

Situational Awareness Matters in 3D Vision Language Reasoning

Supplementary Material

A. Implementation Details

Here we provide more details about our model. The code will be publicly released.

Visual and Textual Encoders. We use OpenScene [51] (the 3D Distilled variant) as our visual encoder, which incorporates distilled CLIP features into a 3D Minkowski CNN backbone originally designed for 3D semantic segmentation task. We use the default 0.02m voxel size to discretize the point cloud into 3D voxels and disable the scaling and elastic distortion augmentation methods during the voxelization process. The 3D architecture is the predefined *MinkUNet18A* [12]. The number of visual tokens N_v is 256. Additionally, we use Sentence-BERT [55] MPNet variant as our text tokenizer and encoder. We use the fixed batch padding strategy and set the length to be 100 for both situation tokens N_s and question tokens N_q . The feature embedding sizes for all three types of tokens are set to 768. The 256 dimensional output of OpenScene backbone is projected to the 768 hidden size with a 1x1 Conv layer. We freeze the OpenScene backbone, and finetune only the last layer of the textual backbone during our training process.

Fusion and Decoder Models. We use a 4-layer MCAN Transformer [59, 67] as the fusion block of the visual and situational tokens. The learnable positional embeddings and situational embeddings are composed of a 2-layer MLP: first from dimension 3 to 128, and then from 128 to the target dimension. We use BLIP-2 [38] as the large multi-modal transformer for final response generation, similar to 3D-LLM [26].

Training Details. We train our model with AdamW [43] optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 1e-8$. We use batch size 16 and set initial learning rate to be $2e-5$. Weight decay is set to 0.05, and we disable the weight decay on the layernorm layers and all bias parameters following [40]. We decrease the learning rate by 10 times after the 10-th and the 20-th epoch. We train the model for a total 50 epochs on a single NVIDIA A100 GPU.

B. Enhanced Vision Token Activation through Situational Re-encoding

In Figure B, we provide an insightful visualization of the activation changes in 3D visual tokens x^{3D} , before and after undergoing our situational-guided visual re-encoding process. This visualization employs the *viridis* colormap, where a brighter token representation indicates a higher activation value. The effectiveness of situational guidance in amplifying the relevance of crucial tokens is evident from

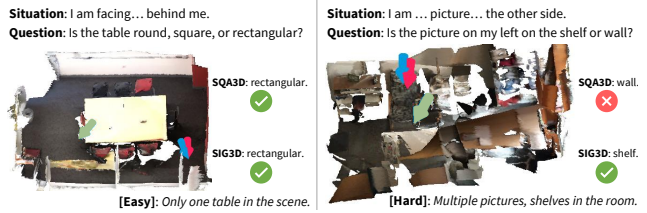


Figure A. Visualization of cases where situation prediction benefits the most. Arrow colors: GT, SIG3D, and SQA3D predictions.

this depiction.

For instance, the visualization in the second row reveals a notable shift in focus. Initially, the tokens predominantly concentrate on the bed area. However, post re-encoding, there is a discernible shift in attention towards areas closely aligned with the situational vector and those directly related to the query. Similarly, in the third row, the situational re-encoding process results in the window region “on the left” receiving increased emphasis. And in the fourth row, the attention initially focuses on the vanity region. Then it shifts to the toilet on the left of the agent, as suggested by the situation vector and the question prompt. This experiment provides a clear demonstration of how our method, by harnessing enhanced situational awareness, contributes to improved performance in downstream reasoning tasks with an explainable manner. The ability of our model to dynamically adjust focus in response to situational cues is a key factor in its enhanced reasoning capabilities.

C. More Qualitative Results

We show more qualitative results of our model in Figures C and D. The visualization encompasses a diverse array of tasks, including queries about object orientation, characteristics of specific objects, the count of objects within a scene, and yes/no questions based on commonsense reasoning. A key observation from these results is that in numerous instances, absolute precision in situational estimation is not a prerequisite for our model to accurately deduce the answers to the posed questions. This finding highlights the model’s robustness and its capacity to effectively handle a variety of query types, even with less optimal situational awareness.

