What Do We Learn from Using Text Captions as a Form of 3D Scene Representation?

Vladyslav Humennyy[1†] Volodymyr Kuzma[1†] Ruslan Partsey[1†]
Sergio Arnaud[2] Franziska Meier[2] Oleksandr Maksymets[2]
[†] Equal contribution. 1. Ukrainian Catholic University 2. Meta AI

Abstract

Large language models (LLMs) encode a wealth of semantic knowledge about the world, which could be useful for embodied agents to understand their environment. However, current LLMs are not grounded in the real world and cannot directly perceive it. This work investigates the extent to which representing 3D scenes with text descriptions (scene captions) can bridge this gap; we focus on the Embodied Question Answering (EQA) task, explore different types of scene captioning and evaluate the performance of LLMs on a subset the OpenEQA benchmark episodes. Our findings show that (1) detailed captions explicitly describing object attributes, spatial relationships, and potential interactions provide significant benefits in EQA performance, even surpassing the state-of-the-art method in OpenEQA (GPT4-V); and (2) despite this great performance, even the best textual representations fall short of the perceptual and reasoning abilities demonstrated by humans when given visual data. These results suggest inherent limitations in using purely text-based scene descriptions and highlight the need for multimodal approaches that integrate visual data for more robust scene understanding.

1. Introduction

Large transformer architectures [15] trained to autoregressively predict the next token based on the previous tokens achieve unprecedented generalization capabilities [11–14, 17]. These capabilities gained by processing large text corpora (and exploring statistical correlations) during training make large language models (LLMs) state-of-the-art approaches in language modelling tasks. With the scale of model parameters, dataset size, and distributed training, abilities far beyond language modelling are starting to emerge [16]. Scale lets the model encode more semantic information that can be used as "world knowledge" to generate relevant answers. As a result, LLMs prompted (in a form of text instruction) to execute a particular task, achieve a superior performance across a wide range of tasks in various domains zero shot. One such domain is Robotics/Embodied AI, where LLMs were initially used for planning [1, 2, 7]. However, with advances in multi-modal research, large models have become able to perceive the world via the visual modality [3, 8, 9, 18]. This unlocked the possibility of using one model as an intelligent agent for perception and decision-making [6]. Nevertheless, according to Majumdar *et al*. [10], LLM-based agents are far from Humans on the Embodied Question Answering (EQA) task [5], which requires good 3D world understanding and reasoning capabilities. This gap motivates us to study how humans describe scenes to learn what information the scene representation should contain to achieve a human-level performance. Thus, in scope of this work we aim to answer the following **research questions**:

1. What information about the scene should the representation contain to enable LLM achieve human level performance on OpenEQA? Is it possible to create a "perfect" scene representation (Definitions 3.1 and 3.2)?
2. How the choice of LLM impacts overall performance?
3. What is a context size of the best scene representation?

2. Benchmark and evaluation

Benchmark ScanNet [4] subset of the OpenEQA benchmark: 17 scenes with highest number of questions (each with 14 questions, 2 per category) per scene.
Evaluation LLM-Match score (Appendix A.2).

3. Approach

Definition 3.1. ***Perfect scene representation*** *- representation that contains enough information about the scene that with alignment enables LLM to reach human level performance on broad range of Embodied AI tasks.*

Definition 3.2. ***Perfect scene caption*** *- form of perfect scene representation in text modality.*

Assumptions:
1. If there is a perfect scene representation then it can be used to produce a perfect scene caption (*i.e*. text can be used as a proxy to represent the encoded information about the scene).
2. LLMs properly aligned with perfect scene (3D) representation should perform at least as good as Socratic LLMs with perfect scene caption(*i.e*. Socratic LLMs with perfect

#	Method	EQA Category							LLM-Match	Scores Per scene
		attribute recognition	functional reasoning	object localization	object recognition	object state recognition	spatial understanding	world knowledge		
Free-form Scene Caption										
1	GPT-4	69.4±0.4	63.7±1.5	48.5±1.9	54.4±2.2	69.4±0.4	45.1±3.4	52.9±0.7	57.6±0.5	Tab. 3
2	LLaMA-2	64.9±1.6	67.4±1.1	48.3±2.1	50.5±1.1	74.8±3.4	48.5±4.4	55.1±2.2	58.5±0.7	Tab. 5
3	Human	58.0	58.5	39.3	45.5	44.9	39.7	52.2	48.3	Tab. 7
Structured-form Scene Caption										
4	GPT-4	59.5±1.7	54.4±2.3	44.6±0.7	44.9±1.4	63.7±1.1	41.6±5.7	45.0±0.8	50.5±1.5	Tab. 4
5	LLaMA-2	56.1±1.2	63.7±1.2	41.3±1.3	48.2±1.6	67.9±2.6	47.2±3.5	48.2±2.4	53.2±0.6	Tab. 6
6	Human	53.2	62.1	44.5	51.6	47.3	44.9	51.2	50.7	Tab. 8
VLM-generated Scene Caption										
7	GPT-4 w/ GPT-4V	54.6±0.8	56.4±1.1	38.0±0.4	41.4±0.4	71.3±2.0	38.7±0.4	49.0±0.4	49.9±0.2	
8	LLaMA-2 w/ GPT-4V	56.3±2.3	59.3±2.8	32.4±1.3	41.9±0.7	66.7±1.7	37.0±3.6	47.3±1.9	48.7±1.9	
Scene Video										
9	GPT-4V (50 frames)*	65.2	63.8	53.3	51.4	57.7	42.6	52.3	55.3±1.1	
10	Human*	87.9	81.8	77.3	87.9	98.7	86.7	87.2	86.8±0.6	

Table 1. **Category-level performance.** * - results reported in the OpenEQA paper [10].

scene captions provide a lower bound performance of LLMs with perfect scene representations).

1-scene study We start from one scene (`0164_02`) and manually iteratively improve the scene description to check if it is possible to achieve the maximum LLM-Match score. We call such hand-crafted description format "structured-form scene caption".

Human study We then scale 1-scene study to multiple scenes by involving independent annotators to describe 17 scenes. But before guiding the annotators with 1-scene study example, we ask them to describe the scene in a free format with the goal: "written scene description could be used to answer any question about the given scene". We call this description format "free-form scene caption". Afterwards, we ask annotators to write a scene description following the "structured" format from 1-scene study. Eventually, we get 2 captions formats per 17 scenes.

We also ask another group of annotators to answer the episode questions based on provided "free" and "structured" scene captions (Appendix B).

Models study We use LLaMA-2 {7, 70}B and GPT-4 as embodied agents and GPT-4V for generating image captions. First, we evaluate LLM-based agents on "free" and "structured" scene captions to compare the results with humans. Then, we use these captions as a few-shot examples for the GPT-4V to generate the scene descriptions. Afterwards, we use generated captions to benchmark the LLMs on OpenEQA. We compare the results with captions without examples and also quering VLM directly using only images (Appendix C).

Measuring the context length Actual length of the description - 782 tokens. We use Eq. (2) to estimate the context length (in GPT tokens) for the perfect scene caption. Longest object description: "a white-blue-red beer cardboard box with label "Samuel Adams, Boston Lager" - 17 tokens. Longest relative location description: "a cabinet with a sink is near the other wall of a room to the right of the cabinet with sparkling water maker." - 24 tokens. Given that there are 37 objects (including 7 receptacles), estimated caption length is - 1517 tokens (41 tokens per scene object).

