A Cognitive Evaluation Benchmark of Image Reasoning and Description for Large Vision Language Models

Anonymous ACL submission

Abstract

Large Vision Language Models (LVLMs), despite their recent success, are hardly tested for their cognitive abilities. Inspired by the prevalent use of the "Cookie Theft" task in human cognition test, we propose a novel evaluation benchmark to evaluate high-level cognitive ability of LVLMs using images with rich semantics [1]. It defines eight reasoning capabilities and consists of an image description task and a visual question answering task. Our evaluation on well-known LVLMs shows that there is still a large gap in cognitive ability between LVLMs and humans.

1 Introduction

Recently, with the emergence of Large Language Models (LLMs) such as GPT-4 (OpenAI, 2023), the cognitive abilities of language models have reached a new level (Zhuang et al., 2023). They demonstrated remarkable performance in many tasks (Bubeck et al., 2023). In Vision Language (VL), several researchers (Zhu et al., 2023b; Liu et al., 2023b; Ye et al., 2023) endeavor to boost Vision Language Pre-trained Models (VLPMs) by integrating powerful LLMs (Touvron et al., 2023; Chiang et al., 2023), referred to as Large Vision-Language Models (LVLMs). With LLM as the "brain", the cognitive abilities of LVLMs are also improved and more challenging tasks, such as table, chart, and document reasoning etc, can be solved (Yang et al., 2023). Some state-of-the-art LVLMs, such as GPT-4V (OpenAI, 2023), are progressing towards human-level cognitive abilities. Thus, there's an increasing interset in evaluating the cognitive abilities of LVLMs. Though some LVLM evaluation benchmarks, such as MME (Fu et al., 2023), MMBench (Liu et al., 2023c), SEED Bench (Li et al., 2023a), etc., also consider cognitive reasoning ability as one aspect of their evaluation, they do not provide a comprehensive evaluation of higher-level reasoning ability via a concentrated task and most of the images they use require relatively little reasoning to understand. However with humans, we have such tasks like Cookie Theft picture description, which is broadly used to assess one's cognitive function.

Among different picture description tasks, the Cookie Theft picture description task (Figure 1) from the Boston Diagnostic Aphasia Examination (Goodglass et al., 2001) has dominated work in speech-language pathology (Cummings, 2019). Surprisingly, this picture has been developed and utilized for more than 50 years and continues to be applied and discussed today. The success of Cookie Theft has sparked our curiosity to investigate the magic behind this picture and explore ways to apply this human-oriented cognitive ability assessment to evaluate cognitive abilities of LVLMs.

Cummings (2019) analyzed the reasons for the success of of Cookie Theft. It was found that it contains information of varying degrees of salience and contains more semantic categories. It was also found that during describing Cookie Theft, compared to individuals with cognitive impairments, healthy people tend to demonstrate their coginitive ability by reasoning. For instance, in Figure 1, by comparing the descriptions produced by the healthy man and the woman with probable Alzheimer's dementia, we can identify the following differences:

- The description produced by a healthy man uses "mother" instead of "lady", indicating the reasoning about character relationship.
- The healthy man used "stealing cookies" instead of "taking cookies", indicating his rea-

[1]The code and part of the data is temporarily released at: https://anonymous.4open.science/r/CogBench-D1E7, while the full version of the code and data will be released after the paper is accepted.

[2]The two samples are from DementiaBank (https://dementia.talkbank.org/)

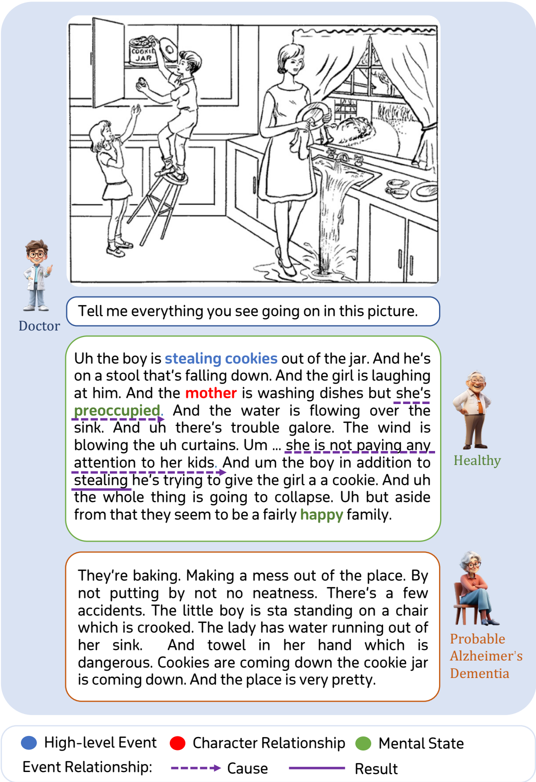

Figure 1: Cookie Theft picture description task. The descriptions in the green frame and the orange frame were respectively produced by a 75-year-old healthy man and a 66-year-old woman with probable Alzheimer's dementia[2].

