UDC 004.8:004.9:629.8
DOI: 10.15587/1729-4061.2025.323757

# ENHANCING 3D SCENE UNDERSTANDING VIA TEXT ANNOTATIONS

**Ruslan Partsey**
*Corresponding author*
PhD Student*
E-mail: ruslan.v.partsei@lpnu.ua
**Vladyslav Humennyy****
**Volodymyr Kuzma****
**Oleksandr Maksymets**
PhD
Meta AI Research
Meta Platforms, Inc.
Hacker Way, 1, Menlo Park, United States, CA 94025
**Vasyl Teslyuk**
Doctor of Technical Sciences*
*Department of Automated Control Systems
Lviv Polytechnic National University
S. Bandery str., 12, Lviv, Ukraine, 79013
**Department of Computer Sciences
and Information Technologies
Ukrainian Catholic University
Ilariona Sventsitskoho str., 17, Lviv, Ukraine, 79011

***The object of this study is the use of text annotations as a form of 3D scene representation. The paper investigates the task of integrating large-scale language models (LLMs) into complex 3D environments. Using the Embodied Question Answering task as an example, we analyze different types of scene annotations and evaluate the performance of LLMs on a subset of test episodes from the OpenEQA dataset. The aim of the study was to evaluate the effectiveness of textual scene descriptions compared to visual data for solving EQA tasks. The methodology implied estimating the optimal context length for scene annotations, measuring the differences between free-form and structured annotations, as well as analyzing the impact of model size on performance, and comparing model results with the level of human comprehension of scene annotations. The results showed that detailed descriptions that include a list of objects, clearly described attributes, spatial relationships, and potential interactions improve EQA performance, even outperforming the most advanced method in OpenEQA – GPT-4V (57.6 % and 58.5 % for GPT-4 and LLaMA-2, respectively, vs. 55.3 % for GPT-4V). The results can be explained by the high level of detail that provides the model with contextually clear and interpretable information about the scene. The optimal context length for textual descriptions was estimated to be 1517 tokens. Nevertheless, even the best textual representations do not reach the level of perception and reasoning demonstrated by humans when presented with visual data (50 % for textual data vs. 86.8 % for video). The results are important for the development of multimodal models for Embodied AI tasks, including Robotics and Autonomous Navigation Systems. They could also be used to improve user interaction with three-dimensional spaces in the field of virtual or augmented reality***

***Keywords: large language models, vision language models, multimodal learning, spatial understanding***

*Received 13.12.2025*
*Received in revised form 03.02.2025*
*Accepted date 29.01.2025*
*Published date 27.02.2025*

***How to Cite:*** *Partsey, R., Humennyy, V., Kuzma, V., Maksymets, O., Teslyuk, V. (2025). Enhancing 3D scene understanding via text annotations. Eastern-European Journal of Enterprise Technologies, 1 (2 (133)), 76–84. https://doi.org/10.15587/1729-4061.2025.323757*

## 1. Introduction

Deep transformer architectures [1], trained to autoregressively predict the next token based on the previous ones, achieve unprecedented generalization capabilities (GPT-4 [2], GPT-4V [3–5], LLaMA-2 [6]). These capabilities, obtained by processing large text corpora (and exploring statistical correlations) during training, make large language models (LLMs) a leading approach in language modeling tasks [7]. By scaling model parameters, dataset size, and distributed learning, capabilities that go far beyond language modeling [8]. The scale allows the model to encode more semantic information, which can be used as "world knowledge" to generate relevant responses. As a result, LLMs can effectively perform a wide range of tasks (from text generation to writing software code) in various fields (from marketing and finance to medicine) without additional training.

One such field is Robotics/Embodied AI, where LLMs were initially used for planning (SayCan [9] and its derived approaches [10, 11]). For context, a robotic platform typically consists of three components (modules): environment perception (recognizing input from onboard sensors), a planning module, and a controller that drives the hardware. However, with the development of multimodal research, large models have become capable of perceiving the world through visual modality (MiniGPT-4 [12, 13] LLaVA [14, 15]). This has opened the possibility of using a single model as an intelligent agent for environment perception and decision-making [16]. However, as noted in [17], LLM-based agents perform considerably below human levels in the Embodied Question Answering (EQA) task [18], which demands both an understanding of the three-dimensional world and advanced reasoning skills. This difference in results motivates us to investigate how humans describe scenes to learn what information a scene representation should contain to achieve human-level performance. By learning how humans describe scenes, AI systems can better represent scenes that capture the semantic and contextual nuances of 3D environments, which will facilitate better environment understanding and interaction.

The need for intelligent agents capable of operating in real-world environments is growing. Such agents have the potential to be used in autonomous navigation, consumer robotics, or industry. Large technology companies (Boston Dynamics, Tesla, and others) are working to design autonomous robots to perform dangerous, repetitive, or boring tasks for humans. And to enable robots to interact with the environment, it is first necessary to ensure their ability to perceive and understand it. Therefore, research on the development of effective representations and integration of LLMs in a 3D environment is relevant.

## 2. Literature review and problem statement

In the Embodied Question Answering (EQA) task [18], an agent must interact with the environment to answer questions about it. In contrast to the Visual Question Answering task, the agent must actively explore its environment to gather the necessary information, making the task both a visual and a physical navigation problem. There is an open vocabulary EQA dataset that contains both episodic memory and active exploration episodes. A subset of the dataset episodes was generated based on the ScanNet dataset [19], a carefully annotated dataset of 3D indoor scenes. Furthermore, paper [17] investigated the capabilities of LLMs based on the OpenEQA benchmark and highlighted the limitations of these models compared to human-level. However, scene annotation and frame selection were not investigated in detail. Text descriptions were generated automatically by randomly selecting 50 frames from the video, applying trained (image captioning) models to describe them, and concatenating the resulting descriptions into one.