D. Performance on Hard Cases

An example of our case study on easy-hard samples is shown in Fig. A. We find that simple examples in the dataset allow existing models to guess the correct answer without any 3D situational understanding. However, our method effectively improve the hard examples with complicated and

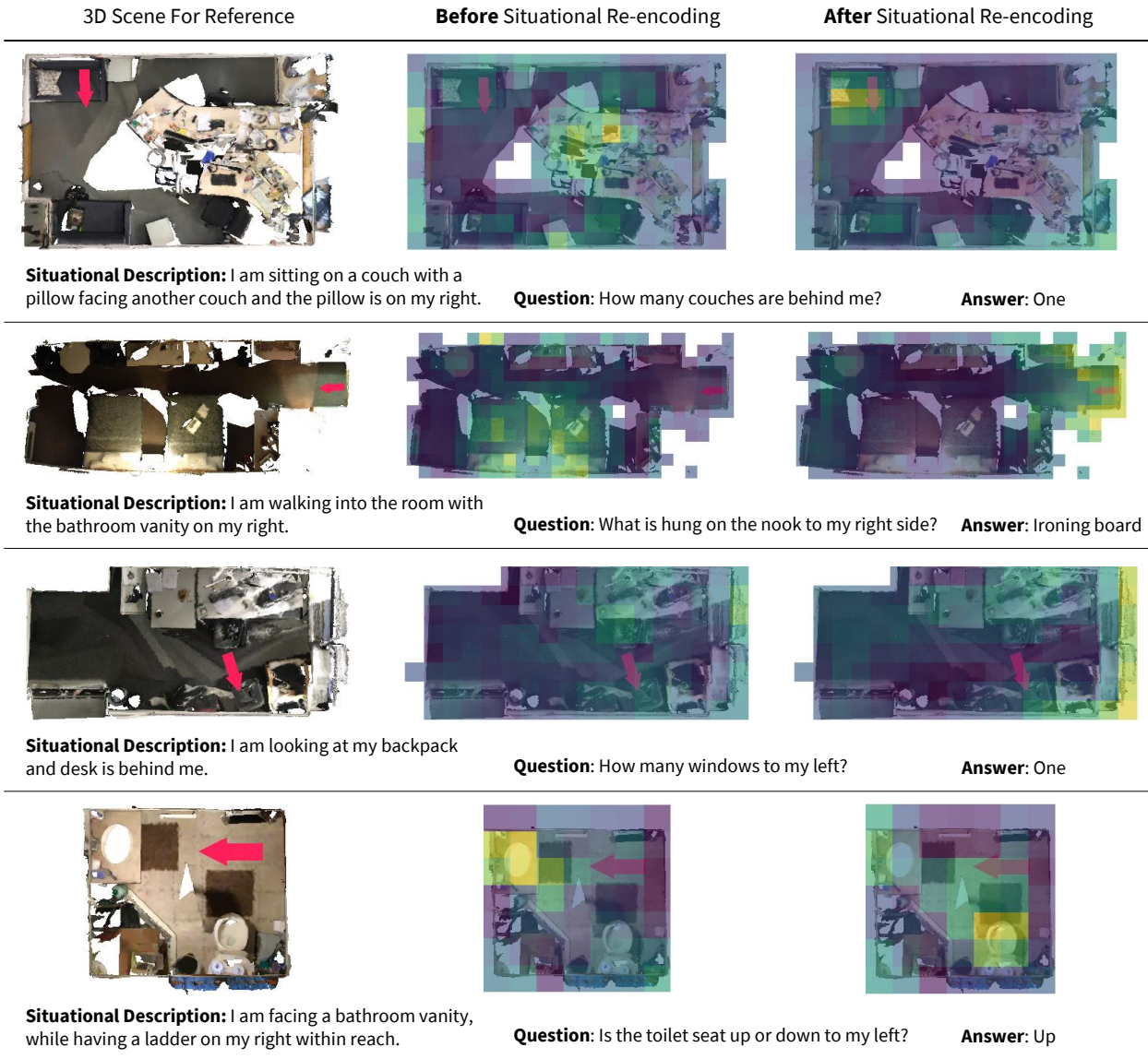


Figure B. **3D vision token activation before and after situational re-encoding.** We can notice that higher weights are assigned to question and situation-related tokens after our proposed situational re-encoding mechanism.

entangled questions and situations.