4. Concluding remarks

First, we disentangle human performance on perceptual information represented as text from visual modality. Our human study shows that the capabilities of text representation bound EQA performance: humans achieve much higher EQA score given videos than captions ($\approx 50\%$ on text *vs.* 86.8% given episode videos (Tab. 1 rows (3), (6) vs row (10)); even though scene captions were written by humans). This suggests that while detailed text helps, integrating visual data might be necessary for models to fully understand and interact with 3D environments. *Second*, our models study provides more evidence that choice of LLM impacts the EQA performance. We do see the difference in the performance when models size differ significantly, meaning that larger model types are better than smaller *e.g.* LLaMA-2 70B > LLaMA-2 7B (Tab. 2). However, when the models pass the certain "performance" threshold, their EQA performance saturates. Some models may be slightly better in one categories of questions and slightly worse in others, but perform pretty close on average (LLaMA-2 70B and GPT-4 show pretty close scores; Tab. 1). *Also*, we estimate the best text representation context length which is 1517 tokens long. Which is $2X$ larger than default VLM generated captions. *Finally*, we show that in the context of EQA task, usage of caption examples for few-shot prompting of the VLM does not improve the quality of generated captions (Tab. 9).

References

[1] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Daniel Ho, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Eric Jang, Rosario Jauregui Ruano, Kyle Jeffrey, Sally Jesmonth, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Kuang-Huei Lee, Sergey Levine, Yao Lu, Linda Luu, Carolina Parada, Peter Pastor, Jornell Quiambao, Kanishka Rao, Jarek Rettinghouse, Diego Reyes, Pierre Sermanet, Nicolas Sievers, Clayton Tan, Alexander Toshev, Vincent Vanhoucke, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Mengyuan Yan, and Andy Zeng. Do as i can and not as i say: Grounding language in robotic affordances. In *arXiv preprint arXiv:2204.01691*, 2022. 1

[2] Boyuan Chen, Fei Xia, Brian Ichter, Kanishka Rao, Keerthana Gopalakrishnan, Michael S. Ryoo, Austin Stone, and Daniel Kappler. Open-vocabulary queryable scene representations for real world planning, 2022. 1

[3] Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning, 2023. 1

[4] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017. 1

[5] Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. Embodied Question Answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1

[6] Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. Palm-e: An embodied multimodal language model, 2023. 1

[7] Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, Pierre Sermanet, Noah Brown, Tomas Jackson, Linda Luu, Sergey Levine, Karol Hausman, and Brian Ichter. Inner monologue: Embodied reasoning through planning with language models, 2022. 1

[8] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023. 1

[9] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 1

[10] Arjun Majumdar, Anurag Ajay, Xiaohan Zhang, Pranav Putta, Sriram Yenamandra, Mikael Henaff, Sneha Silwal, Paul Mcvay, Oleksandr Maksymets, Sergio Arnaud, Karmesh Yadav, Qiyang Li, Ben Newman, Mohit Sharma, Vincent Berges, Shiqi Zhang, Pulkit Agrawal, Yonatan Bisk, Dhruv Batra, Mrinal Kalakrishnan, Franziska Meier, Chris Paxton, Sasha Sax, and Aravind Rajeswaran. Openeqa: Embodied question answering in the era of foundation models. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1, 2

[11] OpenAI. Gpt-4 technical report. *ArXiv*, abs/2303.08774, 2023. 1

[12] OpenAI. Gpt-4v(ision) technical work and authors. https://openai.com/contributions/gpt-4v, 2023.

[13] OpenAI. Gpt-4v(ision) system card. https://openai.com/contributions/gpt-4v, 2023.

[14] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023. 1

[15] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. 1

[16] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models, 2022. 1

[17] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of lmms: Preliminary explorations with gpt-4v(ision), 2023. 1

[18] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 1

What Do We Learn from Using Text Captions as a Form of 3D Scene Representation?

Supplementary Material

A. Towards "perfect" caption (1-scene study)

A.1. "Perfect" caption for one scene

Our first goal was to understand what is the optimal textual representation of the scene for LLM to perform embodied question answering, similar to what is done in OpenEQA [10]. To do this, we decided to perform manual captioning of a single scene, varying the level of detalisation and how relations between the objects are described. We avoided looking at the questions of the particular scene to prevent bias and worked independently to get diverse approaches. By comparing the results of representations we hoped to get some insights on how LLM performs with different text input.

We used an iterative approach to create captions - start with baseline, and then improve it in various aspects. Here are captions, that we created:

1. `Baseline` representation. It contained all objects that were on scene, though with minimal attribute details. Spatial relationships were kept simple, focusing on basic placement such as whether an object was on or below another.
2. `Generic` representation. For that representation we included all attribute information that we could describe and we included basic spatial relations like "near", "along", "next to".
3. `Absolute` representation. Like the generic representation, this one contained detailed attribute information. However, it also defined the position of the agent in the room and established absolute directions, indicating where "north", "south", "east", and "west" were in the scene. Using these directions, the positions of objects in the room were described.
4. `Relative` representation. This representation employed relative spatial directions such as "to the left/right of", "closer", and "further" to describe the relationships between objects.
5. `Tuned` representation. After creating all the above representations, we evaluated the questions that the model couldn't answer and made adjustments to enable the model to provide correct responses. These tweaks were implemented within the relative representation, which had demonstrated the best performance.

Using an iterative approach, we were able to assess the responses of language models to varying captions and, crucially, to identify changes in performance as captions evolved. The most important models' performance over categories is shown at Figure 1, while whole episode performance is shown at Table 2.