A series of studies [9–11] report the use of language models in planning tasks. [9] investigates the integration of LLMs with robotic systems for executing high-level instructions in natural language. Namely, the LLM provides high-level knowledge by building a plan for each task, and the robot's value function evaluates the relevance of the available robot skills at each step, providing the necessary basis for performing tasks in specific environments. In [10], a continuation of [9], different sources of feedback – including action success assessment, scene description, and human interaction – are examined to improve the robot's understanding for task execution. By using environmental feedback, the LLM generates an internal monologue that enables real-time plan adjustments, improving task performance. In [11], NLMap is introduced, a framework that enhances robotic task planning by integrating LLMs with a natural language-queryable scene representation. Which allows LLM-based agents to interact with a natural language representation of the environment. An unresolved issue in the above papers is how to represent the 3D environment to an LLM. The LLM is used exclusively for planning, and integration with the environment occurs by combining the most likely step of the plan with the most appropriate available skill. Also, the NLMap representation cannot be interpreted (compared to text) because it is implicit.

With the development of multimodal research, large models have become capable of perceiving the world through the visual modality [2–4, 12–14]. GPT-4V [2–4] at the time of the study demonstrates the best results. Since GPT-4 is a closed model, the technical report [2] only states that its architecture is a large transformer, without details about the exact number of parameters and training information. The document on the authors and their contributions [3], in the context of extending GPT-4 to GPT-4V, mentions data used for fine-tuning and optimizing the model to enhance its ability to understand images alongside text. Which in turn makes it possible to perform tasks such as image annotation, object recognition, and diagram analysis. However, the model's system card [4] provides only illustrative examples of the tasks. [12–14] describe similar approaches for training LLMs to recognize visual data, where a linear layer is trained to project visual embeddings into the LLM's latent space. The authors of MiniGPT-v2 [13] (an improvement on [12]) proposed task-specific training, explicitly indicating the type of task to the model for each training example. Later, in [14], a methodology for improving LLMs with structured visual instructions was introduced, employing a training strategy that optimized the model for queries involving image-based reasoning. And in [15], it was shown that fine tuning on visual instructions improves the leading results on 11 benchmarks. Consequently, improving the capabilities of multimodal models opened the possibility of using a single model as an intelligent agent for perception and decision-making. However, according to [17], LLM-based agents are far from solving the Embodied Question Answering (EQA) task [18].

In summary, the issue of scene representation necessary for achieving human-level EQA performance remains open. Namely, what information about the scene should the representation contain, is it possible to create an "ideal" representation (what would be the length of the context) and how the choice of LLM affects EQA performance. Addressing these questions through effective 3D scene representation methodologies is crucial for enabling LLMs to reach human performance in EQA.

## 3. The aim and objectives of the study

This study aims to explore the potential of using text annotations as a representation of 3D scenes for the Embodied Question Answering (EQA) task. Effective 3D scene representations can enhance Embodied AI systems' ability to perceive the real world, enabling them to make better decisions and answer questions about their environment more accurately.

To achieve the goal of the study, the following tasks were set:

– to investigate ways to create text representations of scenes: to determine the attributes and relationships that should be included in the text representation of a 3D scene to allow EQA to answer OpenEQA questions without mistakes;

– to assess the effectiveness of EQA task performance by both humans and LLMs: examine how different text-based representations of 3D scenes (free and structured forms) affect performance of humans and LLMs on EQA;

– to investigate the capabilities of Vision Language Models (VLMs) to create annotations of 3D scenes.

## 4. The study materials and methods

The object of our research is the techniques for creating textual annotations to accurately represent 3D scenes: attributes and relationships of objects, spatial relations, level of detail, etc. This includes determining the characteristics of "ideal" scene representations, understanding the impact of the choice of LLM on the efficiency of EQA performance, as well as estimating the size of the context for these representations. By "ideal" we mean representations that contain enough information about the scene that, when properly aligned with the LLM, makes it possible to achieve the level of human efficiency in Embodied AI tasks.

The main hypothesis assumes that an ideal textual representation contains enough information about the scene so that the LLM can achieve the level of human efficiency in Embodied AI tasks, in particular EQA. If such a representation can be constructed in textual modality, then it can be used for learning even more efficient implicit representations (for example, artificial neural networks).

The study uses a subset of the ScanNet [19] OpenEQA dataset, which includes 17 video tours of scenes with the largest number of questions (2 questions per category).

The LLM-Match method [17] was used to evaluate the results of the EQA task. It involves assigning each response a score on a five-point scale, where 1 corresponds to the lowest score and 5 to the highest. At the next stage, the score is calculated according to the formula presented in the OpenEQA paper [17]:

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

where $N$ is the number of questions, and $\sigma_i$ is the LLM-Match score for each individual question. The higher the LLM-Match score, the better the method performs the EQA task. To minimize the stochasticity of responses, which is characteristic of LLM and can cause variations between individual experiments, the model is asked the same question several times. This makes it possible to get a set of answer options for each question, which helps increase the accuracy of analysis.

The research begins with the analysis of one scene (0164_02), for which the description (textual representation) is iteratively improved. The goal of this approach is to check the possibility of achieving the maximum score on the LLM-Match scale.

The research is then scaled to 17 scenes, for which independent annotators are involved to create descriptions. First, the annotators are asked to describe the scene in an arbitrary format with the task: "A written description of the scene should be sufficient to answer any question about this scene." This type of annotation is called "free-form scene description." The annotators then create a description of the same scene, following a structured format defined in the single-scene study. This results in two types of text annotations for each of the 17 scenes. In addition, another group of annotators were involved to answer episode questions based on the free-form and structured scene captions, allowing for an assessment of the effectiveness of each approach.

The model study uses LLaMA-2 {7, 70}B and GPT-4 as intelligent agents, as well as GPT-4V for generating image captions. In the first stage, the performance of LLM-based agents (the correspondence between the model responses and the correct responses) for free-form and structured scene descriptions is evaluated, allowing us to compare the results with the scores achieved by humans. Then, these descriptions are used as examples of multi-frame annotations for prompting GPT-4V, to automatically generate scene descriptions. The generated descriptions are used to test the LLM on the OpenEQA task. The results are compared with descriptions obtained without examples, as well as with responses from a Vision Language Model (VLM), which analyzes only images.

To estimate the length of the context (in GPT tokens) required to create an ideal scene description, a formula for calculating the maximum context length is used. This formula makes it possible to determine how many tokens the model needs to efficiently process all relevant details of the scene, ensuring optimal description quality:

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

where $N$ is the number of objects seen in the episode history, $MObj$ is the maximum length of the object description, and $MSpat$ is the maximum length of the object spatial relationship description.

