the Top-3 oracle setting, the VRS further rises to 22.9, whereas GT poses yield a peak performance of 34.6 VRS. This performance gap underscores that more accurate pose estimation remains a promising direction for future performance gains. Furthermore, for directional questions, our method improves EM@R1 from GPT4Scene's 25.2% to 36.9%. These results demonstrate the robustness and effectiveness of the observer-centric representation against viewpoint variations, even under estimated pose conditions.

Results on MSQA. The benefits of OC-SQA extend to MSQA under its GT-pose evaluation protocol (Table 5). Specifically, OC-SQA achieves an overall GPT-Score of 61.6, exceeding MSR3D by 5.1 points and LEO by 5.7 points. Improvements over MSR3D span five of the six question categories, including a 7.5-point gain on spatial questions. These results indicate that OC-SQA supports diverse reasoning tasks beyond SQA3D.

4.4 Ablation Studies

All ablations are conducted on SQA3D using the GT pose setting, with performance measured by EM@1. In Table 4, Qwen3-VL denotes the original model without task-specific adaptation. The LoRA baseline adapts Qwen3-VL to SQA3D by only tuning the LLM with LoRA, keeping all other components unmodified.

To validate the effectiveness of the observer-centric transformation (Eq. 2), we show the performance of using the observer pose-aligned transformation (*PA*) and using any random transformation (*RR*) in Table 4. Applying random rotation yields 63.0% EM@1, showing negligible improvement over the LoRA baseline (62.7%). In contrast, our pose-aligned OC-RoPE achieves a superior performance of 68.1% EM@1. Notably, the pose-aligned variant demonstrates a substantial gain of +19.1 percentage points over the random rotation setting on "Which"-type questions. This contrast highlights the critical role of observer-direction alignment in facilitating directional reasoning.

To evaluate the observer anchor design, we compare the final configuration with the third-to-last and penultimate configurations in Table 4. In our full model, SpaPE assigns each observer anchor the spatial coordinates of the observer at the BEV origin. For ablation, SeqPE instead encodes each anchor according to its position in the input sequence. With SpaPE fixed, replacing random rotation with pose alignment improves EM@1 from 63.1% to 71.4%. This suggests that grounding the anchor at the observer's location is insufficient without aligning the spatial coordinate axes with the observer's viewing direction. With pose alignment fixed, replacing SeqPE with SpaPE improves EM@1 from 69.2% to 71.4%, with gains across all question types. This supports encoding the anchor in the same spatial coordinate system as visual tokens, making their positional differences reflect observer-relative geometry.

Table 4: Component ablations on SQA3D. EM@1(%) is reported. RR, PA, SeqPE, and SpaPE denote random rotation, pose alignment, sequence positional encoding, and spatial positional encoding, respectively.

Qwen3-VL	LoRA	OC-RoPE		Anchor		What	Is	How	Can	Which	Others	All
		RR	PA	SeqPE	SpaPE							
✓						47.3	61.0	46.5	58.3	47.3	50.9	51.4
✓	✓					55.2	80.6	64.7	71.2	56.7	56.9	62.7
✓	✓	✓				55.5	80.1	59.1	73.1	57.5	59.0	63.0
✓	✓		✓			58.6	79.9	62.2	75.7	76.6	68.9	68.1
✓	✓	✓			✓	56.1	79.4	59.6	73.1	55.3	60.1	63.1
✓	✓		✓	✓		60.7	80.2	63.7	74.6	77.5	69.8	69.2
✓	✓		✓		✓	**63.1**	**81.6**	**66.2**	**78.4**	**78.6**	**72.3**	**71.4**

4.5 Spatial Analysis of Visual Evidence Reliance

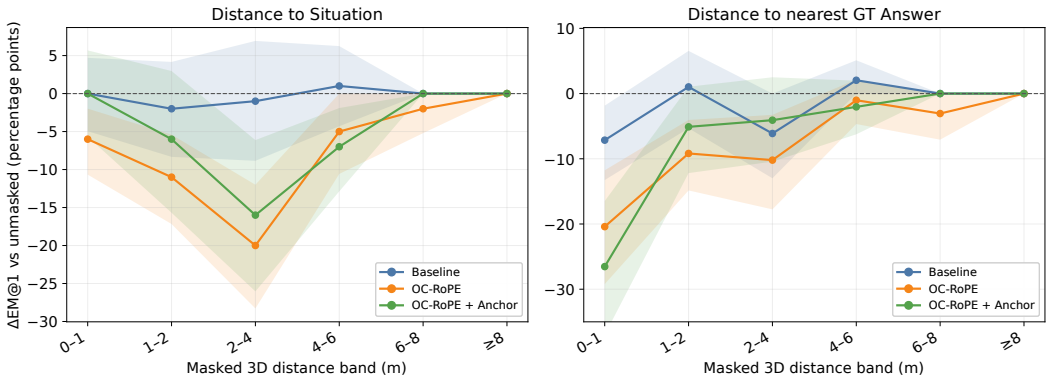

Figure 4: Change in EM@1 relative to the the original unmasked input when masking visual tokens in successive distance bands on the same subset of 100 SQA3D test samples. The left panel measures distance from the situation pose; the right panel measures distance from the manually annotated answer-relevant object location. Shaded bands represent 95% bootstrap confidence intervals around the mean change relative to the unmasked setting.

Motivated by the textual bias identified in Real3DQA (Ma et al., 2026), we analyze the reliance of SQA models on visual evidence through distance-based visual token masking. Specifically, we manually construct a subset of 100 SQA3D test samples, each annotated with the situation pose and the target object associated with the ground-truth answer. We then progressively mask visual tokens according to their distances from either the situation position or the target answer object and evaluate how the removal of visual evidence at different spatial ranges affects SQA performance. Additional attention visualizations and qualitative analyses are provided in Appendix E.

Sensitivity to a Situation Position. Intuitively, visual tokens close to a situation position (the observer's location) should contribute more substantially to the final answer. Figure 4 (left) illustrates the performance of three variants, including the LoRA-tuning variant (baseline), the OC-RoPE variant (also finetuned with LoRA), and the full architecture with all modules. Within a spatial radius of 0-8m from the situation position, the OC-RoPE variant exhibits markedly higher sensitivity to token masking than the LoRA baseline. Specifically, masking visual tokens within the 2-4m range degrades EM@1 by approximately 20.0 percentage points for OC-RoPE and 16.0 percentage points for the full model. In contrast, the baseline's performance fluctuates by at most 2.5 percentage points across all masking thresholds. This performance contrast indicates that observer-centric positional encoding significantly enhances the model's reliance on localized visual evidence around the specified situation.