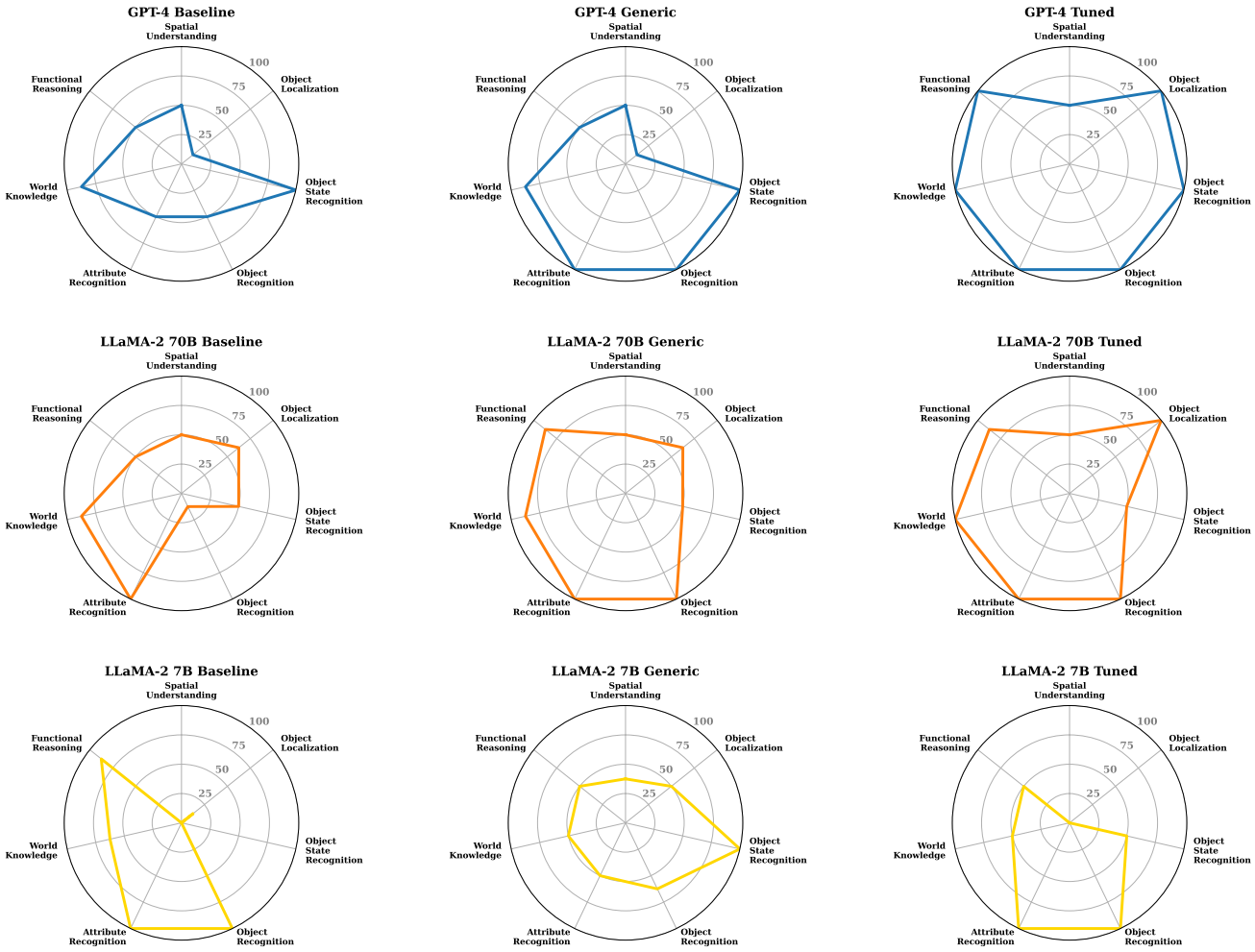
After evaluating everything and attempts to get "perfect" scene representation we got a few of insights about performance of models at that task. Here are they:

- **Complex 3D relationships require precise formulation of questions and a lot of context space in caption.** The best example of that is the only question, that we couldn't make the model to answer consistently - "Is there space for me to put something right next to the sink on the right hand side?". To answer such question it is not sufficient to have knowledge about direction where we can find the object, we should also have the information of the distance between the objects. And to construct such relations in text description is very token-exhaustive.
- **Initial Object Enumeration Helps.** By comparing evaluation results, we found that beginning captions with a comprehensive list of objects improves model performance on EQA tasks. We observed this improvement even when adding the enumeration to existing captions without introducing new information.
- **Same objects should preserve same attributes.** Object attributes in descriptions serve as unique identifiers. Thus, changing an attribute over the caption, even slightly, can disrupt this unique identification. Consistent use of attributes across the caption is essential to maintain correspondence.
- **Different model have very different results.** During initial experiments, we used GPT-4 and LLaMA-2 7B, noting that GPT-4 consistently outperformed LLaMA-2 7B, especially on captions with complex spatial relationships. However, when testing LLaMA-2 70B, we found its performance closely matched that of GPT-4, indicating that model size and complexity play a significant role.

Based on these insights, we developed a hypothesis for constructing captions with higher performance in EQA tasks for large language models. The structured caption approach follows these guidelines:

1. Captions should start with an enumeration of all objects.
2. Spatial relationships should be specified following the initial enumeration.
3. Attributes used as object identifiers must be assigned carefully and maintained consistently.

Due to the defined structure of such captions we define it as a "structured-form caption".



(a) `Baseline` (b) `Generic` (c) `Tuned`

Figure 1. **One-scene caption design study.**

#	Method	`Baseline`	`Generic`	`Tuned`
1	GPT-4	57	71.5	92.75
2	LLaMA-2 70B	59	76.75	84
3	LLaMA-2 7B	51.75	57.25	50

Table 2. **One scene study results.** Comparison of GPT-4, LLaMA-2 {7B, 70B} at different captions of the scene.

A.2. Evaluation method

For evaluation of performance on the EQA task, we used LLM-Match method [10] to assign a score from 1 to 5 to each answer, where 1 represents the lowest rating and 5 the highest. After that we calculated correctness score with exact formula from OpenEQA paper:

$$C = \frac{1}{N} \sum_{i}^{N} \frac{\sigma_i - 1}{4} \times 100\% \quad (1)$$

Here N is number of questions and σ_i is the LLM-Match score for every individual question.

To mitigate stochasticity of LLM's answers, which could be very different from one run to other, we decided to ask the model the same question multiple times to gather multiple data points for each question.

In the initial study, we took the average of three scores for each question, which allowed us to calculate mean answer value, but lacked other statistics.

For subsequent studies, we asked the model each question three times, recording each score individually. This approach allowed us to create dataset of scores for each question, which made it easy to compute mean and standard deviation values of LLM-Match score for each scene and each question category.