Experiments with the LLaMA-2 7B model were performed on a computer with two NVIDIA GeForce RTX 3090 graphics cards. Experiments with large models LLaMA-2 70B, GPT-4, and GPT-4V were performed using the OpenAI API and Hugging Face API.

## 5. Results of research on the use of text annotations as form of 3D scene representation

### 5. 1. Exploring techniques for creating text representations (for one scene)

The following types of text representations of a scene were considered:

– baseline representation: lists all objects present in the scene, with minimal details about attributes. Spatial relationships are described in simple terms, focusing on basic placement, such as whether an object is "on" or "under" another object;

– general representation: includes detailed information about attributes for each object and basic spatial relationships such as "near", "along" and "next to", to provide a more descriptive overview of the scene;

– absolute representation: like the general representation, this approach includes detailed attributes of objects but also specifies the position of the agent in the room. It introduces absolute directions – "north", "south", "east" and "west" – to more accurately describe the location of objects in a scene;

– relative representation: uses relative spatial landmarks such as "left/right", "nearer" and "further" to formulate relationships between objects;

– refined representation: the relative representation is adjusted based on information about questions that the model could not answer.

Using an iterative approach, the responses of language models to different ways of representing scenes were systematically evaluated, which allowed us to identify key changes in performance depending on changes in the forms of description. The performance indicators of the models in different categories of questions are shown in Fig. 1, and the results averaged over episodes are given in Table 1.

Based on the results of our experiments, a technique for creating text descriptions has been proposed, which implies compliance with the following requirements:

1. List of objects as the basis of the description: the initial stage of creating a description should include a complete list of all objects present in the scene. This provides a

clear understanding of all elements that may be important for interpreting the scene.

2. Clear definition of spatial relationships: after listing the objects, spatial relationships between them should be formulated. This allows us to establish their mutual location.

3. Consistency and alignment of attributes: the attributes of objects used for their identification (for example, size, shape, color) should be carefully selected, their consistency (ordering) should be the same throughout the text. This minimizes the risk of ambiguity and ensures the stability of object identification.

This technique of scene annotation is referred to as "structured description". An example of a structured description for the ScanNet scene with the identifier 0100_02 is shown in Fig. 2.

Table 1

Comparison of LLM EQA results on different scene descriptions (0164_02 scene)

| No. | Method | Baseline | General | Refined |
|---|---|---|---|---|
| 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 |

The structured scene description shown in Fig. 2 resembles a graph structure, in which object descriptions are represented as vertices and descriptions of spatial relationships between them form edges. This description technique, according to the hypothesis, should improve spatial understanding by providing a clearer and more systematic representation of the scene.