Sensitivity to Answer-relevant Positions. Inserting the learnable observer anchor token helps the model better localize answer-relevant objects. We show the relative changes with respect to the masking spatial bands in Figure 4 (right). Specifically, masking visual tokens within the 0-1m spatial band degrades EM@1 by approximately 27.0 percentage points for the full model, compared to 21.0 percentage points for the OC-RoPE variant and merely 7.0 percentage points for the LoRA baseline. This pronounced accuracy drop near the target object, coupled with greater robustness beyond 2m demonstrates that the observer anchor effectively concentrates the model's visual reliance onto answer-relevant regions.

Table 5: MSQA GPT-Score (%). Following the official MSQA protocol and prior methods, we evaluate under the GT-pose setting.

Method	Count	Exist.	Attr.	Spatial	Navi.	Others	Overall
LEO	36.2	88.5	51.7	46.9	56.8	80.3	55.9
MSR3D	33.5	87.5	53.7	48.9	**61.9**	75.0	56.5
Ours	**37.3**	**90.3**	**65.6**	**56.4**	57.2	**85.0**	**61.6**

Table 6: Results on the challenging subset. EM@1 (%) is reported.

Method	What	Is	How	Can	Which	Others	All
3DRS	24.4	38.6	27.7	38.5	36.8	28.1	29.9
SR-3D	22.9	38.6	29.5	40.4	41.4	36.0	31.8
Ross3D	24.0	40.7	30.1	37.5	39.1	30.7	31.0
Ours (Top-1 pose)	24.6	42.8	27.8	38.5	55.2	32.0	33.7
Ours (Top-3 oracle)	27.5	49.0	30.1	42.3	60.3	37.3	37.7
Ours (GT pose)	**33.6**	**53.8**	**38.2**	**51.9**	**72.4**	**51.3**	**46.4**

4.6 Evaluation on the Challenging SQA3D Subset

To validate whether the gains persist on more difficult cases, we construct a challenging subset from SQA3D test set using a LoRA-tuned Qwen3-VL-8B-Instruct that receives no situation descriptions during either training or inference. We exclude questions that this model answers correctly, yielding a challenging subset of 1,279 questions.

On this challenging subset, OC-SQA achieves 33.7 EM@1 with Top-1 predicted poses, exceeding SR-3D, the strongest competing baseline, by 1.9 points (see Table 6). Using the Top-3 oracle setting and the GT setting raises OC-SQA's EM@1 to 37.7 and 46.4, respectively. The large improvement in the GT setting suggests that the accuracy of estimated situation poses remains important on the filtered questions. The results show that the largest gains occur on "Which"-type questions, indicating that the observer-centric representation is key to observer-relative directional reasoning.

5 Conclusion

We studied a key limitation of existing SQA models: the situation information is usually neglected by VLMs despite its key role in defining the observer's reference frame. To address this issue, we introduced OC-RoPE to embed the position of visual tokens in an observer-centric BEV coordinate frame and a learnable observer anchor for explicitly representing the observer as a spatial reference for visual reasoning. Experiments on three benchmarks demonstrate consistent improvements in situated spatial reasoning, particularly for directional questions and debiased evaluations. Further analyses show that our representation enhances the alignment between situation information and relevant visual evidence. These results highlight the effectiveness of observer-centric representations for enabling pretrained VLMs to better understand situated environments. Future work will explore longer observation sequences, dynamic scenes, and more diverse environments.

AI USE STATEMENT

We used generative AI tools to assist with language editing during manuscript preparation. Specifically, these tools were used to rephrase sentences, check grammar, improve readability, and enhance the overall flow and clarity of the manuscript. We also used AI tools to assist with code development. All AI-assisted code was reviewed and verified by the authors.

The initial manuscript was written by the authors. Generative AI tools were not used for research ideation, hypothesis formulation, methodological or experimental design, data analysis, interpretation of results, or the generation of scientific claims or conclusions. All research concepts, methodological decisions, experimental designs, analyses, and scientific conclusions were developed by the authors.

All AI-assisted edits were reviewed and revised by the authors to ensure accuracy and preserve the intended scientific meaning. The authors take full responsibility for the final content of the manuscript, including all text edited with the assistance of generative AI.

ETHICS STATEMENT

This work adheres to the ICLR Code of Ethics. No human subjects or animal experimentation were involved. All datasets used in this study, including SQA3D, Real3DQA, MSQA, ScanNet, 3RScan, and ARKitScenes, were obtained and used in accordance with their respective usage guidelines. No new private or personally identifiable data were collected. We also acknowledge that the models and datasets may contain inherent biases, and that errors in perception or spatial reasoning may affect model outputs. The authors are committed to responsible data use, transparency, and research integrity.

REPRODUCIBILITY STATEMENT

We will publicly release the implementation code and the manually crafted dataset to facilitate reproduction and further research.

REFERENCES

Bartlomiej Baranowski, Dave Zhenyu Chen, and Matthias Nießner. OneCanvas: 3D scene understanding via panoramic reprojection. *arXiv preprint arXiv:2606.19253*, 2026.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment anything with concepts. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/e0982cbc81401df3430ee1ff780dc7a2-Abstract-Conference.html.

Haijier Chen, Bo Xu, Shoujian Zhang, Haoze Liu, Jiaxuan Lin, and Jingrong Wang. Vid-LLM: A compact video-based 3D multimodal LLM with reconstruction–reasoning synergy. In *International Conference on Learning Representations*, 2026.

Pingyi Chen, Yujing Lou, Shen Cao, Jinhui Guo, Lubin Fan, Yue Wu, Lin F. Yang, Lizhuang Ma, and Jieping Ye. SD-VLM: Spatial measuring and understanding with depth-encoded vision-language models. In *Advances in Neural Information Processing Systems*, 2025.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

An-Chieh Cheng, Yang Fu, Yukang Chen, Zhijian Liu, Xiaolong Li, Subhashree Radhakrishnan, Song Han, Yao Lu, Jan Kautz, Pavlo Molchanov, Hongxu Yin, Xiaolong Wang, and Sifei Liu. 3d aware region prompted vision language model. In *International Conference on Learning Representations*, 2026.

Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas A. Funkhouser, and Matthias Nießner. ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

Afshin Dehghan, Gilad Baruch, Zhuoyuan Chen, Yuri Feigin, Peter Fu, Thomas Gebauer, Daniel Kurz, Tal Dimry, Brandon Joffe, Arik Schwartz, and Elad Shulman. ARKitScenes - a diverse real-world dataset for 3D indoor scene understanding using mobile RGB-D data. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*, 2021.

Jiajun Deng, Tianyu He, Li Jiang, Tianyu Wang, Feras Dayoub, and Ian D. Reid. 3D-LLaVA: Towards generalist 3D LMMs with omni superpoint transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Zhiwen Fan, Jian Zhang, Renjie Li, Junge Zhang, Runjin Chen, Hezhen Hu, Kevin Wang, Peihao Wang, Huaizhi Qu, Shijie Zhou, Dilin Wang, Zhicheng Yan, Hongyu Xu, Justin Theiss, Tianlong Chen, Jiachen Li, Zhengzhong Tu, Zhangyang Wang, and Rakesh Ranjan. VLM-3R: Vision-language models augmented with instruction-aligned 3D reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Rao Fu, Jingyu Liu, Xilun Chen, Yixin Nie, and Wenhan Xiong. Scene-LLM: Extending language model for 3D visual reasoning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2025.