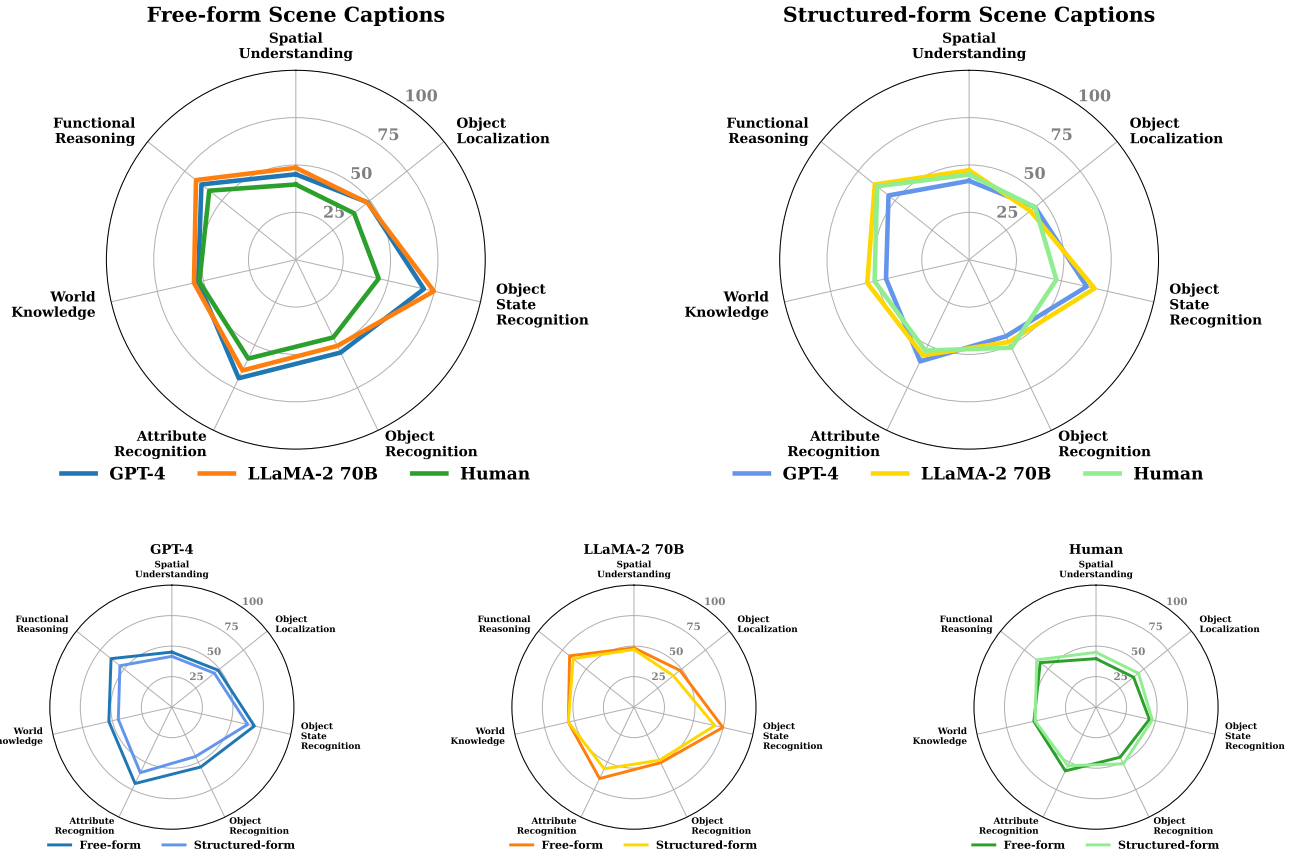


Figure 2. **Model results on free-form and structured-form captions.**

B. Human study

To explore the maximum performance potential of VLMs in generating captions from video content and LLMs in interpreting these captions, we designed a human-centered study. In this study, participants were asked to perform the same tasks as the models. The outcomes from these human participants were then used as a benchmark to compare against the model performances in identical scenarios.

Since each experiment required different tasks, we designed the next multi-phase experiment:

1. Request participants to create captions for various scenes in a way that feels most natural to them. We would refer to those captions as "free-form caption".
2. Instruct participants to create captions for the same scenes but using a more structured approach as specified by our guidelines. We would refer to those captions as "structured-form caption".
3. Ask participants to answer questions about the scenes, relying on the captions produced in the earlier stages.

Also, we ensured, that if the person got the scene at certain phase, they would not have the same scene at the future phase.

B.1. Captioning

The processes for creating both free-form and structured-form captions were quite similar. Each participant was assigned a scene to describe using provided source information. This source material included:

- A video of the scene that required captioning.
- An optional scan of the scene to offer additional context.

Participants were tasked with creating a comprehensive description that would contain enough information to answer questions about any object within the scene. Additionally, they were provided with sample questions to guide them on the types of details their descriptions should include.

For the structured-form captioning they had additional instructions except the previously mentioned. The main change that was requested from people was a determined structure of their caption:

- Describe from which point of view the room is described.
- Give a general overview of the room (its purpose and appearance).
- Enumerate all objects in the scene, with as much detail as possible.

- Describe the spatial relationships between objects.

Furthermore, participants were advised to maintain consistent use of adjectives throughout their descriptions. For example, if an object was described as "red-blue-white", then it should be described in that way throughout the whole description.

This structured-form approach was designed to create captions resembling graphs, where objects are nodes and their spatial relationships form the edges. We suggested, that such way of representation would improve spatial understanding of LLMs through a clearer organizational structure.

B.2. Human EQA

The next phase of our research focused on assessing how accurately people could understand and respond to questions about scenes with generated captions. We created pairs of scenes and related questions, then asked participants to answer those questions. To ensure more consistent data, each scene with each type of caption (free-form or structured-form) was answered by two different participants.

A key aspect of this stage was identifying cases where participants couldn't answer questions based on the information provided in the captions. In such situations, we instructed them to skip the question rather than attempt to guess an answer. This approach helped us understand how many questions were unanswerable using only the captions. After evaluating the results, we found that around 40% of the questions were left unanswered when using free-form captions, compared to 30% for structured-form captions.

B.3. Results

Our study found that the best performance across all caption/model combinations was 58.5% accuracy, based on the LLM-Match metric. This result was achieved using the LLaMA-2 70B model with free-form captions, and GPT-4 delivered a similar outcome, with 57.6% accuracy on these same scenes.

However, when we used structured-form captions, the performance of the LLMs declined. LLaMA-2's accuracy dropped to 53.2%, and GPT-4's accuracy fell to 50.5%.

Despite this drop, our results remain quite strong compared to similar approaches mentioned in the OpenEQA paper. LLMs using human-created captions outperformed those using auto-generated captions from LLaVA-1.5 model or scene-graphs. The best result in the OpenEQA paper was 45.1% accuracy, achieved using automatically generated captions by LLaVA-1.5 answered with GPT-4. In contrast, our study's accuracy exceeded 57% with both LLaMA-2 and GPT-4 when using human-created captions. That suggests, that VLMs are still behind people in perception of the world and correct description of it through text.

But best-performing OpenEQA approach - direct answering on questions with use of multiframe VLM as GPT-4V - got high result of 57.4%, comparable to ours. That is probably caused by a great data loss of textual caption of a scene in comparison to visual data like video frames.

When people completed the same EQA tasks, their performance with structured-form captions was higher (50.7%) compared to free-form captions (48.3%). Notably, the largest difference appeared in question categories involving spatial aspects (object localization, object recognition, and spatial understanding). On these questions, the average human scores with structured-form captions were 5% higher than with free-form captions, which can be seen at Table 1. We believe this difference could be due to the clearer depiction of spatial relationships in structured captions, which likely helped people to better understand the context of a room, leading to improved accuracy.

A comparison of human and model performance on the same dataset reveals that humans tend to achieve lower scores when analyzing the same captions. However, this could be because large language models (LLMs) are better at generalization and making educated guesses even when certain information isn't explicitly provided in the captions. They typically respond with the most likely answer, even if the caption doesn't mention it, or there's some uncertainty. In contrast, humans tend to rely strictly on the information available in the text, skipping questions when they can't find the relevant details. This difference in approach might be a key factor explaining why LLMs often outperform humans in these tasks.

Aggregated scores for all mentioned EQA results are shown in tables 5-8.

