Empowering Large Language Models with 3D Situation Awareness

Supplementary Material

1. Dataset Statistics

We present the detailed statistics of our dataset. The View2Cap dataset comprises 231,184 situation descriptions and 553,779 question answers, paired with corresponding viewpoint positions and rotations. Fig. 1(a) illustrates the word cloud of the situation descriptions, where the size of each word indicates its frequency in the dataset. Prominent words such as “located,” “cloud,” “point,” “table,” “wall,” and “light” appear in larger fonts, while less frequent words like “metal,” “lamp,” “shelf,” and “keyboard” are smaller. Many terms represent spatial descriptions or object placements, including “left,” “right,” “window,” and “front.” Fig. 1(b) displays the distribution of questions based on their prefixes, demonstrating the diversity of our questions, which encompass various inquiries about object attributes, quantities, and spatial relationships.

Fig. 2 presents the distribution of answers categorized by their types and contents. The answers are divided into four types: “Yes,” “No,” “Object,” and “Other.” Fig. 2(b) visualizes the distribution of answers in the “Other” category, showcasing a wide range of specific values such as colors and numbers. Fig. 2(c) displays the distribution of answers within the “Object” category, emphasizing the variety of object types mentioned, with notable contributions from common items such as “chair,” “table,” and “door.” These visualizations provide a comprehensive overview of the types of answers in the dataset and their respective distributions, reflecting the semantic diversity presented in the data.

Fig. 3 visualizes the diversity and frequency of answers corresponding to various question categories such as ”How,” ”Is,” ”Where,” ”What,” and ”Can.” Each stacked bar represents the range of unique responses, with more frequent answers prominently displayed at the base of the stack. Notable patterns include the dominance of simple affirmative (”Yes”) or negative (”No”) answers in certain categories (e.g., ”Is” and ”Can”) and the high variability in object-related responses under ”What” and ”Where.” This visualization underscores the semantic richness and heterogeneity of the dataset while highlighting common response tendencies for specific question types.

2. Dataset Refinement Process

We show the details of using human-annotated object labels to refine the View2Cap dataset. After generating image caption using Llava-onevision [1], we feed the original caption and ground truth 3D object labels from EmbodiedScan [3] to GPT4o and ask it to score the caption and refine it. In Fig. 4, we show the image and region point cloud on the left. On the right, we show the ground truth labels, the original caption, and the refined caption. It can be seen that our original caption can capture more fine-grained objects and detailed information than 3D labels, while refinement, can further improve the correctness.

In Fig. 5 and Fig. 6, we demonstrate a structured approach for evaluating and refining a generated caption against a given ground truth. The process includes a detailed scoring prompt, chain of thought analysis for identifying correctness and hallucinations, and a scoring rationale leading to a final score. The refinement prompt outlines specific instructions to correct hallucinated or omitted objects, ensuring alignment with the ground truth. The refined caption achieves a higher score by eliminating errors and improving descriptive accuracy.