Bo Gu, Zhikang Zhang, Zizhuang Wei, Zhenyuan Chen, Lingyun Li, and Zhuoyi Song. SpaceMind++: Toward allocentric cognitive maps for spatially grounded video MLLMs. *arXiv preprint arXiv:2605.09449*, 2026.

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 *Advances in Neural Information Processing Systems*, 2023.

Haifeng Huang, Yilun Chen, Zehan Wang, Rongjie Huang, Runsen Xu, Tai Wang, Luping Liu, Xize Cheng, Yang Zhao, Jiangmiao Pang, and Zhou Zhao. Chat-Scene: Bridging 3D scene and large language models with object identifiers. In *Advances in Neural Information Processing Systems*, 2024a.

Jiangyong Huang, Silong Yong, Xiaojian Ma, Xiongkun Linghu, Puhao Li, Yan Wang, Qing Li, Song-Chun Zhu, Baoxiong Jia, and Siyuan Huang. An embodied generalist agent in 3D world. In *Proceedings of the 41st International Conference on Machine Learning*, 2024b.

Xiaohu Huang, Jingjing Wu, Qunyi Xie, and Kai Han. 3DRS: MLLMs need 3D-aware representation supervision for scene understanding. In *Advances in Neural Information Processing Systems*, 2025.

Xiang Li, Jian Ding, Zhaoyang Chen, and Mohamed Elhoseiny. Uni3DL: Unified model for 3D and language understanding. In *Computer Vision - ECCV 2024 - 18th European Conference*, 2024.

Xiongkun Linghu, Jiangyong Huang, Xuesong Niu, Xiaojian (Shawn) Ma, Baoxiong Jia, and Siyuan Huang. Multi-modal situated reasoning in 3D scenes. In *Advances in Neural Information Processing Systems*, 2024.

Quan Liu, Weihao Xuan, Junjue Wang, Naoto Yokoya, Ling Shao, and Shijian Lu. Direction-aware 3D large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Xianzheng Ma, Tao Sun, Shuai Chen, Yash Bhalgat, Jindong Gu, Angel X. Chang, Iro Armeni, Iro Laina, Songyou Peng, and Victor Adrian Prisacariu. Do 3D large language models really understand 3D spatial relationships? In *International Conference on Learning Representations*, 2026.

Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. SQA3D: Situated question answering in 3D scenes. In *International Conference on Learning Representations*, 2023.

Yunze Man, Liang-Yan Gui, and Yu-Xiong Wang. Situational awareness matters in 3D vision language reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Zhangyang Qi, Zhixiong Zhang, Ye Fang, Jiaqi Wang, and Hengshuang Zhao. GPT4Scene: Understand 3D scenes from videos with vision-language models. In *International Conference on Learning Representations*, 2026.

Kevin Qu, Haozhe Qi, Mihai Dusmanu, Mahdi Rad, Rui Wang, and Marc Pollefeys. Loc3R-VLM: Language-based localization and 3D reasoning with vision-language models. *CoRR*, 2026.

Jianlin Su, Murtadha H. M. Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 2024. doi: 10.1016/J.NEUCOM.2023.127063.

Qwen Team. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Qwen Team. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025b.

Johanna Wald, Armen Avetisyan, Nassir Navab, Federico Tombari, and Matthias Nießner. RIO: 3D object instance re-localization in changing indoor environments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.

Haochen Wang, Yucheng Zhao, Tiancai Wang, Haoqiang Fan, Xiangyu Zhang, and Zhaoxiang Zhang. Ross3D: Reconstructive visual instruction tuning with 3D-awareness. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *CoRR*, 2024.

Zehan Wang, Haifeng Huang, Yang Zhao, Ziang Zhang, and Zhou Zhao. Chat-3D: Data-efficiently tuning large language model for universal dialogue of 3D scenes. *CoRR*, 2023. doi: 10.48550/ARXIV.2308.08769.

Zhihao Yuan, Yibo Peng, Jinke Ren, Yinghong Liao, Yatong Han, Chun-Mei Feng, Hengshuang Zhao, Guanbin Li, Shuguang Cui, and Zhen Li. Empowering large language models with 3D situation awareness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Chushan Zhang, Ruihan Lu, Jinguang Tong, Yikai Wang, and Hongdong Li. 3D-IDE: 3D implicit depth emergent. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Yue Zhang, Zhiyang Xu, Ying Shen, Parisa Kordjamshidi, and Lifu Huang. SPARTUN3D: Situated spatial understanding of 3D world in large language model. In *International Conference on Learning Representations*, 2025.

Ruosen Zhao, Zhikang Zhang, Jialei Xu, Jiahao Chang, Dong Chen, Lingyun Li, Weijian Sun, and Zizhuang Wei. SpaceMind: Camera-guided modality fusion for spatial reasoning in vision-language models. *CoRR*, 2025.

Duo Zheng, Shijia Huang, Yanyang Li, and Liwei Wang. Learning from videos for 3D world: Enhancing MLLMs with 3D vision geometry priors. In *Advances in Neural Information Processing Systems*, 2025a.

Duo Zheng, Shijia Huang, and Liwei Wang. Video-3D LLM: Learning position-aware video representation for 3D scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025b.

Shengli Zhou, Minghang Zheng, Feng Zheng, and Yang Liu. Scalable object relation encoding for better 3D spatial reasoning in large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Chenming Zhu, Tai Wang, Wenwei Zhang, Jiangmiao Pang, and Xihui Liu. LLaVA-3D: A simple yet effective pathway to empowering LMMs with 3D capabilities. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

Ziyu Zhu, Xiaojian Ma, Yixin Chen, Zhidong Deng, Siyuan Huang, and Qing Li. 3D-VisTA: Pre-trained transformer for 3D vision and text alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

Ziyu Zhu, Zhuofan Zhang, Xiaojian Ma, Xuesong Niu, Yixin Chen, Baoxiong Jia, Zhidong Deng, Siyuan Huang, and Qing Li. Unifying 3D vision-language understanding via promptable queries. In *European Conference on Computer Vision*, 2024.

Table A.1: The stage-I training configuration.

Parameter	Value
Backbone	Qwen3-VL-8B-Instruct
Numerical precision	bfloat16
Frames per scene	32, uniformly sampled
LoRA rank / α	32 / 64
LoRA dropout	0.05
Optimizer	AdamW
Learning rate	1×10^{-4}
Weight decay	0
Learning-rate scheduler	Cosine decay
Warmup ratio	0.03
Maximum gradient norm	1.0
Training epochs	1
Training batch size	16

Table A.2: The stage-II adaptation configuration.