E. Analysis of Failure Cases

We conduct a failure case analysis on our model in Figure E. We categorize and visualize three types of failure cases.

Accurate Situation Estimation, Incorrect Question Answering. This scenario demonstrates that accurate situational understanding does not necessarily guarantee correct responses to queries. A significant proportion of failures within this category can be attributed to complex question prompts that demand multi-stage reasoning or the integration of commonsense knowledge. For instance, the initial

example necessitates the model’s comprehension of the spatial relationship between the viewer’s perspective and the couch, followed by an additional reasoning phase focused on the couch to accurately respond to the query. The subsequent example demands an understanding of the concepts of “odd” and “even”, and their application to the count of objects in a 3D environment.

Inaccurate Situation Estimation, Correct Question Answering. This category reveals that errors in situational estimation are more likely when the scene description involves minor or less common objects. Furthermore, it is observed that the model might incidentally arrive at the correct answer without fully grasping the complex situational and

















<p>Situational Description: I am sitting on the left cushion of the couch and to my right is a pillow.</p> <p>Question: Which direction should I go if I want to open the curtain?</p>  <p>Answer: Left</p>	<p>Situational Description: I just walked into the room through the doors.</p> <p>Question: How many armchairs are directly in front of me?</p>  <p>Answer: Zero</p>	<p>Situational Description: I am standing in front of the door and facing the file cabinet.</p> <p>Question: Is the door behind me open or closed?</p>  <p>Answer: Closed</p>	<p>Situational Description: I am picking up my jacket on the chair while facing the blackboard and there is a desk with no monitor in my six o'clock direction.</p> <p>Question: What color is the desk behind me?</p>  <p>Answer: Brown</p>
<p>Situational Description: I am fixing the cabinet with a few bags by my right foot.</p> <p>Question: If I want to take a break and take a nap, is there a bed I could sleep on?</p>  <p>Answer: Yes</p>	<p>Situational Description: I am taking out the white ball from the side pocket while my beer on the table is being drunk by me. I am also facing a white board.</p> <p>Question: Can I see chair if I turn around?</p>  <p>Answer: Right</p>	<p>Situational Description: I am facing a chair and there is a printer on my right.</p> <p>Question: If I turned directly around and walked straight, what would I hit first?</p>  <p>Answer: Bed</p>	<p>Situational Description: I am looking at mirror and combing my hair with a hairbrush that was in the bathroom vanity.</p> <p>Question: Is the door to my right open or closed?</p>  <p>Answer: Wall</p>
<p>Situational Description: I am standing in the middle of the kitchen and the stove is on my left.</p> <p>Question: What is to my left that can be used to cook food?</p>  <p>Answer: Stove</p>	<p>Situational Description: I am opening the door.</p> <p>Question: Can I see a window if I turn my head rightwards?</p>  <p>Answer: Yes</p>	<p>Situational Description: I am sitting on the toilet facing a bathtub.</p> <p>Question: Does the bathtub in front of me have a shower curtain?</p>  <p>Answer: No</p>	<p>Situational Description: I am turning on the lamp by the chair with the chair on my right within reach.</p> <p>Question: If I look to my right, can I see my reflection?</p>  <p>Answer: Wall</p>
<p>Situational Description: I am leaning on the door facing the blinds.</p> <p>Question: Where should I walk if I want to find a book to read during my break?</p>  <p>Answer: Forward</p>	<p>Situational Description: I am sitting on the toilet facing a bathtub.</p> <p>Question: Can I reach the sink from where I am sitting?</p>  <p>Answer: Right</p>	<p>Situational Description: I am facing the shelf closest to the armchair.</p> <p>Question: What color is the ledge in front of me?</p>  <p>Answer: White</p>	<p>Situational Description: I am opening the door.</p> <p>Question: Are there pillows on the couches?</p>  <p>Answer: Wall</p>

Figure C. We demonstrate more successful examples of our method.