Scene id	EQA Category							LLM-Match
	attribute recognition	functional reasoning	object localization	object recognition	object state recognition	spatial understanding	world knowledge	
0709_00	100.0±0.0	87.5±0.0	12.5±0.0	37.5±0.0	100.0±0.0	12.5±0.0	0.0±0.0	46.1±0.0
0745_00	0.0±0.0	50.0±0.0	50.0±0.0	37.5±0.0	100.0±0.0	54.2±7.2	62.5±0.0	50.6±1.0
0598_00	50.0±0.0	62.5±21.6	8.3±7.2	75.0±0.0	100.0±0.0	12.5±0.0	62.5±0.0	53.0±3.7
0050_00	100.0±0.0	50.0±0.0	37.5±0.0	12.5±0.0	83.3±28.9	0.0±0.0	50.0±0.0	47.6±4.1
0684_01	100.0±0.0	58.3±7.2	79.2±14.4	75.0±0.0	100.0±0.0	75.0±21.6	66.7±14.4	79.2±4.5
0193_00	100.0±0.0	95.8±7.2	41.7±7.2	0.0±0.0	0.0±0.0	58.3±36.1	58.3±7.2	50.6±7.4
0494_00	62.5±0.0	66.7±28.9	66.7±14.4	100.0±0.0	50.0±0.0	100.0±0.0	50.0±0.0	70.8±5.5
0461_00	62.5±0.0	62.5±0.0	100.0±0.0	66.7±26.0	50.0±0.0	50.0±0.0	50.0±0.0	63.1±3.7
0356_00	50.0±0.0	75.0±0.0	37.5±0.0	50.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	30.4±0.0
0583_00	87.5±0.0	37.5±0.0	25.0±0.0	33.3±14.4	66.7±28.9	79.2±26.0	20.8±14.4	50.0±1.8
0406_00	50.0±0.0	45.8±7.2	50.0±0.0	50.0±0.0	100.0±0.0	25.0±0.0	33.3±28.9	50.6±3.7
0655_01	50.0±0.0	50.0±0.0	66.7±14.4	87.5±0.0	50.0±0.0	100.0±0.0	62.5±0.0	66.7±2.1
0608_02	83.3±7.2	50.0±0.0	0.0±0.0	87.5±0.0	50.0±0.0	41.7±50.5	79.2±14.4	56.0±8.8
0685_02	100.0±0.0	50.0±0.0	50.0±21.6	50.0±0.0	50.0±0.0	58.3±14.4	58.3±7.2	59.5±3.7
0100_02	50.0±0.0	91.7±14.4	87.5±0.0	0.0±0.0	100.0±0.0	0.0±0.0	95.8±7.2	60.7±3.1
0655_02	100.0±0.0	100.0±0.0	100.0±0.0	100.0±0.0	87.5±0.0	45.8±28.9	95.8±7.2	89.9±3.7
0500_00	50.0±0.0	50.0±0.0	12.5±0.0	62.5±0.0	91.7±7.2	54.2±7.2	54.2±7.2	53.6±1.8
All	69.4±0.4	63.7±1.5	48.5±1.9	54.4±2.2	69.4±0.4	45.1±3.4	52.9±0.7	57.6±0.5

Table 3. **GPT-4 /w free-form scene caption.**

Scene id	EQA Category							LLM-Match
	attribute recognition	functional reasoning	object localization	object recognition	object state recognition	spatial understanding	world knowledge	
0709_00	25.0±0.0	50.0±0.0	12.5±0.0	37.5±0.0	100.0±0.0	17.5±6.8	0.0±0.0	35.4±1.1
0745_00	0.0±0.0	35.0±20.5	50.0±0.0	50.0±0.0	85.0±20.5	50.0±0.0	62.5±0.0	47.5±4.5
0598_00	100.0±0.0	62.5±0.0	25.0±0.0	42.5±6.8	100.0±0.0	40.0±5.6	100.0±0.0	67.1±1.6
0050_00	62.5±0.0	50.0±0.0	7.5±6.8	37.5±0.0	50.0±0.0	0.0±0.0	45.0±11.2	36.1±1.5
0684_01	100.0±0.0	50.0±0.0	67.5±6.8	37.5±0.0	60.0±22.4	70.0±16.8	70.0±16.8	65.0±5.3
0193_00	50.0±0.0	82.5±11.2	72.5±5.6	37.5±0.0	0.0±0.0	40.0±22.4	100.0±0.0	54.6±3.0
0494_00	62.5±0.0	50.0±0.0	87.5±0.0	62.5±0.0	50.0±0.0	100.0±0.0	0.0±0.0	58.9±0.0
0461_00	72.5±13.7	52.5±5.6	47.5±13.7	87.5±0.0	50.0±0.0	50.0±0.0	50.0±0.0	58.6±3.4
0356_00	0.0±0.0	67.5±16.8	2.5±5.6	50.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	17.1±2.7
0583_00	0.0±0.0	12.5±0.0	20.0±6.8	25.0±0.0	50.0±0.0	77.5±22.4	27.5±5.6	30.4±4.2
0406_00	87.5±0.0	25.0±8.8	0.0±0.0	30.0±16.8	100.0±0.0	25.0±0.0	0.0±0.0	38.2±1.6
0655_01	37.5±0.0	12.5±0.0	50.0±0.0	37.5±0.0	50.0±0.0	50.0±0.0	62.5±0.0	42.9±0.0
0608_02	50.0±0.0	62.5±0.0	62.5±0.0	12.5±0.0	50.0±0.0	20.0±27.4	87.5±0.0	49.3±3.9
0685_02	57.5±6.8	37.5±0.0	50.0±0.0	2.5±5.6	100.0±0.0	12.5±0.0	60.0±5.6	45.7±1.6
0100_02	100.0±0.0	75.0±0.0	87.5±0.0	75.0±0.0	100.0±0.0	50.0±8.8	12.5±0.0	71.4±1.3
0655_02	90.0±22.4	100.0±0.0	70.0±11.2	75.0±0.0	50.0±0.0	67.5±27.4	62.5±0.0	73.6±7.8
0500_00	100.0±0.0	100.0±0.0	45.0±11.2	62.5±0.0	87.5±0.0	37.5±0.0	25.0±0.0	65.4±1.6
All	59.5±1.7	54.4±2.3	44.6±0.7	44.9±1.4	63.7±1.1	41.6±5.7	45.0±0.8	50.5±1.5