soning about this high-level event. The description produced by the patient even did not mention this event at all.

- The healthy man also used "the mother is preoccupied" and "happy" to describe people's mental state.

- The description also reflects the causal relationships between events. Because "the mother is preoccupied" and "not paying attention to her kids", "kids are stealing cookies."

Through these reasoning processes, the difference of cognitive abilities between the two individuals is reflected in their descriptions.

Tasnim et al. (2022) introduced guidelines for drawing pictures similar to Cookie Theft, which is generally consistent with findings mentioned above. Generally speaking, compared to typical images, Cookie Theft-like images feature a prominent story theme, richer content, display relationships between components and require stronger cognitive abilities to understand and describe.

The above studies provided us with directions for collecting Cookie Theft-like images and developing assessment criteria to construct a cognitive evaluation benchmark, so that the idea of this human cognition assessment can be leveraged to evaluate cognition of LVLMs comprehensively. To fill this research gap, we propose to construct a **Cog**nitive Evaluation **Bench**mark, named as CogBench, which consists of high quality Cookie Theft-like images to evaluate the cognitive reasoning ability of LVLMs. CogBench defines eight core cognitive reasoning capabilities, including reasoning about special time, location, character, character relationship, event, event relationship, next moment event and mental state. Both a generative Image Description task and a discriminative Visual Question Answering (VQA) task are designed.

Our main contributions are as follows:

- To the best of our knowledge, this is the first-of-its-kind attempt to incorporate the cognitive test Cookie-Theft picture description task designed for humans into the evaluation of LVLMs.

- We created the first ever and the largest to-date VL dataset with Cookie Theft-like images to evaluate LVLM's cognitive abilities.

- Our evaluation on existing well-known LVLMs shows that there is still a large gap between the cognitive ability of LVLMs and human beings, indicating CogBench is a challenging evaluation benchmark.

2 Dataset Construction

In this section, we will introduce the construction of CogBench. We will first introduce image collection, annotation and the statistics of CogBench. Then, we will introduce tasks in CogBench.

2.1 Image Collection

Based on findings of previous studies (Cummings, 2019; Tasnim et al., 2022), we set the following image collection criteria:

- **Rule 1: Stroy-telling** The image depicts an interesting story clearly. For instance, the Cookie Theft picture tells the story of a mother busy washing dishes while two kids takes the opportunity to stand on a stool and sneakily steal cookies.

- **Rule 2: Rich Chain-of-Reasonings** Images should display rich Chain-of-Reasonings (CoRs) in a scene. A CoR connects low-level observations in an image to produce a high-level reasoning conclusion or connects the cause and effect of events. For example, "The mother is busy washing dishes. + The boy is standing on the stool behind the mother. + The girl standing by the boy is shushing him. + The boy is fetching cookies from the jar in the carbinet. → The boy and girl are stealing cookies." is a CoR about high-level event "stealing cookies". Note that the story is actually constructed via these CoRs.

- **Rule 3: Content Complexity Restriction** Images should contain rich content but not be overly complex. The number of subjects should be limited for better emphasis on the key points of the story, facilitating description by humans or models.

Figure 2: The comparison between our images and those from the previous visual reasoning tasks. Compared to our images, image (a) has fewer entities and CoRs, image (b) and (c) have some entities, but fewer CoRs.

With the above criteria, we aim to select high-quality images for cognition evaluation of LVLMs. Most of the images are manully collected from Pinterest[3], and the Cookie Theft is also included in CogBench. Figure 2 illustrates the differences between our images and those from other datasets.

2.2 Image Annotation

Human annotators are hired to annotate the collected images. As shown in Figure 3, the annotation includes three parts: [Entities], [CoRs] and [Description].