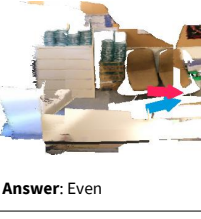






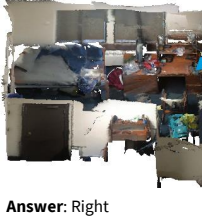

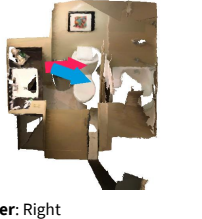
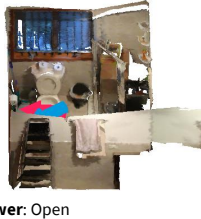
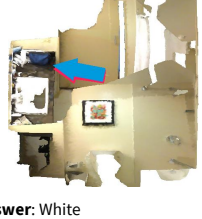
<p>Situational Description: I am facing a table, while there is a chair on my right and two chairs behind me.</p> <p>Question: Is the room clean or messy?</p>  <p>Answer: Messy</p>	<p>Situational Description: I am picking up my backpack with a chair to my right within reach.</p> <p>Question: How many chairs are behind me?</p>  <p>Answer: Two</p>	<p>Situational Description: I am looking for a cold drink to have and there is a cabinet on my right side.</p> <p>Question: What is on top of the cabinet that is on my 1 o'clock?</p>  <p>Answer: TV</p>	<p>Situational Description: I am facing the toilet, and the sink is behind me.</p> <p>Question: Is the amount of bar I am facing odd or even?</p>  <p>Answer: Odd</p>
<p>Situational Description: I am standing by the paper cutter on my right within reach.</p> <p>Question: Is the number of chairs on my left odd or even?</p>  <p>Answer: Even</p>	<p>Situational Description: I am facing the whiteboard, and there are some chairs and a table behind me.</p> <p>Question: What is to my left that you can use to see the outside?</p>  <p>Answer: Windows</p>	<p>Situational Description: I am throwing trash with a chair very close to me behind me.</p> <p>Question: Is the shape of table behind me round, square, or rectangular?</p>  <p>Answer: Rectangular</p>	<p>Situational Description: I am standing right in front of the sink, and I can see table across the room.</p> <p>Question: Can I see the TV without moving much?</p>  <p>Answer: No</p>
<p>Situational Description: I am facing the wall with large windows, and there is a trashcan behind me.</p> <p>Question: How many chairs are accounted for in the room total?</p>  <p>Answer: Eight</p>	<p>Situational Description: I am opening one door while there is another one adjacent on my right.</p> <p>Question: Is the door in front of me closed or open?</p>  <p>Answer: Closed</p>	<p>Situational Description: I am standing in front of a counter, and there is a recycling bin to my left within reach.</p> <p>Question: Which direction should I go if I want to open the mailbox?</p>  <p>Answer: Right</p>	<p>Situational Description: I just entered the room, while having a desk with a shelf on top of it at my right and a door behind me.</p> <p>Question: Where do I go to get to the other bed?</p>  <p>Answer: Right</p>
<p>Situational Description: I am placing garbage into a trash can, while having a toilet to my left within reach.</p> <p>Question: What is the color of the towel behind me?</p>  <p>Answer: White</p>	<p>Situational Description: I am facing the toilet, and the sink is behind me.</p> <p>Question: Which direction should I go if I want to exit the room?</p>  <p>Answer: Right</p>	<p>Situational Description: I am opening the door.</p> <p>Question: Does the toilet to my right have an open or closed lid?</p>  <p>Answer: Open</p>	<p>Situational Description: I am looking at mirror and combing my hair with a hairbrush that was in the bathroom vanity.</p> <p>Question: What color is the sink in front of me?</p>  <p>Answer: White</p>

Figure D. In addition to Figure C, we demonstrate more successful examples of our method.