Table 4. **GPT-4 /w structured-form scene caption.**

Scene id	EQA Category							LLM-Match
	attribute recognition	functional reasoning	object localization	object recognition	object state recognition	spatial understanding	world knowledge	
0709_00	100.0±0.0	75.0±21.6	0.0±0.0	37.5±0.0	100.0±0.0	12.5±0.0	0.0±0.0	42.3±3.3
0745_00	0.0±0.0	4.2±7.2	45.8±7.2	79.2±14.4	87.5±0.0	75.0±21.6	54.2±7.2	49.4±3.7
0598_00	33.3±14.4	58.3±14.4	12.5±0.0	83.3±7.2	33.3±28.9	0.0±0.0	62.5±0.0	40.5±4.5
0050_00	100.0±0.0	41.7±7.2	33.3±7.2	12.5±0.0	83.3±28.9	33.3±28.9	62.5±0.0	52.4±9.8
0684_01	100.0±0.0	87.5±0.0	75.0±0.0	62.5±21.6	100.0±0.0	50.0±0.0	58.3±31.5	76.2±5.2
0193_00	100.0±0.0	87.5±0.0	37.5±0.0	16.7±28.9	50.0±0.0	100.0±0.0	50.0±21.6	63.1±3.7
0494_00	41.7±14.4	66.7±28.9	83.3±28.9	87.5±0.0	50.0±0.0	50.0±50.0	33.3±28.9	58.9±3.6
0461_00	25.0±0.0	79.2±14.4	100.0±0.0	12.5±0.0	50.0±0.0	50.0±0.0	50.0±0.0	52.4±2.1
0356_00	50.0±0.0	100.0±0.0	79.2±14.4	50.0±0.0	100.0±0.0	37.5±0.0	16.7±28.9	61.9±5.5
0583_00	87.5±0.0	12.5±0.0	79.2±7.2	75.0±0.0	83.3±28.9	87.5±0.0	25.0±21.6	64.3±5.4
0406_00	50.0±0.0	45.8±7.2	37.5±0.0	50.0±0.0	50.0±0.0	50.0±21.6	41.7±7.2	46.4±1.8
0655_01	87.5±21.6	50.0±0.0	58.3±14.4	54.2±7.2	33.3±28.9	100.0±0.0	87.5±0.0	67.3±5.2
0608_02	50.0±0.0	87.5±0.0	0.0±0.0	37.5±0.0	50.0±0.0	20.8±26.0	75.0±21.6	45.8±1.0
0685_02	83.3±28.9	62.5±0.0	41.7±14.4	50.0±0.0	100.0±0.0	37.5±0.0	70.8±26.0	63.7±2.7
0100_02	50.0±0.0	87.5±0.0	75.0±21.6	20.8±14.4	100.0±0.0	4.2±7.2	100.0±0.0	62.5±4.7
0655_02	100.0±0.0	100.0±0.0	50.0±12.5	70.8±14.4	100.0±0.0	79.2±19.1	91.7±7.2	84.5±4.1
0500_00	62.5±0.0	100.0±0.0	12.5±0.0	58.3±7.2	100.0±0.0	37.5±0.0	58.3±14.4	61.3±1.0
All	64.9±1.6	67.4±1.1	48.3±2.1	50.5±1.1	74.8±3.4	48.5±4.4	55.1±2.2	58.5±0.7

Table 5. **LLaMA-2 /w free-form scene caption.**

Scene id	EQA Category							LLM-Match
	attribute recognition	functional reasoning	object localization	object recognition	object state recognition	spatial understanding	world knowledge	
0709_00	0.0±0.0	87.5±15.3	0.0±0.0	37.5±0.0	100.0±0.0	12.5±0.0	0.0±0.0	36.5±2.4
0745_00	0.0±0.0	7.5±6.8	50.0±0.0	50.0±0.0	95.0±6.8	87.5±0.0	50.0±0.0	48.6±1.5
0598_00	100.0±0.0	100.0±0.0	35.0±10.5	37.5±0.0	100.0±0.0	37.5±0.0	100.0±0.0	72.9±1.5
0050_00	62.5±0.0	45.0±6.8	0.0±0.0	37.5±0.0	100.0±0.0	72.5±20.5	50.0±0.0	52.5±3.2
0684_01	100.0±0.0	60.0±5.6	60.0±5.6	37.5±0.0	50.0±0.0	62.5±0.0	62.5±0.0	61.8±1.0
0193_00	50.0±0.0	87.5±0.0	50.0±8.8	0.0±0.0	40.0±22.4	50.0±0.0	62.5±23.4	48.6±3.2
0494_00	62.5±0.0	100.0±0.0	80.0±11.2	92.5±6.8	50.0±0.0	60.0±22.4	30.0±27.4	67.9±4.5
0461_00	75.0±12.5	77.5±13.7	57.5±24.4	72.5±20.5	50.0±0.0	50.0±0.0	0.0±0.0	54.6±5.7
0356_00	50.0±0.0	87.5±0.0	7.5±6.8	50.0±0.0	50.0±0.0	0.0±0.0	0.0±0.0	35.0±1.0
0583_00	0.0±0.0	17.5±11.2	17.5±6.8	55.0±16.8	50.0±0.0	45.0±6.8	35.0±13.7	31.4±4.1
0406_00	72.5±20.5	0.0±0.0	12.5±0.0	0.0±0.0	50.0±0.0	50.0±0.0	0.0±0.0	26.4±2.9
0655_01	27.5±13.7	0.0±0.0	42.5±6.8	72.5±13.7	40.0±22.4	100.0±0.0	65.0±13.7	49.6±7.7
0608_02	62.5±0.0	75.0±0.0	57.5±6.8	37.5±0.0	50.0±0.0	32.5±11.2	85.0±5.6	57.1±1.8
0685_02	0.0±0.0	50.0±0.0	47.5±5.6	22.5±5.6	80.0±27.4	35.0±5.6	87.5±0.0	46.1±4.1
0100_02	100.0±0.0	87.5±0.0	72.5±20.5	75.0±0.0	100.0±0.0	22.5±20.5	100.0±0.0	79.6±1.0
0655_02	100.0±0.0	100.0±0.0	50.0±0.0	85.0±5.6	50.0±0.0	47.5±33.5	62.5±0.0	70.7±4.7
0500_00	62.5±0.0	100.0±0.0	62.5±8.8	57.5±16.8	100.0±0.0	37.5±0.0	30.0±11.2	64.3±3.1
All	56.1±1.2	63.7±1.2	41.3±1.3	48.2±1.6	67.9±2.6	47.2±3.5	48.2±2.4	53.2±0.6

Table 6. **LLaMA-2 /w structured-form scene caption.**

Scene id	EQA Category							LLM-Match
	attribute recognition	functional reasoning	object localization	object recognition	object state recognition	spatial understanding	world knowledge	
0709_00	100.0	68.8	12.5	37.5	43.8	12.5	25.0	38.5
0745_00	6.2	0.0	25.0	18.8	43.8	25.0	43.8	23.2
0598_00	87.5	81.2	18.8	62.5	100.0	0.0	75.0	60.7
0050_00	75.0	56.2	37.5	12.5	43.8	25.0	37.5	41.1
0684_01	75.0	75.0	18.8	56.2	75.0	50.0	50.0	57.1
0193_00	93.8	31.2	43.8	62.5	6.2	75.0	87.5	57.1
0494_00	75.0	50.0	81.2	75.0	6.2	87.5	0.0	53.6
0461_00	43.8	87.5	43.8	25.0	43.8	56.2	50.0	50.0
0356_00	0.0	75.0	43.8	37.5	0.0	6.2	0.0	23.2
0583_00	87.5	18.8	56.2	43.8	68.8	87.5	25.0	55.4
0406_00	6.2	43.8	0.0	0.0	75.0	25.0	6.2	22.3
0655_01	81.2	50.0	75.0	87.5	0.0	75.0	81.2	64.3
0608_02	18.8	62.5	0.0	37.5	0.0	68.8	62.5	35.7
0685_02	75.0	56.2	56.2	62.5	81.2	37.5	87.5	65.2
0100_02	31.2	100.0	81.2	25.0	0.0	12.5	100.0	50.9
0655_02	100.0	87.5	68.8	87.5	100.0	25.0	62.5	75.9
0500_00	50.0	50.0	6.2	37.5	75.0	6.2	93.8	45.5
All	58.0	58.5	39.3	45.5	44.9	39.7	52.2	48.3