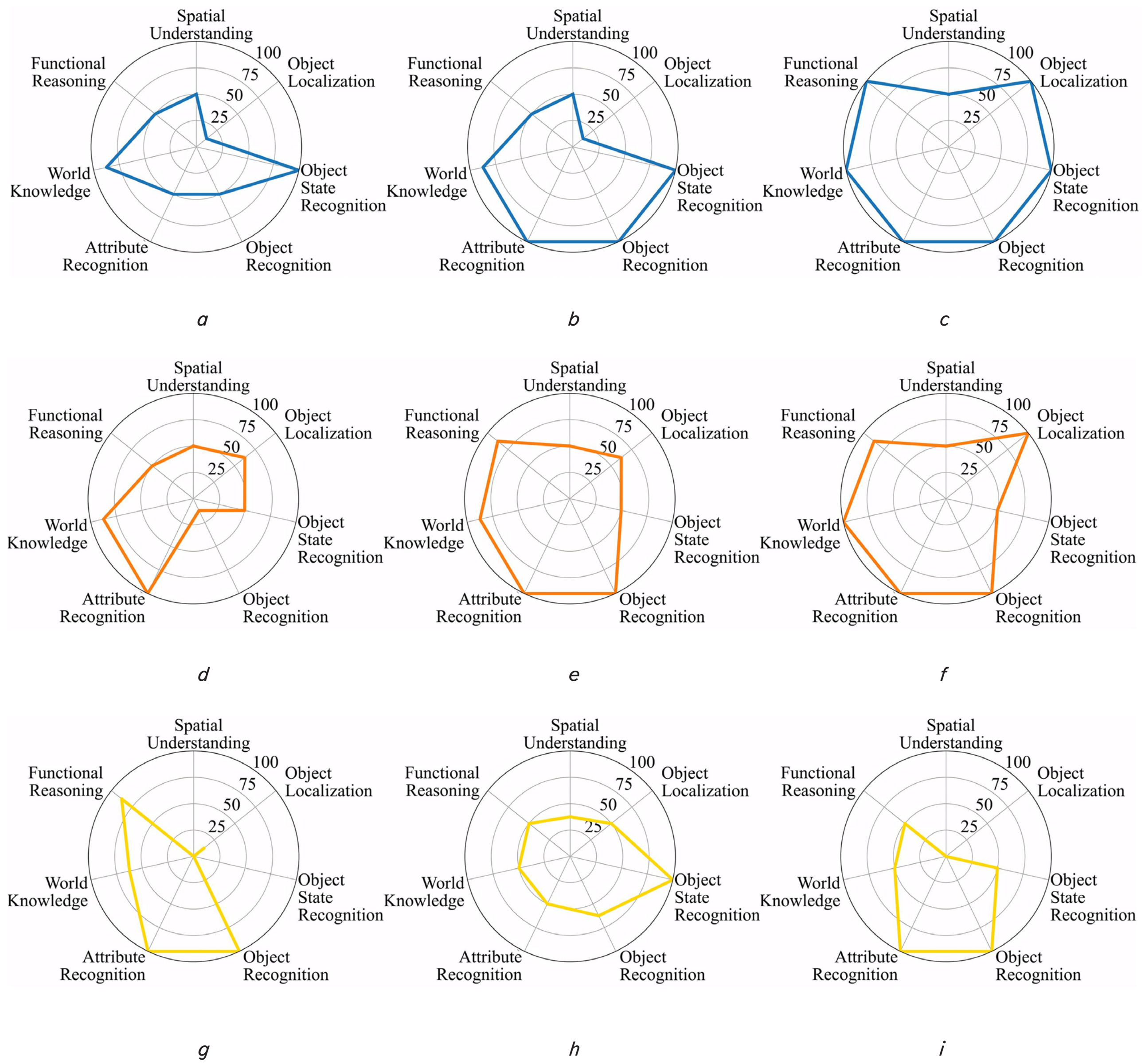
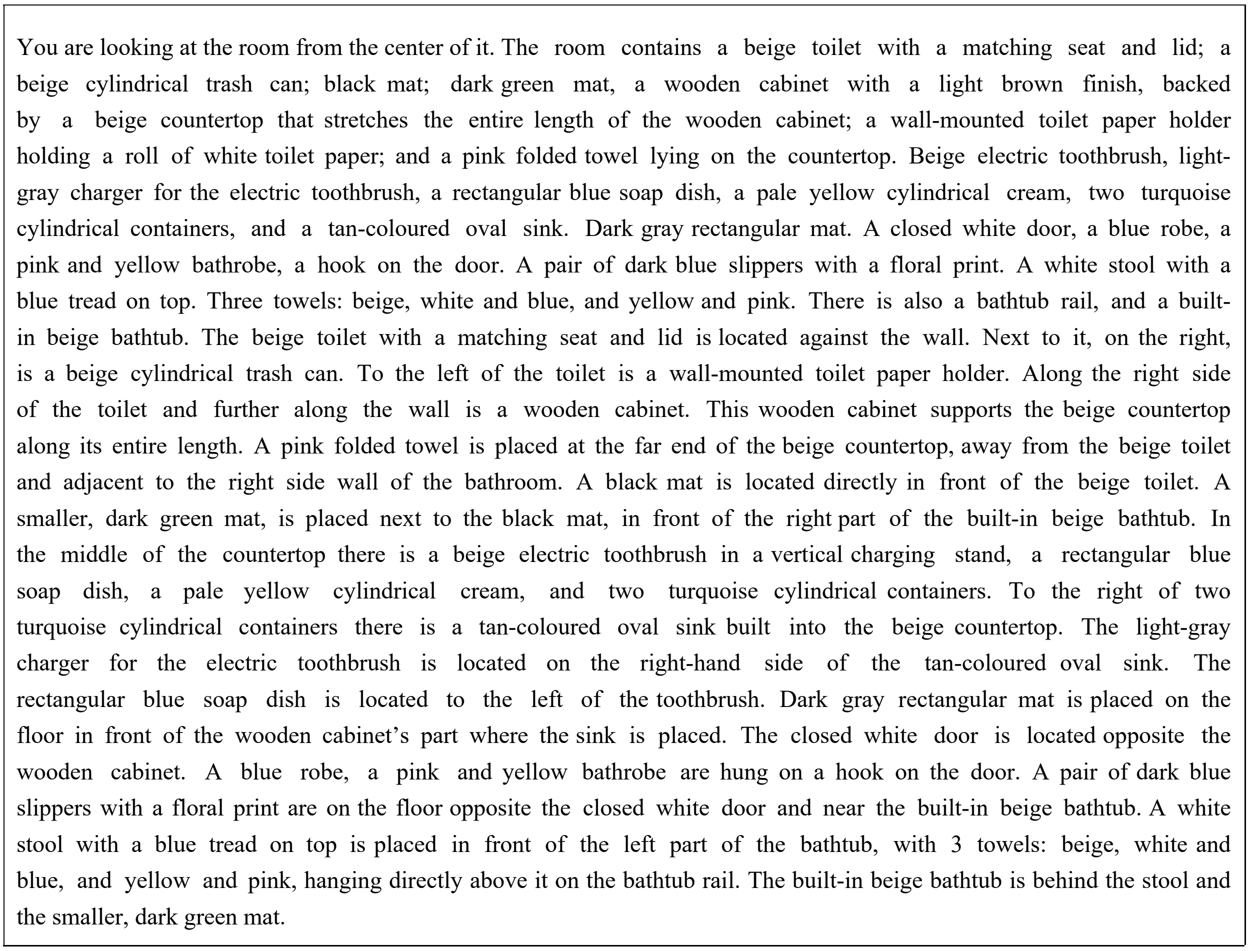


Fig. 1. Scene description design study (one scene): *a* — GPT-4 results using the baseline representation; *b* — GPT-4 results using the general representation; *c* — GPT-4 results using the refined representation; *d* — LLaMA-2 70B results using the baseline representation; *e* — LLaMA-2 70B results using the general representation; *f* — LLaMA-2 70B results using the refined representation; *g* — LLaMA-2 7B results using the baseline representation; *h* — LLaMA-2 7B results using the general representation; *i* — LLaMA-2 7B results using the refined representation



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.

Fig. 2. Scene description (0100_02) in structured form

**5. 2. Assessing the effectiveness of EQA task performance by humans and LLM**

As part of the study on the effectiveness of EQA task performance by humans and LLM, a three-stage human-in-the-loop experiment was conducted. In it, participants performed the same tasks as the models, and the results obtained were used as a reference for comparison with the results of the models.

The experiment included the following stages:

1. Free-form scene description: Participants were asked to describe scenes in a natural form that contained enough detail to answer questions about any object present. The input for creating the description consisted of a video tour of the scene, a 3D scan of the scene, and samples of possible questions about the scene.

2. Structured scene description: Participants were asked to describe the same scenes but following a structured description technique (Section 5. 1).

3. Question answering: Participants were asked to answer questions about the scenes using only the descriptions (created in previous stages). To prevent bias, it was ensured that participants worked with different scenes at each stage. If the information in the descriptions was not enough to answer, the question was allowed to be skipped.

The length of the most detailed 3D scene annotation by a human was 782 tokens (in GPT tokens). The length of the longest object description: "a white-blue-red beer cardboard box with label Samuel Adams, Boston Lager" was 17 tokens. The length of the 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" was 24 tokens. Given that the number of objects is 37 (including 7 surfaces), the approximate length of the detailed description is 1517 tokens (41 tokens per scene object).

The results of the study are illustrated in Fig. 3.