By annotating [Entities] and [CoRs], we aim to evaluate the low-level recognition ability and high-level cognitive reasoning ability of models respectively based on description. [Description] is annotated as the reference description for the image. The three parts are annotated in a sequential order.

Entity Annotation We ask annotators to list as many [Entities] in the image as possible and entities that are difficult to recognize should be omitted.

CoR Annotation In order to evaluate model cognition in a fine-grained manner, the following eight reasoning capabilities are annotated with CoRs:

- **Special Time Reasoning**: reasoning about the special time of the story in the image, e.g. festivals, seasons etc.

- **Location Reasoning**: reasoning about the location of the story in the image, e.g. near the school etc.

- **Character Reasoning**: reasoning about the character of the subjects in the image, e.g. police officer etc.

- **Character Relationship Reasoning**: reasoning about relationship between characters in the image, e.g. "the woman is the mother of the kids."

- **Event Reasoning**: reasoning about high-level events happened in the current moment and previous moments in the image based on clues in the picture. The difference between high-level and low-level lies in how much semantic information the event contains, e.g. "stealing cookies" is a higher-level event compared to "taking cookies" as it additionally conveys the semantic of "taking advantage without permission or knowledge."

- **Event Relationship Reasoning**: reasoning about causal and temporal relationship between different events in the image. For instance, "the sink is overflowing **because** the mother left the tap on."

- **Next Moment Event Reasoning**: reasoning about the event that will happen in the next moment. For example, "The police officer will reprimand the boy who violates the rules."

[3]https://www.pinterest.com/

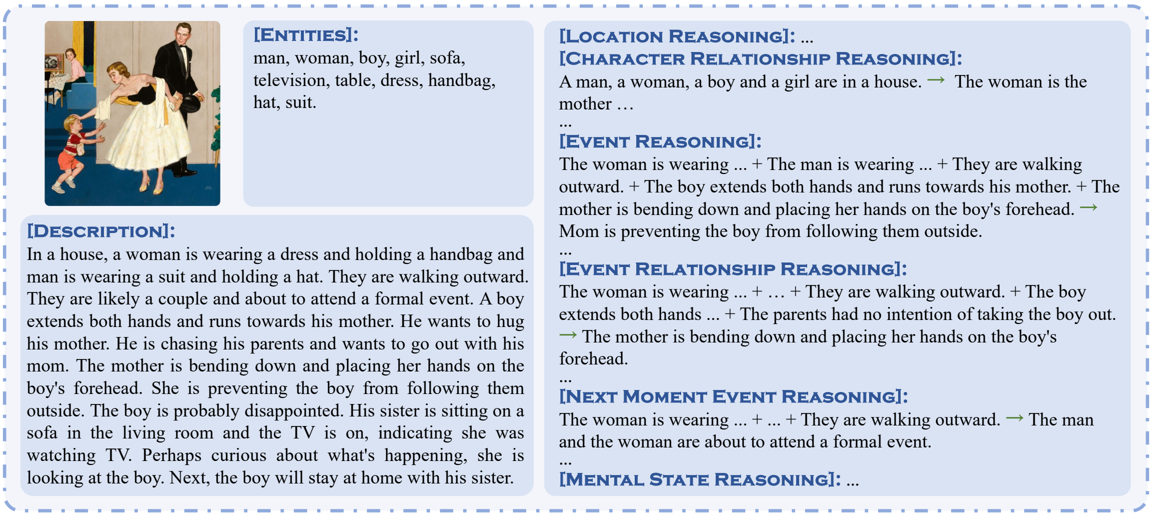

Figure 3: An example of Description task from CogBench.

- **Mental State Reasoning**: reasoning about mental state of subjects in the image, e.g. daydreaming etc.

Description Summary Annotators are last asked to write a description to tell the story in the image based on the annotated [Entities] and [CoRs]. The annotation instruction for annotators is shown in Appendix A.

Considering different people may have different understanding about some images, we ask three annotators to annotate each image. Then, we manually merge the three annotations as one final annotation based on the principle of majority vote. For the [Description], if two or more annotators understand the story in the image in the same way, we accept one of the best descriptions as the final description. For [Entities] and [CoRs], we accept the entities and CoRs that are annotated by at least two annotators.

2.3 CogBench Statistics

To build CogBench, we have gathered 95 images with 1041 entities, 881 CoRs, 95 descriptions and 1091 questions. Table 1 shows the distribution of these CoRs and questions. Though the number of images in CogBench is not large, the content contained in each picture is very rich. The number of event-related CoRs and [Mental State Reasoning] is large, which is a manifestation of the rich interesting stories in the images.