Table 7. **Human /w free-form scene caption.**

Scene id	EQA Category							LLM-Match
	attribute recognition	functional reasoning	object localization	object recognition	object state recognition	spatial understanding	world knowledge	
0709_00	50.0	50.0	12.5	37.5	56.2	12.5	18.8	32.7
0745_00	0.0	75.0	56.2	31.2	68.8	50.0	87.5	52.7
0598_00	81.2	68.8	18.8	62.5	75.0	18.8	81.2	58.0
0050_00	56.2	56.2	6.2	18.8	50.0	25.0	25.0	33.9
0684_01	100.0	68.8	75.0	93.8	75.0	50.0	81.2	77.7
0193_00	50.0	43.8	62.5	18.8	0.0	25.0	87.5	41.1
0494_00	75.0	50.0	93.8	100.0	68.8	75.0	43.8	72.3
0461_00	62.5	68.8	50.0	50.0	50.0	56.2	31.2	52.7
0356_00	0.0	43.8	0.0	43.8	25.0	6.2	43.8	23.2
0583_00	6.2	25.0	12.5	25.0	25.0	81.2	37.5	30.4
0406_00	37.5	56.2	25.0	37.5	50.0	43.8	6.2	36.6
0655_01	37.5	12.5	50.0	87.5	0.0	37.5	75.0	42.9
0608_02	50.0	87.5	62.5	37.5	0.0	100.0	87.5	60.7
0685_02	31.2	50.0	31.2	18.8	93.8	18.8	56.2	42.9
0100_02	100.0	93.8	87.5	75.0	18.8	62.5	43.8	68.8
0655_02	75.0	93.8	75.0	75.0	25.0	62.5	56.2	66.1
0500_00	81.2	100.0	50.0	75.0	75.0	62.5	37.5	68.8
All	53.2	62.1	44.5	51.6	47.3	44.9	51.2	50.7

Table 8. **Human /w structured-form scene caption.**

C. Models study

```
You are an intelligent agent.  Your task
is to describe the scenes on photos in
the next messages.  In the future, you
will be given questions about the scenes
and you will be able to use only this
generated descriptions, so make them very
detailed.

When you describe the scenes focus on
describing the following aspects:
What objects do you see and how you can
describe them in detail?
Where are they located?
Where are they located in comparison to
other objects?
What you can do with those objects?

Your response should start with "Answer:"
and then continue with your description.
```

Figure 3. Prompt used for few-shot captioning

For model study, we used 3 best free-form captions as few-shot examples for GPT-4V. For frame selection, two approaches were used: random uniform sampling and JPEG+SSIM approach (described in Subsection C.1).

The model was given the main prompt, shown on Figure 3, and pairs of frames with corresponding free-form captions, starting with "Answer", then it was given frames from the unseen episode to create caption. The obtained caption was evaluated on the questions corresponding to the episode. The experiments included evaluation of JPEG-based frame-selection algorithm relative to the random baseline, and comparison of zero-shot, one-shot, and two-shot captions on OpenEQA. All of the tests were performed on 15 episodes, each with 11 questions (these are not the same that were used for human study). Here is the full list of parameters chosen for experiments:

1. 0,1,2-shot caption generation on 5 frames with JPEG+SSIM frame selection.
2. One-shot caption generation on 10 frames with JPEG+SSIM frame selection.
3. One-shot caption generation on 10 frames with uniform frame selection.

C.1. Frame selection

Uniform selection served as a baseline. A uniform sampling of predefined number of frames was performed over all length of the episode.

JPEG+SSIM approach utilized JPEG weight as a ranking score to determine the most detailed frames, and used Structure Similarity Index Measure (SSIM) to avoid selection of identical views, therefore capturing the biggest possible visual context. The pipeline of the algorithm follows:

- Frames of the episode are chronologically separated into n parts of equal length where n is the number of frames that we want to select.
- In each formed part, the frames are sorted by the weight of their JPEG-compressed forms in decreasing order (the biggest frames are considered the best).
- Starting from the first, we iterate over parts, selecting the one biggest (in terms of JPEG weight) frame from each, such that its SSIM score against previous selected frame is less than 0.4 (adjacent frames which are more likely to show similar view, are selected to be different).

C.2. Results

Aggregated results of these experiments can be seen in Table 9. The specified model is the one that was used for evaluating caption, not generating it.

As seen in the table, the differences between zero-shot and one-shot captions are relatively small with zero-shot captioning capturing the most information. Two-shot captioning obviously lacked the detail due to usually being very short (more general conclusion about this in Subsection C.3), though structurally close to the human-written caption.

The JPEG+SSIM selection method did not show a positive change relative to uniform sampling, but should be further investigated as a way to minimize number of retries during captioning (blurry frames that are often selected with uniform sampling can cause LLM's "I can't assist with request" answer which is then captured by the code and causes retry).

C.3. Observations

Along with these numbers, we include several observations, made during the experiments, that we believe can help the future research in the field:

1. Inclusion of bigger number of frames does not make the description more detailed. To this point, when LLM was given 20 frames per episode in 1-shot case (40 frames in total) it failed to generate caption at all in most captions ("I can't assist in this case" answer) and generated a very short overview of all frames in other cases, while for 5 and 10 frames it usually provided a description of each frame and general overview of the scene.
2. Prompts for captioning should be task-specific. This is because VLM descriptions of context rich 3D spaces are very general and lack detail, unless the model was given specific attributes to capture. Even then it would not capture most details, but it will perform better on the given task.
3. A typical generated caption of the episode (in our experiments) included a list of descriptions for given frames.

#	Method	N frames	Frame selection	EQA Category							LLM-Match
				object recognition	object localization	attribute recognition	spatial understanding	object state recognition	functional reasoning	world knowledge	
Zero-shot											
1	GPT-4	5	JPEG+SSIM	51.7±0.6	52.6±1.3	42.4±5.0	36.2±1.5	73.3±2.9	33.6±1.7	50.4±1.8	48.3±0.8
2	LLaMA-2	5	JPEG+SSIM	51.7±0.6	44.7±4.7	38.8±0.6	39.4±1.7	83.7±2.1	40.7±2.0	50.4±0.7	50.1±0.6
One-shot											
3	GPT-4	5	JPEG+SSIM	57.0±1.0	59.6±0.8	32.2±3.1	26.9±1.0	77.0±2.6	31.1±2.0	44.8±0.7	46.5±0.5
3	GPT-4	5	Uniform	37.0±1.7	56.6±1.3	40.9±2.5	26.9±1.0	84.7±1.1	33.0±0.6	54.4±0.7	47.0±0.1
3	GPT-4	10	JPEG+SSIM	46.7±0.6	50.9±2.7	31.2±1.7	36.2±2.2	83.0±0.0	39.1±0.6	56.0±1.2	48.8±0.2
3	GPT-4	10	Uniform	50.0±0.0	50.9±0.8	35.5±0.6	26.9±1.0	71.0±0.0	29.8±1.7	55.2±3.0	45.1±0.4
4	LLaMA-2	5	JPEG+SSIM	51.7±3.2	47.8±4.0	24.6±2.3	34.0±2.2	70.3±2.9	29.2±3.6	48.8±3.1	43.6±0.9
4	LLaMA-2	5	Uniform	41.0±3.5	57.5±4.6	30.8±1.3	23.4±3.1	95.0±0.0	44.2±0.0	54.8±1.2	49.1±0.2
4	LLaMA-2	10	JPEG+SSIM	42.0±1.0	57.0±2.0	34.1±1.3	27.6±1.5	78.7±4.0	31.7±1.0	50.0±3.1	45.3±0.5
4	LLaMA-2	10	Uniform	49.0±0.0	49.1±2.0	29.0±1.7	21.8±1.1	90.7±2.3	35.6±1.0	61.9±1.2	47.8±0.1
Two-shot											
5	GPT-4	5	JPEG+SSIM	40.0±1.0	53.1±2.0	38.0±2.2	31.7±2.5	75.3±0.6	35.9±0.6	47.6±0.0	45.6±0.8
6	LLaMA-2	5	JPEG+SSIM	48.3±1.5	45.2±3.0	26.1±0.0	27.6±0.6	80.0±4.0	27.2±2.9	54.8±3.1	43.9±0.7

Table 9. **Model study results.** Comparison of zero-shot, one-shot, and two-shot prompting on 5 frames per episode, using JPEG+SSIM frame selection method.