3. Implementation Details

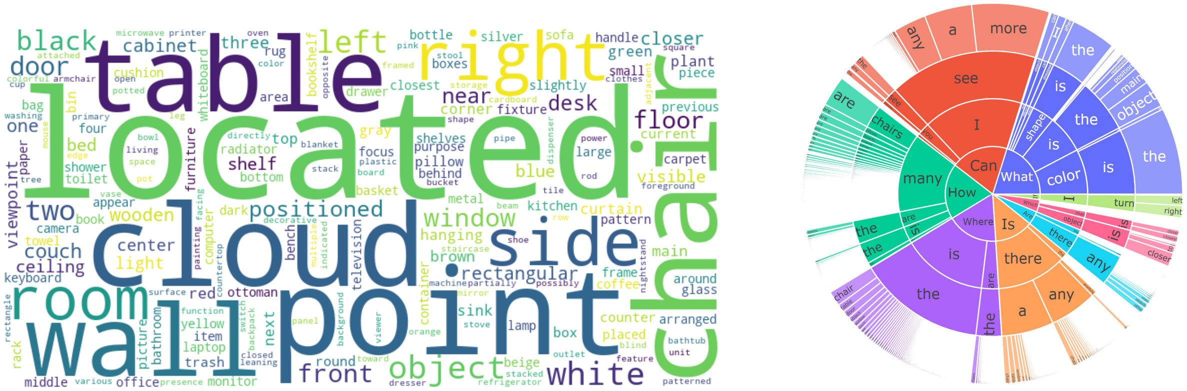
We set the maximum input token length and output new token length of Llama 3.1 7B [2] to 256. For each 3D scene, we sample up to 60 instances. For each instance, the maximum point number is 1024. During the training, the point cloud encoder is frozen, and the LLM is fine-tuned with Low-Rank Adaptation (LoRA) super-parameters rank and alpha 16. We add LoRA parameters on all linear layers. The learning rate is $2e^{-4}$, and the batch size is 8 on each GPU. We use 4 NVIDIA A100 GPUs with 80G memory in all experiments. During inference, we use the beam search strategy to generate text. The beam size is set to 5.

References

[1] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 1

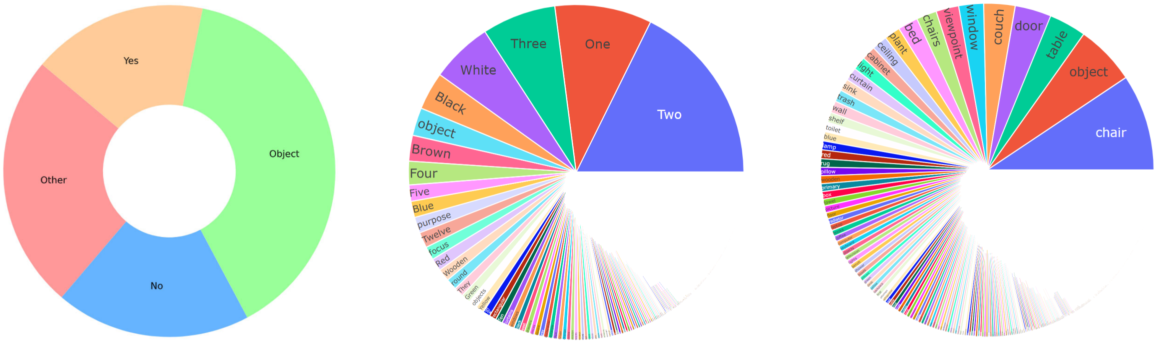
[2] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 1

[3] Tai Wang, Xiaohan Mao, Chenming Zhu, Runsen Xu, Ruiyuan Lyu, Peisen Li, Xiao Chen, Wenwei Zhang, Kai Chen, Tianfan Xue, et al. Embodiedscan: A holistic multimodal 3d perception suite towards embodied ai. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19757–19767, 2024. 1



(a) Word cloud of situation descriptions (b) Distribution of questions

Figure 1. Dataset statistics.



(a) Type of Answers (b) Answer distribution of type “Other” (c) Answer distribution of type “Object”

Figure 2. Answer distributions.