2.4 Tasks in CogBench

To comprehensively evaluate cognitive abilities of LVLMs, we designed a generative Image Description task and a discriminative Multiple-Choice Question Answering task in CogBench.

2.4.1 Image Description Task

Inspired by Cookie Theft picture description task, we propose to assess the cognitive abilities of LVLMs through an Image Description task using Cookie Theft-like images. Recently, some researchers (Xie et al., 2022; Zhu et al., 2023a; Zhuge et al., 2023) are also trying to improve model performance on image description task. The difference between our description task and existing image description task is that we expect LVLMs understand and describe the story in the image through high-level cognitive reasoning. For instance, in Figure 3, the description of the image should not only include what is in the picture but also focus on elucidating the story of "parents are going out to attend a formal event, and the mother is refusing the boy to accompany them" through a series of reasonings.

2.4.2 Visual Question Answering Task

All questions in VQA task are Multiple-Choice Questions with four options, which simplifies evaluation. Similar to Description task, the questions in VQA task are also related to high-level cognitive reasoning. For example, in Figure 4, we can ask question about the event "What is the woman doing?" or question about the cause of the event "Why is the mom placing her hands on the boy's forehead?", which require models to answer with reasoning.

We use a semi-automatic GPT-assisted approach to generate questions for images based on CoRs

	Time	Location	Character	Character Relationship	Event	Event Relationship	Mental State	Next Moment Event
CoR	23	79	42	102	245	172	162	56
QA	47	94	70	159	258	168	221	74

Table 1: Distribution of CoRs and questions in CogBench.

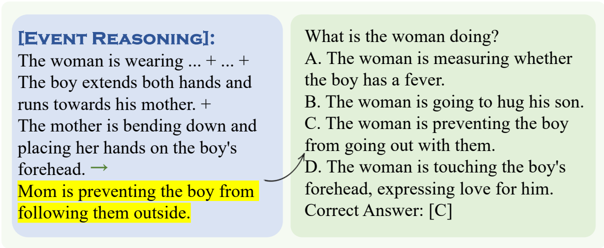

Figure 4: An example of VQA task from CogBench.

we annotated in the previous section, referred to as CoR-based GPT-assisted Question Generation. The main idea of the question generation is that for each CoR, we can ask questions about both the conclusion (right part of →) and reasons to draw the conclusion (left part of →), as shown in Fiure 4. The process can be divided into two stages.

In the first stage, we use GPT-4 (OpenAI, 2023) to generate questions for images in CogBench. For each category of reasoning capability, we design a sepcific prompt for GPT-4 to generate questions, so that questions related to different types of CoRs can be generated. For Question Generation, the key point is to generate high quality distractors. Thus, we encourage GPT-4 to hallucinate to generate more perplexing distractors in our prompt. Appendix B shows an example of prompt we use for GPT-4 to generate questions.

Though GPT-4 is powerful, it is still difficult for it to generate 100% satisfying questions for all images. Thus, in the second stage, we manually select and modify the questions generated by GPT-4. The principle of selection and modification is that distractors should be as closely related to the correct option as possible, thereby possessing a certain level of misleadingness. During selection, ChatGPT is also utilized to help detect those simple questions that can be easily answered without having to accept the image as input.

3 Evaluation Strategy of CogBench

In this section, we will introduce the evaluation strategy of tasks in CogBench.

3.1 Evaluation for Description Task

For the evluation of Description task, we consider model performance comprehensively on two levels: low-level Recognition ability and high-level Cognition ability. Evaluation metrics for both levels are calculated based on recall score, referred to as Recognition Score and Cognition Score, respectively.

The **Recognition Score** is calculated as the ratio of the number of recognized [Entities] to the number of annotated [Entities] in all images.

First, we use SpaCy[4] to extract nouns from the model generated description, and then calculate cosine similariy between embeddings[5] of annotated [Entities] and extracted nouns. For each entity, if the cosine similarity score between the entity and any noun is greater than a threshold (0.5 in this paper), we consider the entity is recognized by the model.

For **Cognition Score**, We first calculate a score for each of the eight cognitive reasoning abilities, and then compute an overall score.