Based on our experiments, the highest metrics among all combinations of descriptions and models – 58.5 % score – were achieved by the LLaMA-2 70B model using free-form descriptions. GPT-4 showed a similar result, achieving 57.6 % score on the same scenes. When structured descriptions were used, the performance of LLMs decreased: LLaMA-2 accuracy dropped to 53.2 %, and GPT-4 accuracy dropped to 50.5 %. People performing the same tasks achieved higher accuracy on structured descriptions (50.7 %) than on free-form descriptions (48.3 %). The most significant improvement was observed in the categories of questions related to spatial aspects, such as object localization, object recognition, and spatial understanding, where structured descriptions increased the average human score by 5 % compared to free-form descriptions.

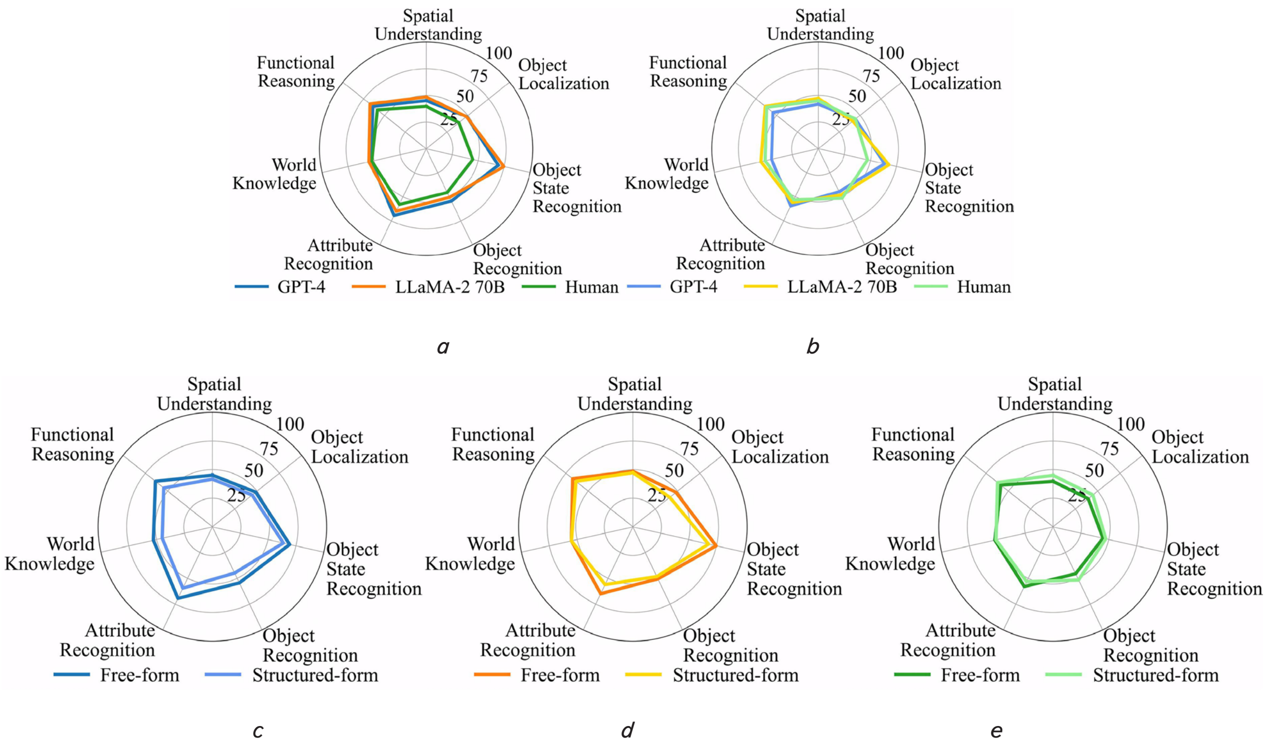
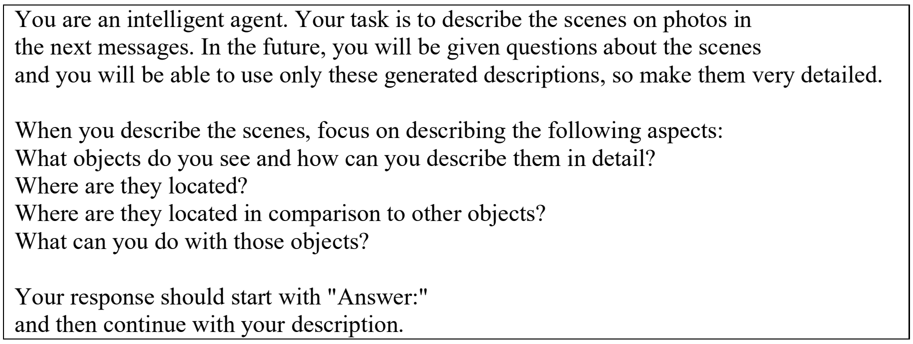


Fig. 3. Results of EQA task performance by humans and LLM on descriptions in free and structured forms: *a* — results of humans and LLMs on descriptions in free form; *b* — results of humans and LLMs on descriptions in structured form; *c* — results of GPT-4 on descriptions in free and structured form; *d* — results of LLaMA-2 70B on descriptions in free and structured form; *e* — results of humans on descriptions in free and structured form

**5. 3. Assessing the ability of VLMs to create annotations of 3D scenes**

GPT-4V was used to create annotations of 3D scenes, and the three best descriptions were used as examples for prompting. Two different approaches were used for frame selection: random uniform sampling and an algorithm based on JPG+SSIM (based on visual similarity determined by the SSIM and JPG structural similarity algorithm). The model received a query (as shown in Fig. 4) along with pairs of frames and corresponding descriptions in free form. Each test began with the keyword "Answer", after which frames from a new episode were provided so that the model would generate a corresponding scene description.



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 these 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 can you describe them in detail?
Where are they located?
Where are they located in comparison to other objects?
What can you do with those objects?

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

Fig. 4. Example of a query

The experiments evaluated two key areas:

1. The performance of the JPG+SSIM frame selection algorithm compared to random uniform sampling.

2. Comparison of description generation with no examples, with one and with two examples on the OpenEQA benchmark.

All experiments were conducted on 15 episodes, each containing 11 questions (different from those used in human study). The following parameters were chosen for the experiments:

– 0, 1, and 2 examples generating descriptions on 5 frames using JPG+SSIM frame selection;

– 1 example for generating descriptions on 10 frames using JPG+SSIM frame selection;

– 1 example for generating descriptions on 10 frames with random uniform sampling.