Parameter	Value
Initialization	Stage-I checkpoint (warm-start)
Backbone	Qwen3-VL-8B-Instruct
Numerical precision	bfloat16
Frames per scene	32, uniformly sampled
LoRA rank / α	32 / 64
LoRA dropout	0.05
Optimizer	AdamW
Learning rate	2×10^{-5}
Weight decay	0
Learning-rate scheduler	Cosine decay
Warmup ratio	0.03
Maximum gradient norm	1.0
Training batch size	16
Training steps	4,966

A Detailed Training Configuration

For Stage 1 training on SQA3D, we use Qwen3-VL-8B-Instruct with 32 uniformly sampled RGB-D frames per scene. The BEV grid of OC-RoPE is discretized with 0.1 m cells, without scale normalization or additional yaw offsets. We jointly train the shared anchor token and LoRA adapters inserted for `q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, and `down_proj`. Table A.1 summarizes the training hyperparameters. All experiments are conducted on a single NVIDIA A100 GPU.

For Stage 2, we initialize the model with the Stage-1 checkpoint and further tune the existing LoRA adapters and shared observer anchor token on SQA3D. We run the pose estimator on the SQA3D training split and compare its predictions with GT poses. The resulting local-frame errors $(\Delta r, \Delta f, \Delta\theta)$ form an empirical error bank. During training, local frame error is sampled for each training example from this bank and is used to perturb the training pose's position and yaw. Training settings are listed in Table A.2.

B Situation Pose Estimation

Given a sequence of posed RGB-D observations $\mathcal{V}$ and a situation description $\mathbf{s}$, the auxiliary pose estimator g_ϕ predicts the observer's planar pose $\mathcal{S} = g_\phi(\mathcal{V}, \mathbf{s}) = (\mathbf{a}, \theta)$, where $\mathbf{a}$ and θ denote the observer's ground-plane position and yaw, respectively. The estimated pose provides the observer-centric reference frame for downstream question answering in Section 3.

To infer this pose, we use Qwen3-8B (Team, 2025b) to extract the spatial relations between the observer and the objects mentioned in the situation description (Section B.1), and locate these objects in 3D through SAM 3-based instance segmentation (Carion et al., 2026), followed by mask back-projection and multi-view fusion (B.2). Together, the extracted relations and localized objects allow us to evaluate possible observer poses. We sample candidate positions and headings in the scene and train an MLP scorer to assess how well each candidate agrees with the described relations (B.3). The highest-scoring candidate is then selected as the estimated pose. We train the estimator on SQA3D and directly reuse it on Real3DQA without additional training or adaptation, as the two datasets share a similar style of situation descriptions.

B.1 Spatial Relation Extraction

We extract three types of spatial relations between the observer and the objects mentioned in a situation description:

- **Anchor relations** identify objects that constrain the observer's position independently of heading, such as the chair in "I am sitting on a chair."
- **Coarse directional relations** indicate whether an object is in front of, behind, to the left of, or to the right of the observer.
- **Fine-grained directional relations** specify an object's direction using the twelve clock-face positions.

The extractor first identifies the mentioned objects, then queries each object's anchor and coarse directional relations. Fine-grained directional relations are queried only when the description contains an explicit clock-direction expression. All stages use Qwen3-8B (Team, 2025b), with thinking mode disabled for object extraction and enabled for relation queries. Each query uses the corresponding prompt below as the system message and the situation description as the user message. In these prompts, {`object`} denotes an extracted object phrase, and {`relation`} is instantiated separately with `front`, `behind`, `left`, and `right`.

Object extraction. The object-extraction prompt identifies only physical objects explicitly mentioned in the description:

```
You are an information-extraction system.
Read one situation description and extract every physical
object explicitly mentioned in it. Do not infer objects that
are not mentioned. Include objects that support, contain,
surround, or are occupied by the speaker.

Return exactly one JSON object, with no markdown or explanation:
{"objects":["lowercase object phrase"]}

Use a short lowercase object phrase. Return each object phrase
once. If no physical object is mentioned, return {"objects":[]}.
```

Anchor constraints. For each extracted object, we query the yaw-invariant *anchor* relation:

```
Answer yes only if the situation explicitly places me on, in,
inside, surrounded by, or physically occupying {object}; or if
it says I am directly using, operating, manipulating, repairing,
or fixing {object}. Those direct interactions mean I am
immediately next to the object. Answer yes or no.
```

Cardinal-direction constraints. We apply one shared template to every object–relation pair. In contrast to relation-specific prompts, this shared decision rule avoids prompt-induced asymmetry across the four directions and makes differences reflect the available language and geometric evidence.

```
You are an information-extraction system.
Determine whether {object} is in the speaker's egocentric
{relation} direction. Answer yes only when the situation
explicitly or unambiguously establishes this relation. The
speaker's orientation, view, or movement may establish it
without the literal word "{relation}". Do not infer it from
co-occurrence, world knowledge, another relation, or a phrase
referring to another object.

Return exactly one JSON object, with no markdown or explanation:
{
 "answer":"yes or no",
 "evidence":"exact quote from the situation or null"
}

For yes, evidence must be one contiguous, comma-free exact quote
that names {object} and supports {relation}. For no, evidence
must be null.
```

Clock-direction constraints and validation. We query clock-face relations only if a situation contains an explicit clock-direction expression:

```
Read the situation and determine the current object's clock
direction. Return exactly one JSON object, with no markdown or
explanation:
{"answer":"no or a clock hour from 1 through 12",
 "evidence":"exact quote from the situation or null"}

Return a clock hour only when one contiguous, comma-free exact
quote explicitly names the current object and expresses that same
clock hour. Otherwise answer no with null evidence.

Current object: {object}
```

Malformed JSON, schema violations, invalid relation labels, or invalid evidence trigger a retry and request a complete answer again. We accept a final relation only after JSON parsing, relation-label validation, and object grounding. These checks avoid silently passing malformed or unsupported constraints to the pose scorer.

B.2 3D Object Localization

To evaluate candidate observer poses against the extracted spatial relations, we need to locate the referenced objects in 3D. For each scene, we collect the object phrases extracted from its situation descriptions (Section B.1) and use them as text prompts for SAM3 (Carion et al., 2026) on RGB-D observations. For each phrase, SAM3 produces 2D instance masks. We back-project each accepted mask using the RGB-aligned depth map and camera intrinsics, then transform the resulting points into the world frame using the camera-to-world pose $\mathbf{T}_t$ from Section 3.

To consolidate detections across views, we voxelize these points at 5 cm resolution and merge observations associated with the same object phrase when their 3D intersection-over-minimum is at least 0.10. This produces a scene-level collection of localized object instances, each represented by its voxel set, bounding box, centroid, segmentation confidence, and multi-view support. These instances provide the 3D object references for candidate pose scoring in the next section.