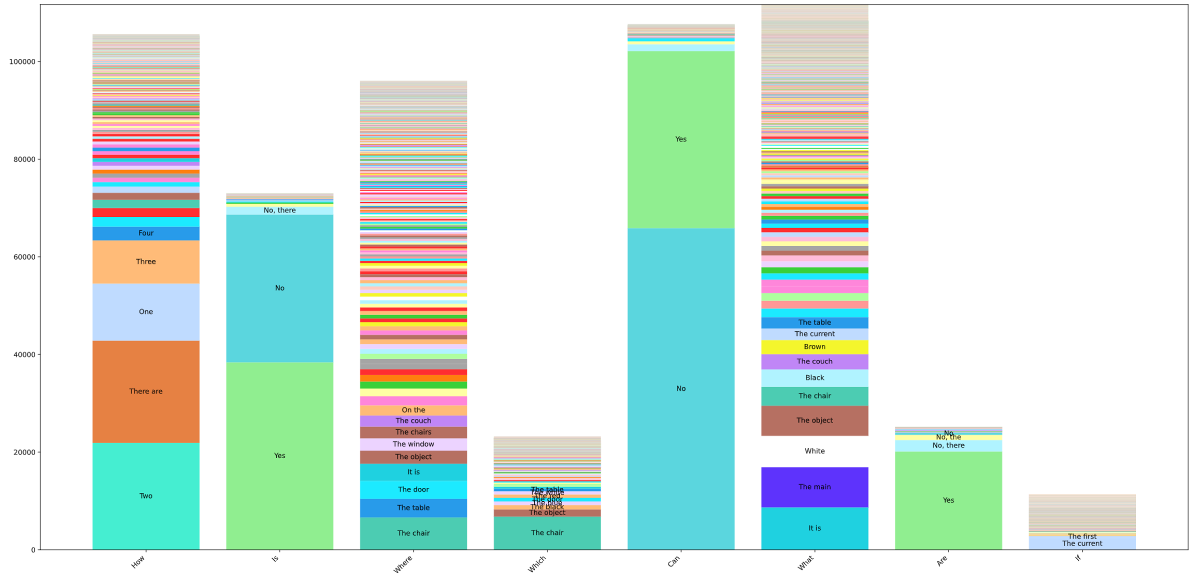
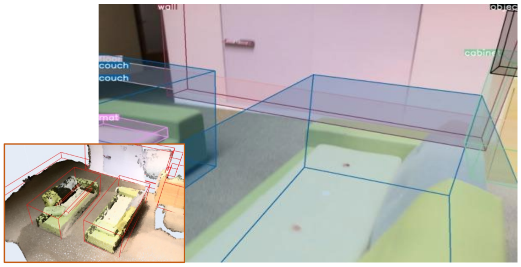


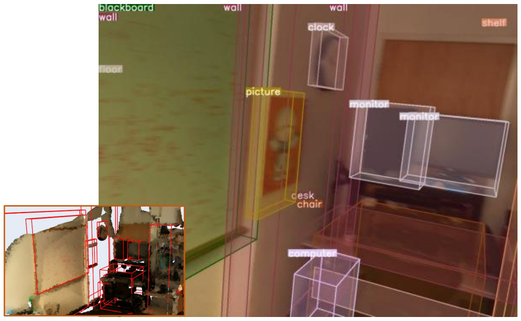
Figure 3. Answer distribution organized by question prefixes.



GT: *couch, mat, couch, cabinet*

Original: I am in a minimalist, modern room with a beige carpet. In the foreground, there's a low, padded bench in soft green and cream colors with small red dots. On the left, part of a similar bench is visible. In the background, a white wall has a small, rectangular object attached to it. To the right, there's a wooden piece of furniture, possibly a cabinet or table. The overall space is clean and uncluttered.

Refined: I am in a minimalist, modern room with a beige mat. In the foreground, there's a low, padded couch in soft green and cream colors with small red dots. On the left, part of a similar couch is visible. In the background, a white wall has a small, rectangular cabinet attached to it. The overall space is clean and uncluttered.



GT: *blackboard, picture, wall, computer, clock, monitor, chair, desk, monitor, shelf*

Original: I am in a room with a whiteboard on the left side and a wooden door on the right, with a clock hanging above the door. On the wall next to the clock is a poster of a blue character. In front of the door, there is a desk with a computer monitor, speakers, and various items on it.

Refined: I am in a room with a blackboard on the left side and a wall on the right, with a clock hanging above the wall. On the wall next to the clock is a picture. In front of the wall, there is a desk with a computer monitor.

Figure 4. Examples of our View2Cap captions.