These parameters allowed us to assess both the impact of the frame selection method and the impact of a small number of examples on the quality of the descriptions and their subsequent performance in EQA tasks.

The frame selection method was based on the uniform selection method, when a predetermined number of frames were evenly selected throughout the entire length of the episode. This approach provided uniform coverage, but did not consider the quality of the content and the uniqueness of the selected frames.

Unlike random uniform sampling, the JPG+SSIM approach focused on selecting the most detailed and visually diverse frames. This approach uses the JPG compression weight as a rating score to identify frames with a high level of detail. In addition, the structural similarity index (SSIM) is used to avoid selecting frames with excessive visual information due to similar views. This approach allows us to create a more complete and representative visual context, ensuring high quality frame selection for further model work. That, in turn, contributes to the generation of better descriptions and increases the accuracy of answers to questions related to episodes. The JPG+SSIM frame selection algorithm is as follows:

1. Split the episode into parts: the frames from the episode are chronologically divided into $n$ parts of equal length, where $n$ is the number of frames to be selected.

2. Rank the frames by JPG weight: within each part, the frames are sorted in descending order of the weight of their JPG-compressed forms (higher weights correspond to more detailed frames).

3. Select visually diverse frames: starting from the first part, the algorithm iteratively goes through the parts, selecting the frame with the highest ranking (by JPG weight) from each, provided that its SSIM score with respect to the previously selected frames is less than 0.4. This threshold ensures that neighboring frames represent different viewpoints, minimizing redundancy.

This algorithm aims to balance detail and diversity, ensuring that the selected frames collectively convey the most important visual information while avoiding similar viewpoints.

The aggregated results of these experiments are given in Table 2.

Table 2 rows 1–2 correspond to experiments without examples, rows 3–10 – with one example, rows 11–12 – with two examples. The results of the experiments show that:

– annotations without examples generally conveyed the most information, with a small performance gap compared to descriptions generated based on a single example;

– descriptions generated from two examples lacked detail and were generally very short, despite being structurally similar to those written by a human;

– the JPG+SSIM frame selection method did not improve the results compared to uniform sampling.

In addition to the quantitative results, the following observations are worth noting:

1. Giving the LLM more frames per episode does not necessarily improve the detail of the captions. For example, in the case of 1 frame with 20 frames in an episode (40 frames total), the model often did not generate captions, defaulting to "I can't assist in this case." When captions were generated, they typically contained a very brief overview of all frames, rather than a detailed description of individual frames or scenes. In contrast, when there were 5 or 10 frames in an episode, the model provided a more detailed frame-by-frame description and an overall overview of the scene.

2. Without clear task-specific instructions, the descriptions generated by the LLM for visually diverse 3D spaces tend to be too general and lacking in the necessary detail. Even when the model was given the attributes it needed to focus on, it only captured a fraction of the necessary detail, but it performed better when it was given targeted instructions.

3. Typical descriptions generated using LLM described objects and relationships for individual frames but could not integrate this information into a holistic representation of the scene.

Table 2

Comparison of results with no examples, with one and two examples using JPG+SSIM: aggregated score over all question categories in LLM-Match column, score per category (attribute recognition, functional reasoning, object localization, object recognition, object state recognition, spatial understanding and world knowledge) in columns entitled with corresponding category acronym

| No. | Method | N | Frame selection | AR | FR | OL | OR | OS | SU | WN | LLM-Match |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | GPT-4 | 5 | JPG+SSIM | 51.7 | 52.6 | 42.4 | 36.2 | 73.3 | 33.6 | 50.4 | 48.3±0.8 |
| 2 | LLaMA-2 | 5 | JPG+SSIM | 51.7 | 44.7 | 38.8 | 39.4 | 83.7 | 40.7 | 50.4 | 50.1±0.6 |
| 3 | GPT-4 | 5 | JPG+SSIM | 57.0 | 59.6 | 32.2 | 26.9 | 77.0 | 31.1 | 44.8 | 46.5±0.5 |
| 4 | GPT-4 | 5 | Uniform | 37.0 | 56.6 | 40.9 | 26.9 | 84.7 | 33.0 | 54.4 | 47.0±0.1 |
| 5 | GPT-4 | 10 | JPG+SSIM | 46.7 | 50.9 | 31.2 | 36.2 | 83.0 | 39.1 | 56.0 | 48.8±0.2 |
| 6 | GPT-4 | 10 | Uniform | 50.0 | 50.9 | 35.5 | 26.9 | 71.0 | 29.8 | 55.2 | 45.1±0.4 |
| 7 | LLaMA-2 | 5 | JPG+SSIM | 51.7 | 47.8 | 24.6 | 34.0 | 70.3 | 29.2 | 48.8 | 43.6±0.9 |
| 8 | LLaMA-2 | 5 | Uniform | 41.0 | 57.5 | 30.8 | 23.4 | 95.0 | 44.2 | 54.8 | 49.1±0.2 |
| 9 | LLaMA-2 | 10 | JPG+SSIM | 42.0 | 57.0 | 34.1 | 27.6 | 78.7 | 31.7 | 50.0 | 45.3±0.5 |
| 10 | LLaMA-2 | 10 | Uniform | 49.0 | 49.1 | 29.0 | 21.8 | 90.7 | 35.6 | 61.9 | 47.8±0.1 |
| 11 | GPT-4 | 5 | JPG+SSIM | 40.0 | 53.1 | 38.0 | 31.7 | 75.3 | 35.9 | 47.6 | 45.6±0.8 |
| 12 | LLaMA-2 | 5 | JPG+SSIM | 48.3 | 45.2 | 26.1 | 27.6 | 80.0 | 27.2 | 54.8 | 43.9±0.7 |

The results of the studies are summarized in Table 3. Rows 1–3 correspond to experiments with free-form annotations, rows 4–6 to structured-form annotations, rows 7–8 to annotations generated by large language models, row 9 to the results of the GPT-4V language-visual model on 50 images of the scene, and 10 to the results achieved by humans with visual data.

Table 3 allows us to compare the results of the EQA task performed by large language models and humans on scene descriptions in arbitrary (rows 1–3) and structured (rows 4–6) forms. And for descriptions in structured form, the results of large language models on descriptions created by annotators (rows 4, 5) and generated by GPT-4V with scene images as input data (rows 7, 8).