B.3 Candidate Scoring and Training

Candidate scoring. Candidate observer poses $\mathbf{z}$ are constructed by sampling a 25 cm spatial grid in the XY plane over the scene bounds, paired with 24 discretized yaw orientations spaced at 15° intervals. For each extracted spatial relation r, we identify compatible localized object instances whose object names match the object mentioned in r. Let i index these instances. For example, the relation "the chair is to my left" is evaluated against each localized chair instance. Relations without compatible instances remain unmatched.

For each candidate pose $\mathbf{z}$, relation r, and compatible object instance i, we compute a spatial feature vector $\mathbf{f}_{r,i}(\mathbf{z})$. This vector includes the distance between the candidate observer position and the object, independently of heading. For coarse and fine-grained directional relations, this vector additionally includes forward and lateral offsets measured in the candidate observer's coordinate frame and thus depends on the candidate heading. We encode these spatial features together with instance information and the relation type with a MLP:

$$\mathbf{h}_{r,i}(\mathbf{z}) = \mathrm{MLP}_{\mathrm{enc}}([\mathbf{f}_{r,i}(\mathbf{z}),\, \mathbf{g}_i,\, \mathbf{e}_r])\,, \tag{A.1}$$

where $[\cdot]$ denotes concatenation, $\mathbf{g}_i$ contains the object's geometry, segmentation confidence, and multi-view support, and $\mathbf{e}_r$ is the embedding of the relation type specified by r. The output $\mathbf{h}_{r,i}(\mathbf{z})$ encodes the information used to assess whether instance i satisfies relation r under candidate pose $\mathbf{z}$.

We aggregate these representations in two steps. First, for each relation, we apply normalized log-sum-exp pooling over its compatible object instances. We then apply mean, maximum, and minimum pooling across relations and concatenate the three outputs into a single vector. Denoting this two-step aggregation by Pool, a final MLP produces the candidate score:

$$s(\mathbf{z}) = \mathrm{MLP}_{\mathrm{score}}\big(\mathrm{Pool}\big(\{\mathbf{h}_{r,i}(\mathbf{z})\}_{r,i}\big)\big)\,. \tag{A.2}$$

This scalar score measures how well the candidate observer pose agrees with the situation description, with higher values indicating stronger agreement.

Training. We train the candidate scorer on the SQA3D training split, using the ground-truth observer position and heading to supervise the candidate scores $s(\mathbf{z})$. To reduce the cost of scoring the full spatial grid during training, we sample a subset of its positions at each optimization step. We prioritize positions within 0.5 m of the ground truth and fill the remaining sampling budget with random positions from the scene grid. Each sampled position is paired with all 24 candidate headings defined above.

For these candidate poses, we convert the scores into a predicted probability distribution $p(x, y, \theta)$ by applying a softmax over the candidates, where (x, y) denotes a grid position and θ denotes a candidate heading. The supervision also takes the form of a probability distribution: candidates closer to the ground-truth pose receive larger target probabilities. Specifically, we construct a Gaussian target $q_{xy}(x, y)$ over the sampled positions and a Gaussian target $q_{\text{yaw}}(\theta)$ over the candidate headings, centered on the ground-truth position and heading, respectively. These are soft targets because they distribute probability across nearby candidates instead of assigning all probability to a single candidate.

We normalize each target over its candidate set and define the target probability of a complete pose as their product:

$$q(x, y, \theta) = q_{xy}(x, y)q_{\text{yaw}}(\theta). \tag{A.3}$$

We supervise both the complete pose and its position and heading components. The joint loss $\mathcal{L}_{\text{joint}}$ is the cross-entropy between the target distribution q and the prediction p. For position supervision, we sum $p(x, y, \theta)$ over headings to obtain the predicted position distribution and compute its cross-entropy with q_{xy}, denoted by $\mathcal{L}_{xy}$. Similarly, summing $p(x, y, \theta)$ over positions gives the predicted heading distribution, whose cross-entropy with q_{yaw} is $\mathcal{L}_{\text{yaw}}$. The base training objective combines these three losses:

$$\mathcal{L}_{\text{base}} = \mathcal{L}_{\text{joint}} + \lambda_{xy}\mathcal{L}_{xy} + \lambda_{\text{yaw}}\mathcal{L}_{\text{yaw}}, \tag{A.4}$$

where λ_{xy} and λ_{yaw} weight the position and heading losses. We apply early stopping when the fraction of SQA3D validation examples, whose highest-scoring pose has both a position error below 0.5 m and a yaw error below 15°, stops improving.

B.4 Inference and Evaluation

At inference time on both SQA3D and Real3DQA, the pose estimator evaluates all cells from the candidate grid and normalizes the logits across all candidates. We then sort the candidates according to their scores and apply non-maximum suppression. This suppresses redundant hypotheses from neighboring cells or multiple heading bins at the same location, ensuring that the Top-K cells represent distinct candidate poses. The highest-ranked candidate is used as the input state for our reasoning model. If the extracted constraints do not match any instance in the object bank, the estimator falls back to a randomly sampled candidate pose.

We evaluate each prediction using the Euclidean position error in the world XY plane and the minimum circular difference between the predicted and ground- truth yaw angles. We report position accuracy at 0.5 and 1.0 m, yaw accuracy at 15° and 30°, Top-1 joint success under the 0.5 m/15° criterion, and mean position and yaw errors. For Top-K evaluation, we report the fraction of samples for which at least one of the K highest-ranked candidates simultaneously satisfies the position and yaw thresholds. Both errors must be computed for this same candidate; combining the position estimate from one candidate with the yaw estimate from another does not count as a success. This avoids artificially combining the best position estimate from one candidate with the best orientation estimate from another.

Table A.3 shows the localization and orientation accuracy of our estimator on the SQA3D test set with prior methods. Our Top-1 pose setting achieves localization accuracies of 40.12% and 62.75% at 0.5 m and 1.0 m, respectively, and orientation accuracies of 32.57% and 50.74% at 15° and 30°. For the Top-3 oracle setting, we first retain the three highest-ranked candidates according to the scorer, without using the ground-truth pose. We then select the candidate closest to the ground-truth pose from these three predictions for evaluation. This Top-3 oracle setting achieves 55.33%,

Table A.3: Situation-pose estimation results on the SQA3D test set. All values are percentages. The Top-3 oracle selects the pose closest to the ground-truth pose among the three highest-scoring hypotheses.