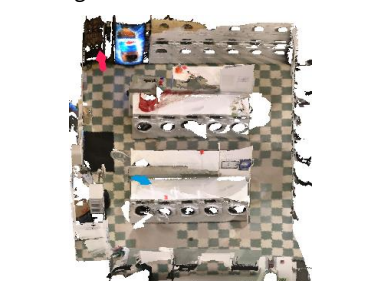

(a) Situational estimation is correct, and QA is wrong.	(b) Situational estimation is wrong, but QA is correct.	(c) Situational estimation is wrong, and QA is wrong.
<p>Situational Description: After an exhausted day, I am lying on the bed closest to the window with my head on the pillows.</p> <p>Question: What is to the right of the couch in front of me?</p>  <p>GT Answer: Desk ✗ Pred Answer: End table ✗</p>	<p>Situational Description: I am standing in front of the door and facing the file cabinet.</p> <p>Question: Is the door behind me open or closed?</p>  <p>GT Answer: Closed ✔ Pred Answer: Closed ✔</p>	<p>Situational Description: I am standing by backpack on my left side and the bed is behind me while the curtain is in my six o'clock direction.</p> <p>Question: Are there more doors to my left than there are to my right?</p>  <p>GT Answer: No ✗ Pred Answer: Yes ✗</p>
<p>Situational Description: Standing by the table with a tray rack to my immediate right.</p> <p>Question: Is the amount of chair on my left side odd or even?</p>  <p>GT Answer: Odd ✗ Pred Answer: Even ✗</p>	<p>Situational Description: I am facing a vending machine that is next to another one, while there is a trash can behind me directly.</p> <p>Question: Can I see clothes where I am standing?</p>  <p>GT Answer: Yes ✔ Pred Answer: Yes ✔</p>	<p>Situational Description: I am sitting on a chair facing the table with the blackboard behind me and a chair on my left within reach.</p> <p>Question: Is the amount of table I am facing odd or even?</p>  <p>GT Answer: Odd ✗ Pred Answer: Even ✗</p>

Figure E. We demonstrate three different categories of failure cases.

multimodal context, particularly in cases where the question involves choosing between two or among multiple given options. Therefore, a blend of qualitative and quantitative assessments is crucial for a comprehensive evaluation of the model’s performance.

Both Situation Estimation and Question Answering are Incorrect. This group contains the most challenging examples from the dataset, typically encompassing multiple complexities identified in the preceding categories. These cases present a compounded difficulty level, highlighting the model’s limitations in scenarios that require an intricate understanding of both situational context and question in-

terpretation.

F. Limitations and Future Work

Selection of 3D Scenes. The SQA3D [44] and ScanQA [5] datasets, both derived from the ScanNet [13] dataset, exclusively feature indoor household environments. These static scenes limit the model’s applicability to dynamic tasks like manipulation and exploration. Consequently, our current model is tailored to static household settings. This scalability problem is a long-standing challenge for all existing 3D VL reasoning work [5, 26, 44, 74]. We believe that with a more scalable visual representation

(*e.g.*, scene graphs, sparse learnable embeddings), we can extend our model to support larger 3D environment in the future work.

More Comprehensive Visual Encoding. In our approach, the utilization of a voxel-based, open-vocabulary 3D encoder achieves much better overall performance. Nevertheless, for specific queries involving counting or referencing, a detection-based encoder may yield a more advantageous visual token set, owing to its capacity to provide instance-level information pertinent to the questions. This indicates the potential benefits of a multifaceted visual tokenization system that amalgamates the strengths of various encoder types.