Table 3

Research results (* – results reported in the OpenEQA paper [10]): aggregated score over all question categories in LLM-Match column, score per category (attribute recognition, functional reasoning, object localization, object recognition, object state recognition, spatial understanding and world knowledge) in columns entitled with corresponding category acronym

| No. | Method | AR | FR | OL | OR | OS | SU | WN | LLM-Match |
|---|---|---|---|---|---|---|---|---|---|
| 1 | GPT-4 | 69.4 | 63.7 | 48.5 | 54.4 | 69.4 | 45.1 | 52.9 | 57.6±0.5 |
| 2 | LLaMA-2 | 64.9 | 67.4 | 48.3 | 50.5 | 74.8 | 48.5 | 55.1 | 58.5±0.7 |
| 3 | Human | 58.0 | 67.4 | 48.3 | 50.5 | 74.8 | 48.5 | 55.1 | 58.5±0.7 |
| 4 | GPT-4 | 59.5 | 54.4 | 44.6 | 44.9 | 63.7 | 41.6 | 45.0 | 50.5±1.5 |
| 5 | LLaMA-2 | 56.1 | 63.7 | 41.3 | 48.2 | 67.9 | 47.2 | 48.2 | 53.2±0.6 |
| 6 | Human | 53.2 | 62.1 | 44.5 | 51.6 | 47.3 | 44.9 | 51.2 | 50.7 |
| 7 | GPT-4 w/ GPT-4V | 54.6 | 56.4 | 38.0 | 41.4 | 71.3 | 38.7 | 49.0 | 49.9±0.2 |
| 8 | LLaMA-2 w/GPT-4V | 56.3 | 59.3 | 32.4 | 41.9 | 66.7 | 37.0 | 47.3 | 48.7±1.9 |
| 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 |

## 6. Discussion of results on the use of text annotations as form of 3D scene representation

Detailed descriptions, which include a list of objects, clearly characterize their attributes and spatial relationships, provide advantages in EQA performance: 92.75 % versus 57 % compared to the basic representation for GPT-4 and 84 % versus 59 % compared to the basic representation for LLaMA-2 70B (Table 1). The choice of LLM affects the efficiency of the EQA task. Larger models answer more questions correctly, which indicates their superiority over smaller models, for example, LLaMA-2 70B>LLaMA-2 7B (Table 1, Fig. 2). Also, smaller models are not able to distinguish between different ways of textual representation of a 3D scene. LLaMA-2 7B on average answers the same number of questions for the three different representation techniques (row 3, Table 1, column 3, Fig. 2).

The length of the detailed structured description, an example of which is shown in Fig. 2, is estimated as 1517 tokens (in GPT tokens), which is 2 times longer than the length of descriptions generated by VLM by default. The main explanation for this is that complex tasks, such as determining spatial relationships and attributes, require a significantly larger amount of information to reduce ambiguity in the model response. Given the annotations created by humans as input data, the VLM outperforms the best method in OpenEQA [17] – GPT-4V: 57.6 % and 58.5 % for GPT-4 and LLaMA-2, respectively (rows 1, 2, Table 3) versus 55.3 % for GPT-4V (in OpenEQA [17]). However, even a detailed text representation significantly limits the efficiency of the EQA task. Humans show significantly higher results when working with video episodes compared to text descriptions: approximately 50 % accuracy for text versus 86.8 % for video (rows 3 and 6 versus row 10, Table 3). This difference indicates that even detailed human-generated text annotations have limitations in conveying the full picture of the scene. Thus, the integration of visual data becomes key for models to fully understand and interact with the 3D environment. This highlights the importance of combining text and visual modalities to achieve maximum efficiency in EQA tasks. Nevertheless, the results obtained are better than the best performing methods reported in the OpenEQA paper [17], in which the highest accuracy achieved using automatically generated captions using the LLaVA-1.5 model with GPT-4 was 45.1 %. In contrast, the proposed approach using human-generated captions exceeded 57 % accuracy with both LLaMA-2 and GPT-4 (Fig. 3, rows 1, 2, Table 3), which highlights the superiority of human-generated descriptions over automatically generated captions or scene graphs.

When the GPT-4V was used to generate scene annotations, for prompts with no examples and with one example (Table 2), the model sometimes returned a reasonably coherent description of the scene. Although the generated descriptions resembled those written by humans, they still lacked the detail to outperform OpenEQA [17]. For prompts with two examples, the generated description was always very short. The JPG+SSIM frame selection method also did not improve the generated descriptions (Table 2). These results suggest that although VLMs can mimic the structure of human-generated descriptions, they struggle aggregating information from individual frames to create a detailed and coherent representation. After reaching ~50 % correct answers, the model performance reaches a plateau: further increases in model size or complexity yield little improvement. Some models may be slightly better in individual question categories, but on average their performance remains quite close. For example, LLaMA-2 70B and GPT-4 show similar results (Table 3). These observations highlight the importance of not only model size but also optimization of its architecture and training algorithms to achieve high results in the EQA task.

As regards limitations, it is worth noting that the results of the study depend largely on the level of detail of the text annotations, whether created manually or automatically. In practice, obtaining detailed descriptions for complex scenes can be resource- and time-consuming. Also, larger models improve performance, but only up to a certain saturation point, indicating the need for other approaches, such as those that can work with multimodal input data. In addition, the study was conducted on a subset of OpenEQA, which limits its generalization to other problems or environments.

Further development of the work involves the following:

1. Integration of multimodal data: combining textual representations with visual and sensory data would allow us to achieve a more complete understanding of 3D scenes.

2. Expanding the scope of testing by utilizing diverse datasets and environments to assess the generalizability of the results obtained, including evaluations on dynamic or real-world scenes.

Development of these areas could help overcome the identified limitations and make our method more suitable for wide practical application (for example, running on a robotic platform, virtual or augmented reality glasses). In addition to the direct use of textual annotations of scenes as input data to the planning module (virtual or physical agent), it could also be combined with other forms of 3D representations (point cloud, video, tuples of photos of the scene and camera positions and orientations, etc.). The main advantage of the text modality is its interpretability, ease of understanding by humans, and the ability to be combined with or converted into other explicit or non-explicit 3D representations.