We use GPT-4 (OpenAI, 2023) to help calculate the Cognition Score. To avoid the interpretability issues of GPT-based evaluation, instead of using a subjective evaluation method of directly comparing two descriptions, we use GPT-4 to perform a easier, more objective and fine-grained binary classification task, that is, judging whether the generated description contains the semantics of annotated CoRs. Details and prompts of this GPT-based evaluation are shown in Appendix C.

After obtaining the score of each CoR, we calculate a recall score for each reasoning capabilities. The overall Cognition Score is calculated by summing up the scores of all CoRs and dividing by the total number of CoRs. The effectiveness analysis of GPT-based cognition evaluation is shown in Appendix F.

[4]https://spacy.io/

[5]Implemented with sentence-transformers package (https://www.sbert.net) and *all-mpnet-base-v2* is adopted as the model to encode [Entities] and nouns.

3.2 Evaluation for VQA Task

For multiple-choice questions in VQA task, we use accuracy as the evaluation metric. As questions are generated based on CoRs, we can also calculate both the accuracy for each reasoning capability and the overall accuracy.

4 Experiments

In this section, we will introduce selected LVLMs for evaluation, experiment results and analysis.

4.1 Large Vision Language Models

InstructBLIP-7B (Dai et al., 2023), Qwen-VL-Chat (Bai et al., 2023), LLaVA-v1.5-7B and LLaVA-v1.5-13B (Liu et al., 2023a), mPLUG-Owl-2 (Ye et al., 2023), ShareGPT4V-7B and ShareGPT4V-13B (Chen et al., 2023a) and GPT-4V (OpenAI, 2023) are selected in this paper to be evaluated using CogBench. A brief introduction of these models is shown in Appendix D.

4.2 Results of Description Task

We prompt the selected LVLMs with the following instruction to obtain detailed descriptions about images in CogBench.

Describe this image in detail.

Then, we evaluate the performance of LVLMs on Description task in terms of both recognition and cognition ability.

As a reference, we also calculated traditional image captioning evaluation metrics by camparing annotated reference description and model-generated description, and details are shown in Appendix E.

4.2.1 Recognition

Table 2 shows the Recognition Score of models on Description task. GPT-4V achieves the best performance in terms of recognition, which means it can recognize and describe more entities. Besides, it is significantly better than other open-source LVLMs, indicating open-source LVLMs still have some room for development before reaching the recognition capability of GPT-4V. ShareGPT4V-7B and ShareGPT4V-13B achieves better performance than other open-source LVLMs. As it follows the design of LLaVA v1.5, one possible reason is that ShareGPT4V uses a high-quality image-text dataset featuring highly descriptive captions for training, which makes it describe more entities shown in images. Though GPT-4V achieves the best performance, there are still a lot of entities that are not recognized by GPT-4V, indicating a room for improvement.

Model	Recognition Score
InstructBLIP-7B	0.50
Qwen-VL-Chat	0.54
LLaVA-v1.5-7B	0.51
LLaVA-v1.5-13B	0.52
mPLUG-Owl-2	0.48
ShareGPT4V-7B	0.57
ShareGPT4V-13B	0.60
GPT-4V	0.77
Human	0.94

Table 2: Recognition score of LVLMs on Description task. For reference, the Recognition Score of Human is calculated based on the annotated [Description] in CogBench as an estimate of an upper limit score.

4.2.2 Cognition

Table 3 shows Cognition Scores of LVLMs on Description task. Similarly, GPT-4V achieves the best performance and there is a large performance gap between GPT-4V and other open-source models. For open-source models, Qwen-VL-Chat achieves the best performance. Though ShareGPT4V achieves better recognition performance than other open-source LVLMs, it's cognition performance does not show a significant improvement compared with other models. In terms of different reasoning capabilities, all of the LVLMs show better performance on [Location Reasoning] than others, which is probably because it is a kind of relatively lower-level reasoning. Differently, for [Event Reasoning], [Event Relationship Reasoning], and [Mental State Reasoning], all of the open-source LVLMs show very low performance, indicating they almost do not understand the story in the images at all. The significantly large performance gap on these three kinds of reasoning types between GPT-4V and open-source LVLMs could be a manifestation of the capabilities emerging in GPT-4V. Though GPT-4V achieves the best performance, there is also a huge gap between the cognition ability of LVLMs and human beings, and the gap is obviously larger than in the aspect of recognition. This indicates that LVLMs still have a lot of room for development in terms of cognitive abilities.