Method	Localization		Orientation	
	Acc@0.5 m	Acc@1.0 m	Acc@15°	Acc@30°
Random	7.2	25.8	8.4	16.9
SQA3D (Ma et al., 2023)	9.5	29.6	8.7	16.5
SQA3D (separate) (Ma et al., 2023)	10.3	31.4	17.1	22.8
3D-VisTA (Zhu et al., 2023)	11.7	34.5	16.9	24.2
SIG3D+ (Man et al., 2024)	16.8	35.2	23.4	26.3
SIG3D (Man et al., 2024)	27.4	59.1	28.7	42.5
View2Cap (Yuan et al., 2025)	17.4	36.9	24.1	28.5
Loc3R-VLM (Qu et al., 2026)	42.6	75.9	38.4	63.0
Ours (Top-1)	40.12	62.75	32.57	50.74
Ours (Top-3 oracle)	55.33	77.18	48.06	67.40

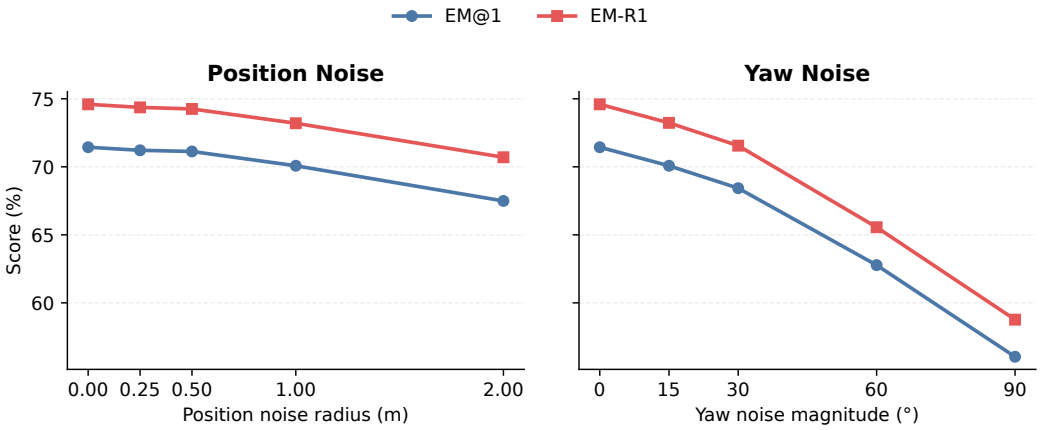

Figure A.1: Sensitivity to perturbations of the ground-truth situation pose on the SQA3D test set. Position and yaw noise are varied independently; the unperturbed setting corresponds to the GT-pose result.

77.18%, 48.06%, and 67.40% on the four metrics. Note that the candidate generation and ranking only rely on the pose estimator.

C Sensitivity to Situation-Pose Perturbations

Our experimental results show a substantial performance gap between the GT-pose setting and the other two settings. To validate the robustness to situation pose estimation, we perturb the ground-truth situation pose at inference time to isolate the reasoning model's sensitivity to position and heading errors. Figure A.1 shows that the performance is relatively stable under moderate translation: for 0.5 m position noise, EM@1 and EM@R1 both decrease by only 0.3 points, and when increasing the noise to 2.0 m, the performance further drops by 3.9 points. Heading noise has a larger effect. A 30° yaw perturbation reduces EM@1 and EM@R1 by 2.9 and 3.1 points, respectively, while a 90° perturbation causes drops of 15.4 and 15.8 points. This asymmetry is expected because yaw directly rotates the observer-centric coordinate frame, leading to an incorrect heading when predicting the answer.

Table A.4: BEV cell-size ablation on SQA3D with GT situation poses (EM@1, %). All rows use OC-RoPE with learnable observer anchors. Bold indicates the best score in each column at the reported precision.

Cell size (m)	What	Is	How	Can	Which	Others	All
0.05	62.6	78.7	61.3	**78.4**	76.4	**72.6**	69.9
0.10 (default)	**63.1**	**81.6**	**66.2**	**78.4**	**78.6**	72.3	**71.4**
0.20	61.9	76.7	62.6	77.2	75.8	70.8	69.0

D Sensitivity to BEV Cell Size

We show the impact of different BEV cell sizes ($0.10\,\mathrm{m}$, $0.05\,\mathrm{m}$, and $0.20\,\mathrm{m}$) on SQA3D. Table A.4 reports per-question-type and the overall EM@1 for the full model with observer anchors under the GT-pose setting. The default setting achieves the highest overall EM@1 of 71.4%, exceeding the finer and coarser grids by 1.5 and 2.4 percentage points, respectively. Meanwhile, this setting ($c = 0.10\,\mathrm{m}$) also achieves the highest scores on *What, Is, How, and Which* questions. The finer grid cell setting matches the default setting on *Can* at the reported precision (78.4%) and performs slightly better on Others (72.6% versus 72.3%).

E Attention Visualization

We further visualize the attention patterns of the LoRA baseline with the other variants incorporating OC-RoPE and the learnable observer anchor. We construct three separate attention maps for the attention between visual tokens and the situation description, the question, and the answer, respectively. For each map, we assign each visual token a score by averaging the attention it receives from the corresponding text tokens at the model's last four layers. We then normalize these scores across all visual tokens within each attention map and project them onto the reconstructed scene. In each visualization figure, we compare the attention map for situation texts, question texts, and answer texts across the LoRA baseline (the left), the OC-RoPE variant (the middle), and the full model (the right). Warmer colors indicate higher normalized attention scores.

Figure A.2 shows a representative example, with **pink arrows** marking the situation position and orientation and **red boxes** highlighting answer-relevant regions. In the first row, the variants with geometric encoding (OC-RoPE and the full model) concentrate situation-token attention more strongly around the situation pose than the baseline. The last row shows a similar pattern for answer tokens, where our method focuses attention more tightly on answer-relevant regions, whereas the baseline distributes attention broadly across the scene. These qualitative observations are consistent with the distance-mask results, suggesting that the proposed components strengthen the association between situation information and visual tokens and help focus answer generation on relevant visual tokens. Five additional examples are provided in Figures A.3–A.7.