## 7. Conclusions

1. In one-scene experiments, the best results among all considered techniques for creating a text representation of a 3D scene were achieved with the refined structured descriptions (92.75 % vs. 57 % compared to the baseline representation for GPT-4 and 84 % vs. 59 % compared to the baseline representation for LLaMA-2 70B). Structured descriptions start with a list of all objects in the scene, include detailed information about their attributes and location relative to other objects. Such representations resemble a graph structure, in which descriptions of objects and their attributes are vertices, and descriptions of spatial relationships between them are edges. It has also been confirmed that larger models, such as LLaMA-2 70B, allow better performance on tasks requiring complex spatial understanding, which highlights the dependence of results on the complexity of the model (84 % LLaMA-2 70B vs. 50 % LLaMA-2 7B).

2. Analysis of the work of our experiment's participants reveals that humans demonstrate better results in spatial understanding tasks compared to language models (participants who used structured scene descriptions answered 5 % more questions correctly on average). The experiment results showed that the GPT-4 and LLaMA-2 models achieve high accuracy when using structured descriptions, outperforming previous methods (53.2 % for LLaMA-2 and 50.5 % for

GPT-4). It should also be noted that on text representations of scenes, the models achieve results on the EQA task even higher than humans (58.5 % vs. 50.7 %). However, the results of both models and humans on text representations of scenes remain lower than the results of humans with videos (86.8 %).

3. The (GPT-4V) VLM can generate annotations of 3D scenes. The generated descriptions resembled those written by a person, but they lacked the detail to surpass the results by OpenEQA [17]. LLMs that received as input the VLM-generated scene annotations did not give more than 50 % correct answers. Prompting with examples and a better algorithm for extracting frames (JPG+SSIM) from the video for scene description did not improve the quality of the annotations.

## Conflicts of interest

The authors declare that they have no conflicts of interest in relation to the current study, including financial, personal, authorship, or any other, that could affect the study, as well as the results reported in this paper.

## Funding

The study was conducted without financial support.

## Data availability

The manuscript has associated data in the data warehouse.

## Use of artificial intelligence

The authors confirm that they did not use artificial intelligence technologies when creating the current work.

## References

1. Vaswani, A., Shazeer, N. M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N. et al. (2017). Attention is All you Need. NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems. Available at: https://dl.acm.org/doi/10.5555/3295222.3295349
2. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L. (2023). Gpt-4 technical report. ArXiv. https://doi.org/10.48550/arXiv.2303.08774
3. Gpt-4v(ision) technical work and authors. OpenAI. Available at: https://openai.com/contributions/gpt-4v
4. Gpt-4v(ision) system card. OpenAI. Available at: https://openai.com/index/gpt-4v-system-card/
5. Yang, Z., Li, L., Lin, K., Wang, J., Lin, C., Liu, Z., Wang, L. (2023). The Dawn of LLMs: Preliminary Explorations with GPT-4V(ision). ArXiv. https://doi.org/10.48550/arXiv.2309.17421
6. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y. et al. (2023). Llama 2: Open Foundation and Fine-Tuned Chat Models. ArXiv. https://doi.org/10.48550/arXiv.2307.09288
7. Annepaka, Y., Pakray, P. (2024). Large language models: a survey of their development, capabilities, and applications. Knowledge and Information Systems, 67 (3), 2967–3022. https://doi.org/10.1007/s10115-024-02310-4
8. Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S. et al. (2022). Emergent Abilities of Large Language Models. ArXiv. https://doi.org/10.48550/arXiv.2206.07682
9. Ahn, M., Brohan, A., Brown, N., Chebotar, Y., Cortes, O., David, B. Et al. (2022). Do As I Can, Not As I Say: Grounding Language in Robotic Affordances. Conference on Robot Learning. ArXiv. https://doi.org/10.48550/arXiv.2204.01691
10. Huang, W., Xia, F., Xiao, T., Chan, H., Liang, J., Florence, P. et al. (2023). Inner monologue: Embodied reasoning through planning with language models. ArXiv. https://doi.org/10.48550/arXiv.2207.05608
11. Chen, B., Xia, F., Ichter, B., Rao, K., Gopalakrishnan, K., Ryoo, M. S. et al. (2023). Open-vocabulary Queryable Scene Representations for Real World Planning. 2023 IEEE International Conference on Robotics and Automation (ICRA). https://doi.org/10.1109/icra48891.2023.10161534
12. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M. (2023). MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. ArXiv. https://doi.org/10.48550/arXiv.2304.10592
13. Chen, J., Zhu, D., Shen, X., Li, X., Liu, Z., Zhang, P., Krishnamoorthi, R. et al. (2023). MiniGPT-v2: large language model as a unified interface for vision-language multi-task learning. ArXiv. https://doi.org/10.48550/arXiv.2310.09478
14. Liu, H., Li, C., Wu, Q., Lee, Y. J. (2023). Visual instruction tuning. NIPS '23: Proceedings of the 37th International Conference on Neural Information Processing Systems. Available at: https://dl.acm.org/doi/10.5555/3666122.3667638
15. Liu, H., Li, C., Li, Y., Lee, Y. J. (2024). Improved Baselines with Visual Instruction Tuning. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 26286–26296. https://doi.org/10.1109/cvpr52733.2024.02484
16. Driess, D., Xia, F., Sajjadi, M. S. M., Lynch, C., Chowdhery, A., Ichter, B. et al. (2023). PaLM-E: An embodied multimodal language model. ICML'23: Proceedings of the 40th International Conference on Machine Learning. Available at: https://dl.acm.org/doi/10.5555/3618408.3618748
17. Majumdar, A., Ajay, A., Zhang, X., Putta, P., Yenamandra, S., Henaff, M. et al. (2024). OpenEQA: Embodied Question Answering in the Era of Foundation Models. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 16488–16498. https://doi.org/10.1109/cvpr52733.2024.01560
18. Das, A., Datta, S., Gkioxari, G., Lee, S., Parikh, D., Batra, D. (2018). Embodied Question Answering. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2135–213509. https://doi.org/10.1109/cvprw.2018.00279
19. Dai, A., Chang, A. X., Savva, M., Halber, M., Funkhouser, T., Niessner, M. (2017). ScanNet: Richly-Annotated 3D Reconstructions of Indoor Scenes. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2432–2443. https://doi.org/10.1109/cvpr.2017.261