4.2.3 Case Study of Description Task

Figure 5 shows a case of Description task. In terms of recognition, GPT-4V shows a good performance by recognizing most annotated entities ("boy", "girl", "man", "woman", "table", "dress", "hat",

Model	Time	Location	Character	Character Relationship	Event	Event Relationship	Mental State	Next Moment Event	Overall
InstructBLIP-7B	0.13	0.54	0.19	0.35	0.10	0.04	0.05	0.17	0.17
Qwen-VL-Chat	0.26	0.54	0.31	0.28	0.18	0.13	0.09	0.20	0.22
LLaVA-v1.5-7b	0.09	0.52	0.19	0.27	0.09	0.05	0.05	0.19	0.16
LLaVA-v1.5-13b	0.04	0.47	0.21	0.33	0.09	0.06	0.04	0.20	0.17
mPLUG-Owl-2	0.04	0.47	0.19	0.28	0.10	0.05	0.05	0.14	0.15
ShareGPT4V-7B	0.22	0.56	0.21	0.25	0.09	0.03	0.05	0.19	0.16
ShareGPT4V-13B	0.22	0.54	0.24	0.29	0.09	0.06	0.09	0.17	0.17
GPT-4V	0.44	0.71	0.45	0.29	0.36	0.37	0.18	0.54	0.41
Human	0.87	0.95	0.95	0.84	0.99	0.94	0.86	0.91	0.93

Table 3: Cognition Scores of LVLMs on Description task evaluated by GPT-4. For reference, the Cognition Score of Human is calculated based on the annotated [Description] in CogBench as an estimate of an upper limit score.

"suit") and only fails to recognize the "sofa", "television", "handbag". LLaVA-v1.5-13B obviously recognizes fewer entities than GPT-4V, it only recognizes "boy", "woman", and "dress". However, GPT-4V fails to understand the story in the image and gets a 0 in terms of cognition. This is because it neither clarifies the location nor the relationships between the characters, nor does it accurately depict the events occurring in the image, etc. LLaVA-v1.5-13B is slightly better and successfully inferred that the woman is the mother.

The case shows CogBench can reflect that current LVLMs fail at aspects like recognition and cognition and the cognitive abilities exhibited in the description still have some gap with the level of humans.

4.3 Results of VQA Task

Table 4 shows the performance of LVLMs on VQA task. Consistent with results in previous sections, GPT-4V achieves the best performance and there is a performance gap between GPT-4V and open-source models. For open-source models, ShareGPT4V-13B and LLaVA-v1.5-13B achieves better performance than other LVLMs based on 7B LLMs, which shows the importance of LLM size to LVLMs. As for LVLMs based on 7B LLMs, the performance of InstructBLIP is the worst and there is also a gap between InstructBLIP and other models. One possible reason is that it only finetunes the Q-former for instruction-tuning as Q-former has limited capacity compared with LLMs. The performance of other models are similar. Consistent with previous findings, reasoning about location is also the easiest for LVLMs and event-related reasoning are more difficult. There is also a large gap between the performance of LVLMs and human beings.

5 Related Work

Evaluation Benchmark for LVLMs. To better understand emerging capabilities of LVLMs, many scholars are dedicated to proposing different evaluation benchmarks. LVLM-eHub (Xu et al., 2023) evaluates LVLMs' multimodal capabilities across six categories using various publicly available computer vision datasets. MME (Fu et al., 2023), MMBench (Liu et al., 2023c) and SEED Bench (Li et al., 2023a) use True/False Questions or Multiple-Choice Questions to evaluate different abilities of LVLMs. MM-VET (Yu et al., 2023) evaluates LVLMs in terms of their integrated VL capabilities via open-ended questions. Different from them, CogBench mainly focus on high-level cognition evaluation of LVLMs. Though some of them also consider cogition as one of the evaluation dimensions, most images they use can only be used to evaluate limited aspects of reasoning.

Image Captioning. Image Captioning is a classical VL task. As model capabilities advance, researchers strive to enhance their ability to describe images in detail. Krause et al. (2017) propose Image Paragraph Captioning, tasking models with generating a descriptive paragraph for an image. Recently, some researchers (Xie et al., 2022; Zhu et al., 2023a; Zhuge et al., 2023; Chen et al., 2023a) are trying to leveraging the ability of LLMs to generate more detailed image descriptions. None of these tasks consider to evaluate the high-level cognitive ability of models through description. HL dataset (Cafagna et al., 2023) requires models to generate high-level captions, but it only considers three aspects (scene, action, rationale). The content of most images in existing datasets does not reach the level of a story. This reveals the need for higher-level datasets like CogBench.