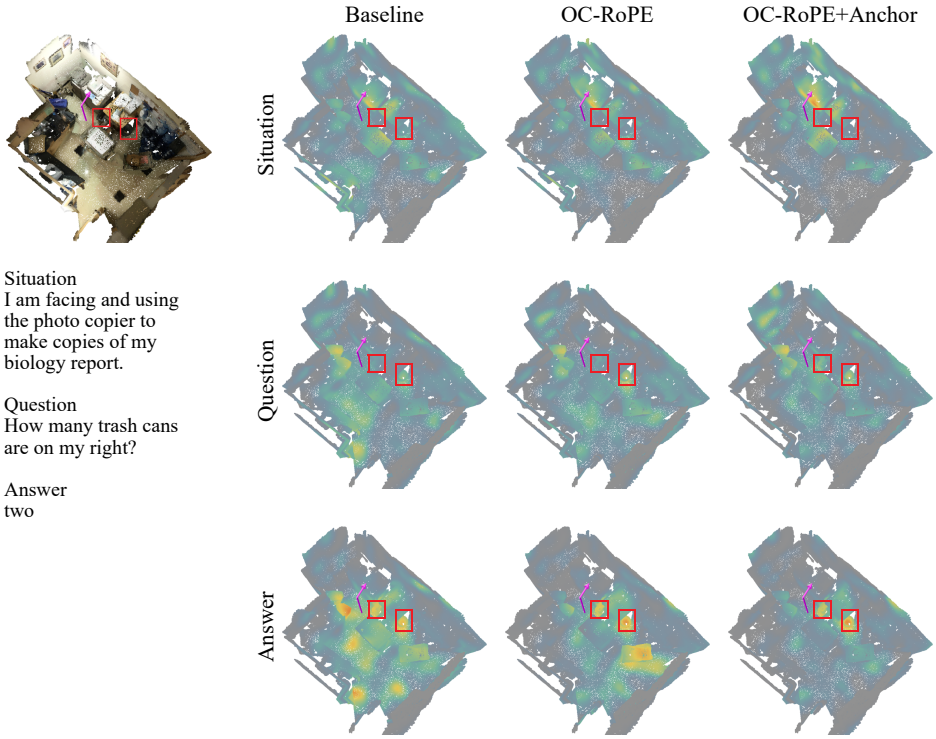

Figure A.2: Attention visualization on a representative SQA3D example.

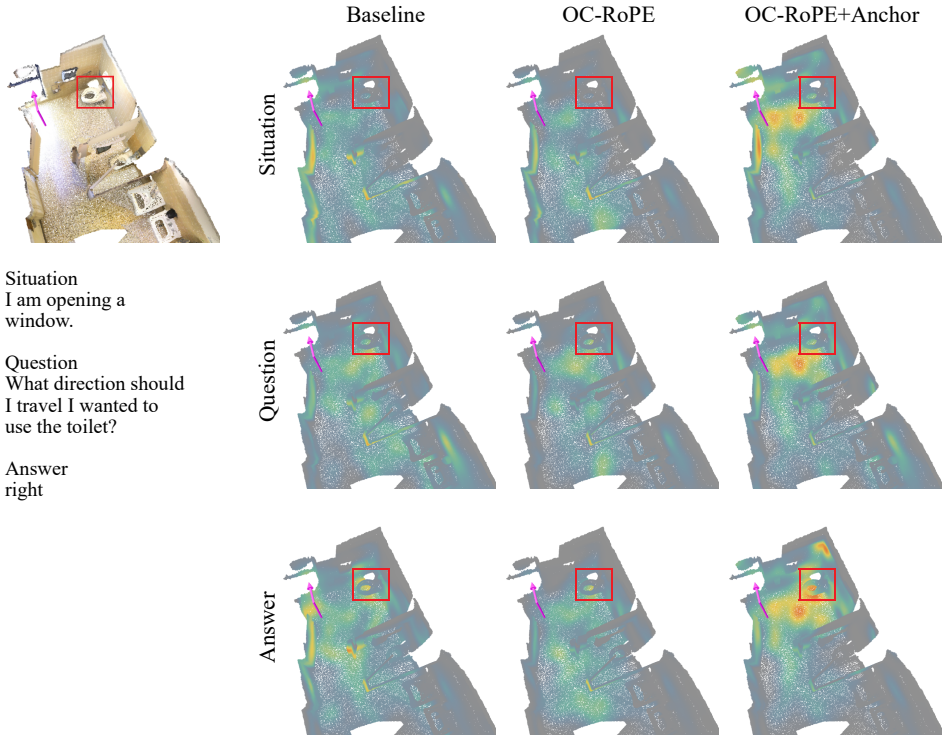

Figure A.3: Additional attention visualization on SQA3D (example 1).

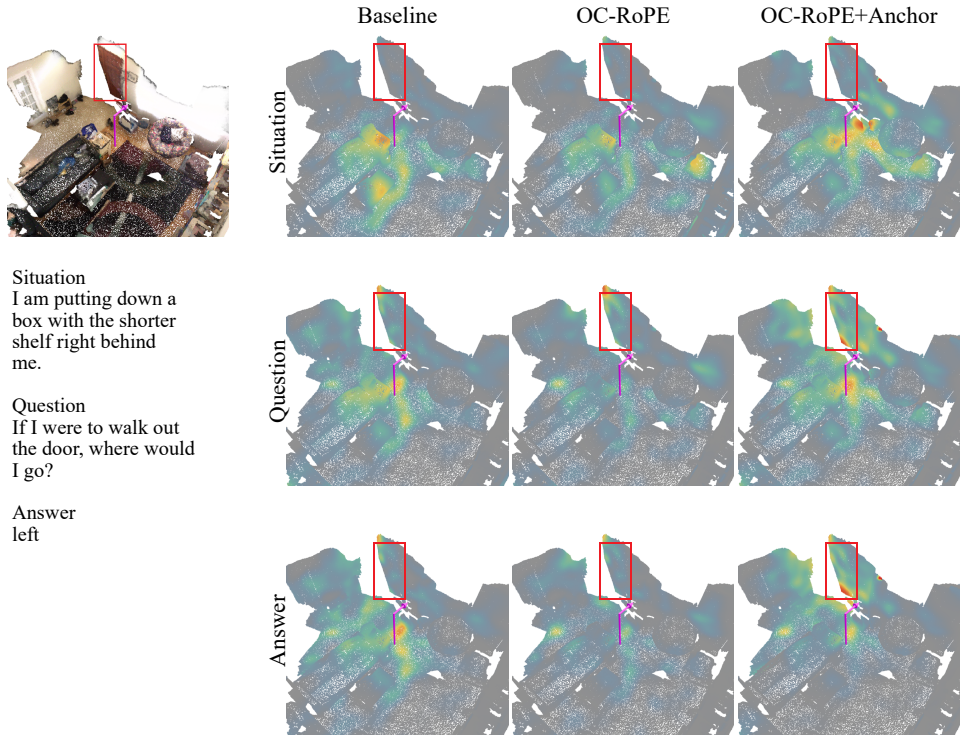

Figure A.4: Additional attention visualization on SQA3D (example 2).

