ENHANCING SQA3D

VLM-Generated and Evaluated 3D Scene Understanding Questions

李沛宸 B10902034, 林祐辰 B10902033, 范秉逸 B109020117

GITHUB PAGE

INTRODUCTION

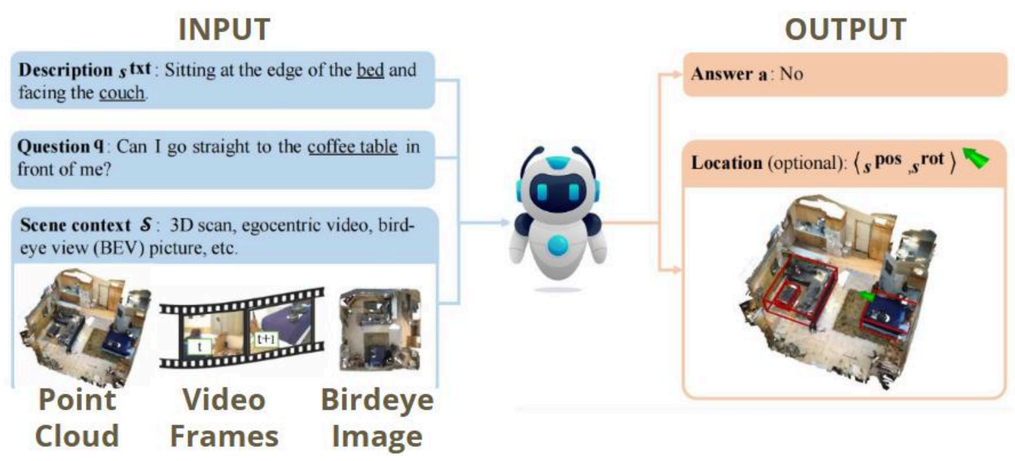

Figure 1. How the SQA3D Dataset works

SITUATION UNDERSTANDING

SPATIAL UNDERSTANDING

SPATIAL REASONING

MOTIVATION & PROBLEM DEFINITION

Initially, we want to do the SQA3D challenge, and we want to improve the Video3DLLM paper by giving it position data for training. However, through our pilot study we found out that the SQA3D dataset has a lot of problems, and that we don't meet the hardware requirement for training Video3DLLM's model. Thus, we thought, "Can Gemini improve the data quality of 3D situation question answering?". We performed a series of prompt tuning to generate Gemini's data based on SQA3D's videos, and compare the generated dataset with the original dataset through human evaluation and Gemini self evaluation.

Our problem is therefore defined as,

Can Gemini generate a dataset better than SQA3D? How can we evaluate the dataset?

PILOT STUDY - INVESTIGATING GEMINI

Our pilot study investigates whether Gemini 2.5 Pro has the ability to answer questions from the SQA3D dataset. Here are the findings.

- **Gemini is great at both Reasoning and Video Recognition**
- **Some questions in the SQA3D dataset are ambiguous or wrong**

Figure 2. Example of Gemini 2.5 Pro's Answer

METHODOLOGY: PROMPT GENERATION

Gemini is provided the scene data(bird's-eye image & scene video). and then guided through a reasoning flow:

1. **Select a viewpoint.**
2. **Observe the surroundings.**
3. **Describe the scene in context.**
4. **Generate three SQA data that includes spatial challenges.**

(Spatial challenges includes visibility, occlusion, relative direction, object alignment, and what can or cannot be seen.)

S: I am at a workstation that has a large stack of papers on it. A whiteboard is mounted on the wall to my left, and a cardboard box is on the floor to my right.

Q: If I walk straight forward, past the desk in front of me, and into the main open area, what object will be on the desk immediately to my left?

A: pink piggy bank.

EVALUATION

We define five criteria to evaluate the quality of each question-answer pair:

1. **You can know where you are based on the situation.**
2. **You can understand the question perfectly.**
3. **The answer is correct based on the situation and question**
4. **To answer this question, the video or birdeye image of the scene is necessary**
5. **To answer this question, you need to understand the 3d relations of the items in the scene and yourself.**

Each criterion is worth 0.2 points, for a total of 1 point per QA pair.

We evaluate performance from two perspectives: human and LLM .

- For human evaluation, we built a web interface with five checkbox —one for each criterion—allowing annotators to rate each QA pair.
- For Gemini, we give a prompt that explains the five criteria and asks the model to self-assess which criteria are satisfied.

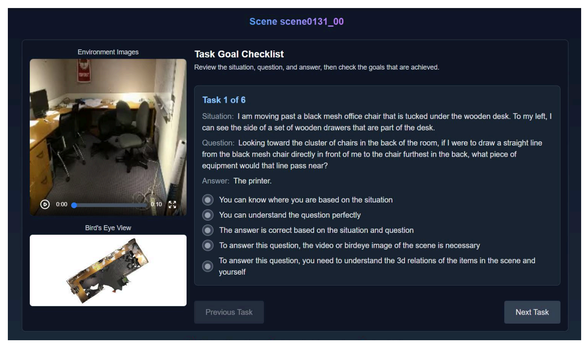

Figure 3. Human Evaluation Web UI

Figure 4. Result

RESULT & CONCLUSION

- Based on our five criteria evaluation, Gemini's dataset shows **marginal improvement** over the original SQA3D dataset.
- Gemini's questions are more **straightforward**, and the answers demonstrate **better alignment** with the scene and viewpoint.
- However, Gemini still falls short in certain aspects.
  - Some questions **misinterpret bird's-eye view images**.
  - Some issues include **misinterpret left & right** and **rotating 180° instead of rotating 90°**.

Gemini is effective at enhancing question quality compared to existing datasets, but it can still improve in generating more challenging and correct SQA problems.

REFERENCE

[1] Zheng, Duo, Shijia Huang, and Liwei Wang. "Video-3d llm: Learning position-aware video representation for 3d scene understanding." Proceedings of the Computer Vision and Pattern Recognition Conference. 2025.