Scoring Prompt:
Ground truth = blackboard, picture, wall, computer, clock, monitor, chair, desk, monitor, shelf.
Description = I am in a room with a whiteboard on the left side and a wooden door on the right, with a clock hanging above the door. On the wall next to the clock is a poster of a blue character. In front of the door, there is a desk with a computer monitor, speakers, and various items on it.
Based on the correctness and hallucination of objects, give a score of the description.

Chain of Thought :

1. Ground Truth Analysis:
   1. The ground truth list includes: **blackboard**, **picture**, **computer**, **clock**, **monitor**, **chair**, **desk**, **monitor**, and **shelf**.

2. Description Analysis:
   1. **Correct Matches** :
      1. **Clock**: The description mentions a **clock hanging above the door**, which is correctly identified as a **clock** in the ground truth.
      2. **Desk**: The description mentions a **desk**, which matches the ground truth.
      3. **Monitor**: The description mentions a **computer monitor**, which corresponds to one of the monitors in the ground truth.
      4. **Picture**: The **poster of a blue character** can be interpreted as a form of **picture**, which is similar to the ground truth **picture**.
      5. **Blackboard/Whiteboard**: The description mentions a **whiteboard**, which is similar to **blackboard**. (As previously considered, this is treated as a correct match).
   2. **Omissions**:
      1. The description does not mention the **computer**, **chair**, or **shelf**.
   3. **Hallucinations**:
      1. The description includes **speakers** and "various items," neither of which are part of the ground truth.
   4. **Incorrect Object**:
      1. The **door** is mentioned in the description, but it is not part of the ground truth list, and this is considered an error.

3. Scoring Rationale:
   1. There are **5 correct objects**, but the description introduces **1 incorrect object** (the door).
   2. There are also **3 omissions** and **2 hallucinations**.
   3. The errors are significant due to the **incorrect door**, but the overall description still gets many objects right.

Final Score: 2

Figure 5. Example of our View2Cap verification process.

Refine Prompt:
Ground truth = blackboard, picture, wall, computer, clock, monitor, chair, desk, monitor, shelf.
Description = I am in a room with a whiteboard on the left side and a wooden door on the right, with a clock hanging above the door. On the wall next to the clock is a poster of a blue character. In front of the door, there is a desk with a computer monitor, speakers, and various items on it.
Based on the ground truth and the provided image, refine the description with following requirements:
Please replace or delete all objects in description that does not appear in ground truth or the image. If object B in description is considered as a hallucination of object A in ground truth, use object A to replace object B. Otherwise, delete object B.
DO NOT delete the color or other adjectives of an object.

Refined Caption:
I am in a room with a blackboard on the left side and a wall on the right, with a clock hanging above the wall. On the wall next to the clock is a picture. In front of the wall, there is a desk with a computer monitor.

Chain of Thought:

1. Ground Truth Analysis:
  1. The ground truth list includes: **blackboard**, **picture**, **computer**, **clock**, **monitor**, **chair**, **desk**, **monitor**, and **shelf**.

2. Description Analysis:
  1. **Correct Matches**:
     1. **Blackboard**: The description mentions a **blackboard** on the left side, which correctly matches the ground truth **blackboard**.
     2. **Clock**: The description mentions a **clock hanging above the wall**, which matches the ground truth **clock**.
     3. **Picture**: The description mentions a **picture** next to the clock, which matches the ground truth **picture**.
     4. **Monitor**: The description mentions a **computer monitor**, which correctly corresponds to one of the monitors in the ground truth.
     5. **Desk**: The description mentions a **desk**, which matches the ground truth **desk**.
  2. **Omissions**:
     1. The description fails to mention the **computer**, **chair**, or **shelf**, which are part of the ground truth.

3. Scoring Rationale:
  1. There are **5 correct objects** and **3 omissions**.
  2. There are no hallucinated or incorrect objects that severely affect the score.

Final Score: 4

Figure 6. Example of our View2Cap refinement process.