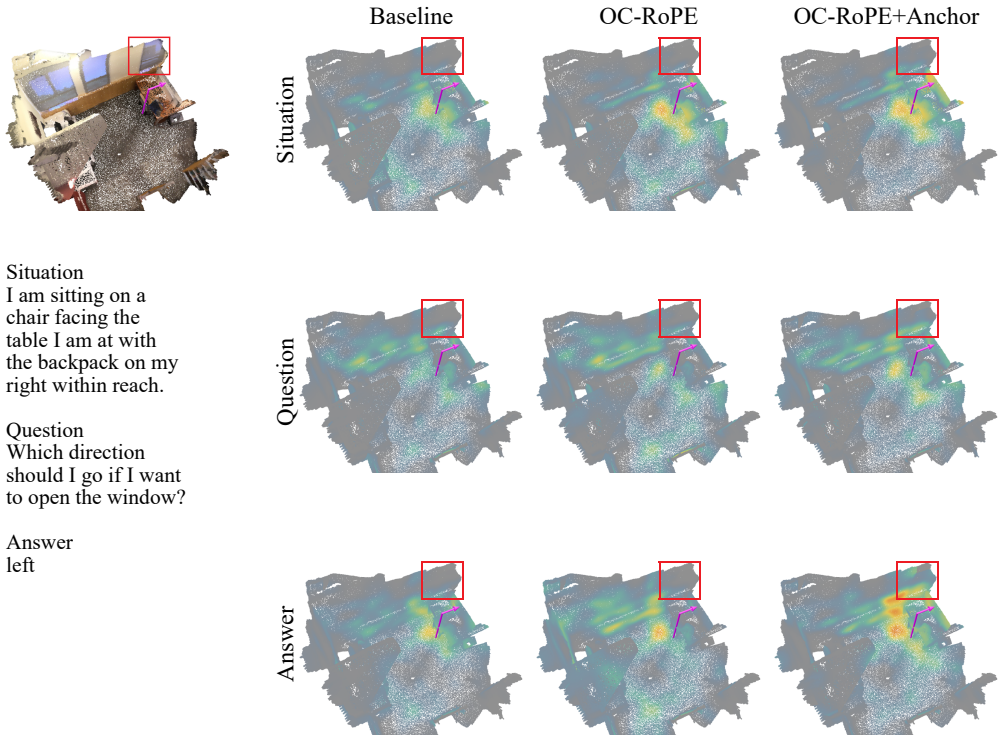

Figure A.5: Additional attention visualization on SQA3D (example 3).

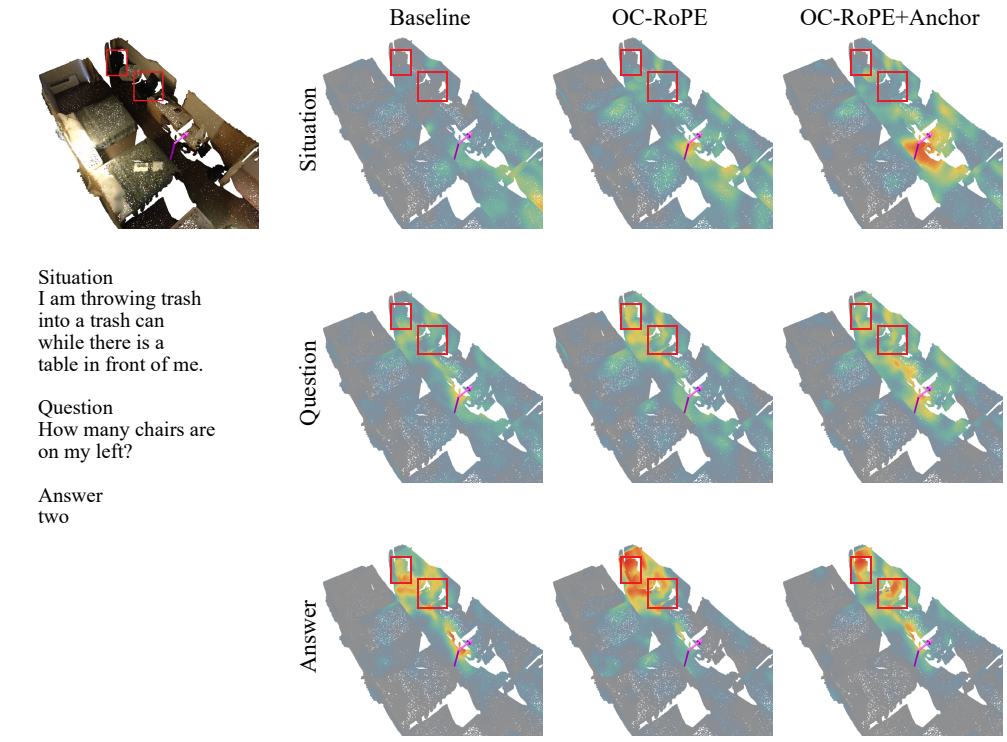

Figure A.6: Additional attention visualization on SQA3D (example 4).

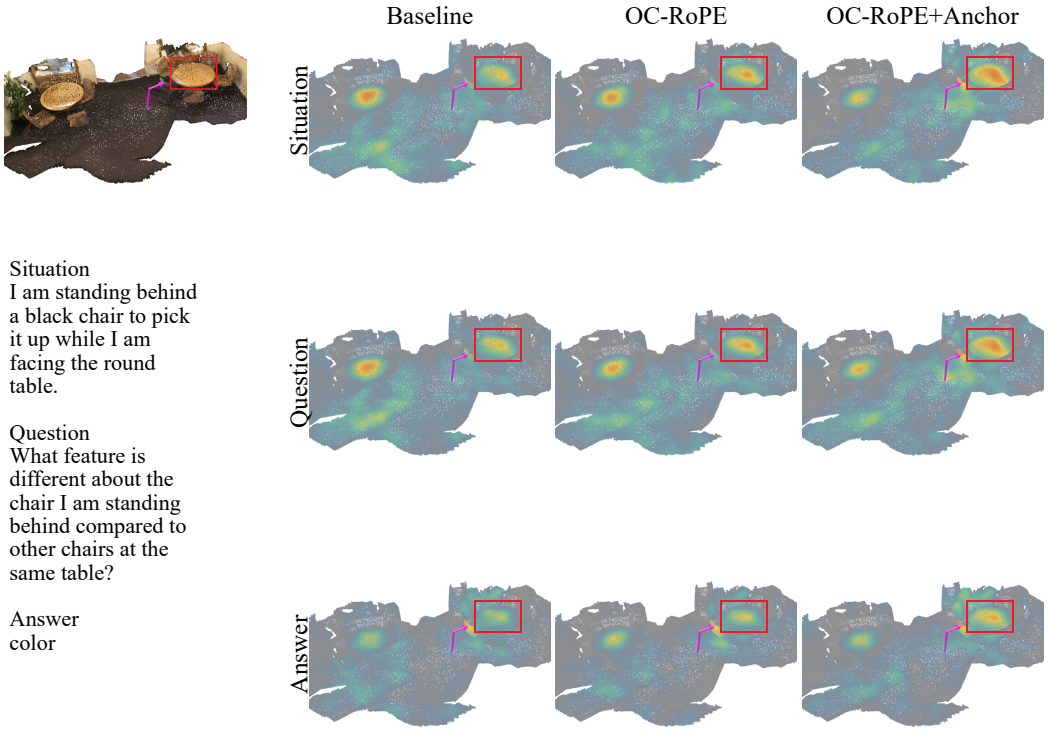

Figure A.7: Additional attention visualization on SQA3D (example 5).