This captured object attributes and most relations on pictures, but was not "stitching" the scene together. In 1-shot and 2-shot cases, the model often created a general description of the whole scene, not separating it into frames, and closely resembling the one created by human. Nevertheless, these attempts captured not enough detail to actually perform better on OpenEQA.

D. Caption context length estimation

Maximum **C**aption **C**ontext **Len**gth estimation formula:

$$MCCLen = (MObj + MSpat) \times N \qquad (2)$$

where:
N - number of objects seen in episode history;
MObj - **M**aximum **Obj**ect description length;
MSpat - **M**aximum object **Spat**ial relation description length.

It is a relatively small room for meetings, presentations, or study rooms. The main object in the center of the room is the big rectangular table with eight black metal legs; more precisely, two square tables were combined into one. To the left of the desk on the wall is a large rectangular whiteboard on a white wall. Behind the table and in front of it, large gray sheets hang almost on the whole white walls, which are used to project images onto. The wall on the right is blank, painted with white paint, but it looks almost gray in this light. The steel is also painted the same color as the wall, and the floor is completely covered with dark gray carpeting. The walls and carpet are linked by a wood-colored skirting board.
There are 20 chairs in the room, in black, red, blue and yellow colors. The red chairs are soft, have a square back, and a metal frame. The black and blue chairs are soft and have a semicircular back, one leg, and 5 wheels. The yellow chairs are made of plastic and have a semicircular back with perforated holes and four legs on wheels. The color of the tables is light gray, with a red border around the table's edges. One of the tables which is left has a recess for sockets framed in black metal in its center. There are three black cables sticking out of it, two of which have blue plugs on the ends and are almost near the outlet, and one has a black plug right at the end, and this cable stretches almost across the table from left to right. To the right of the end of this cable is a white remote control, most likely from a projector. There is enough space between the cable and the remote control, for example, for a computer. To the left of the end of the black cable is a sloppily crumpled white paper, possibly used for notes. You can also see a marker standing in the lower left corner of the table. It is white, with a blue cap. Next to it, closer to us, is a white plug. Under the table, a thick black cable stretches from one of the sockets to the other end of the table.
Around the table and against the walls of the room are chairs in disorder, turned in different directions. 9 of them are around the table. They are all red except one, which is black. If you look from the entrance to the room, there are two red chairs on the near long edge of the table. The one on the right is directly facing the table, and the one on the left is facing the whiteboard on the left wall.
There are two chairs on the right side of the table, the closer black one is completely moved behind the table, and the farther red one is turned toward us. Most likely, it was turned so it would be easier to get up from the table. Behind these chairs, there are 4 more chairs near the wall. The nearest yellow plastic one is right behind the black chair. There is a square white socket on the back of it below. Then there are all three soft black chairs along the wall to the end of the wall.
There are three red chairs pushed behind the table, and another one on the left edge, also red one, but turned slightly toward the whiteboard. Behind these chairs, almost in the middle, there are two chairs set to the wall. The one on the right is black, and the one on the left is red. In the bottom right and left corners of the wall, there are square white plastic sockets. On the left edge of the table, a red chair is almost in the middle, pushed to the table. Behind it is a three-part whiteboard. The largest is rectangular in the middle, and two are square on the sides. There are notes on the board, and at the bottom of the board, under the rectangular and one of the square parts, which is closer to us, there are sponges for wiping and markers. At the top of the board, there is an oblong object, most likely, for lights. Above the whiteboard on the ceiling, there are two shining round lamps that are at an equal distance from the walls and each other. But in general, the room is quite dark. On the left, almost in the corner, under a whiteboard by the wall, there is a black chair. There is a rectangular white socket in the middle of the wall below.
On the left, there are 4 chairs near the wall. The one closest to us is plastic yellow, then a soft black one. Behind it is a soft blue one turned to the wall, and in front of the blue one, covering it, is a soft red table turned to the table. There are two circular vents on the floor, one in the far left corner, half-covered, and the other directly under the table in the lower right corner.

Figure 4. **Free-form (0655_02) scene caption example.**

You are looking at the room from the center of it.
The room contains a beige toilet with a matching seat and lid; a beige cylindrical trash can; black mat; dark green mat, a wooden cabinet with a light brown finish, backed by a beige countertop that stretches the entire length of the wooden cabinet; a wall-mounted toilet paper holder holding a roll of white toilet paper; and a pink folded towel lying on the countertop. Beige electric toothbrush, light-gray charger for the electric toothbrush, a rectangular blue soap dish, a pale yellow cylindrical cream, two turquoise cylindrical containers, and a tan-coloured oval sink. Dark gray rectangular mat. A closed white door, a blue robe, a pink and yellow bathrobe, a hook on the door. A pair of dark blue slippers with a floral print. A white stool with a blue tread on top. Three towels: beige, white and blue, and yellow and pink. There is also a bathtub rail, and a built-in beige bathtub.
The beige toilet with a matching seat and lid is located against the wall. Next to it, on the right, is a beige cylindrical trash can. To the left of the toilet is a wall-mounted toilet paper holder. Along the right side of the toilet and further along the wall is a wooden cabinet. This wooden cabinet supports the beige countertop along its entire length. A pink folded towel is placed at the far end of the beige countertop, away from the beige toilet and adjacent to the right side wall of the bathroom. A black mat is located directly in front of the beige toilet. A smaller, dark green mat, is placed next to the black mat, in front of the right part of the built-in beige bathtub. In the middle of the countertop there is a beige electric toothbrush in a vertical charging stand, a rectangular blue soap dish, a pale yellow cylindrical cream, and two turquoise cylindrical containers. To the right of two turquoise cylindrical containers there is a tan-coloured oval sink built into the beige countertop. The light-gray charger for the electric toothbrush is located on the right-hand side of the tan-coloured oval sink. The rectangular blue soap dish is located to the left of the toothbrush. Dark gray rectangular mat is placed on the floor in front of the wooden cabinet's part where the sink is placed. The closed white door is located opposite the wooden cabinet. A blue robe, a pink and yellow bathrobe are hung on a hook on the door. A pair of dark blue slippers with a floral print are on the floor opposite the closed white door and near the built-in beige bathtub. A white stool with a blue tread on top is placed in front of the left part of the bathtub, with 3 towels: beige, white and blue, and yellow and pink, hanging directly above it on the bathtub rail. The built-in beige bathtub is behind the stool and the smaller, dark green mat.

Figure 5. **Structured-form (`0100_02`) scene caption example.**