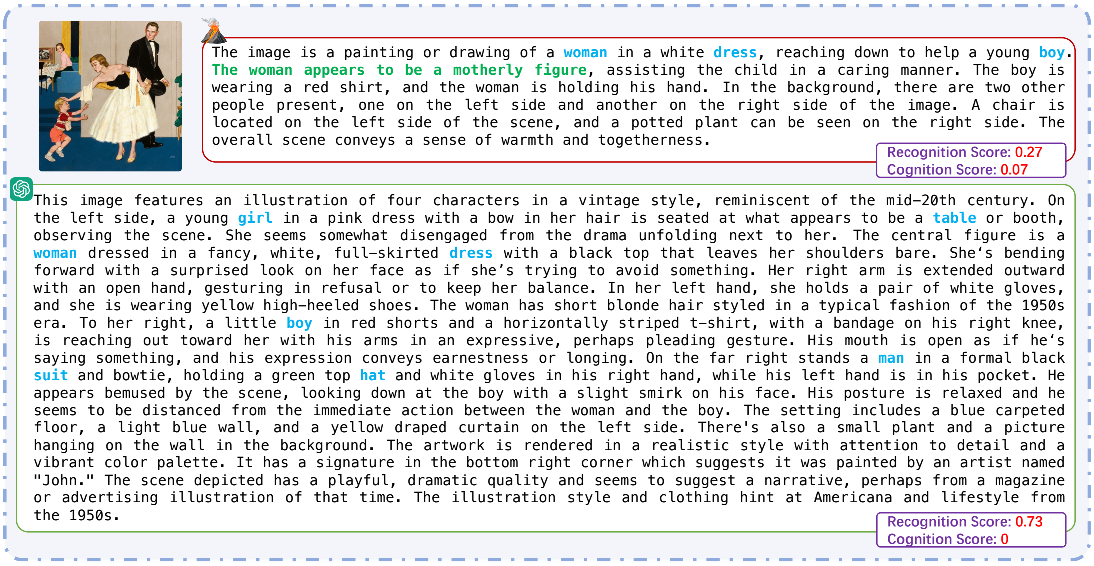

Figure 5: Case study of the Description task. A representative open-source LVLM LLaVA-v1.5-13B (red frame) and GPT-4V (green frame) are selected for analysis. Recognized entities are marked in blue, and CoRs are marked in green.

Model	Time	Location	Character	Character Relationship	Event	Event Relationship	Mental State	Next Moment Event	Overall
InstructBLIP-7B	0.34	0.57	0.47	0.52	0.36	0.44	0.37	0.39	0.43
Qwen-VL-Chat	0.64	0.84	0.63	0.55	0.54	0.40	0.53	0.50	0.55
LLaVA-V1.5-7B	0.55	0.80	0.54	0.54	0.51	0.37	0.53	0.47	0.52
LLaVA-V1.5-13B	0.68	0.85	0.60	0.62	0.48	0.54	0.56	0.68	0.59
mPLUG-Owl-2	0.57	0.82	0.54	0.53	0.51	0.46	0.56	0.53	0.55
ShareGPT4V-7B	0.62	0.78	0.63	0.59	0.50	0.36	0.56	0.53	0.54
ShareGPT4V-13B	0.70	0.85	0.59	0.56	0.50	0.50	0.58	0.72	0.58
GPT-4V	0.72	0.88	0.79	0.69	0.68	0.65	0.71	0.72	0.71
Human	0.98	0.96	0.99	0.96	0.97	0.98	0.95	0.96	0.96

Table 4: Model performance on VQA task. The accuracy of Human is calculated based on the responses of a healthy adult male.

Visual Reasoning. Visual Reasoning is closely related to the cognitive ability of models. VCR (Zellers et al., 2019) tasks models with answering visual questions using commonsense reasoning and justifying their answers. VisualCOMET Park et al. (2020) is a framework of visual commonsense reasoning tasks to predict past, future events, and present intents. Hessel et al. (2022) utilize images from VCR and Visual Genome (Krishna et al., 2017) to evaluate the ability of models to perform abductive reasoning. Fu et al. (2022) propose a task to identify the time and location of a given image. CURE (Chen et al., 2023b) is proposed to measure both the zero-shot reasoning performance and consistency of VLMs. Similarly, compared CogBench, these tasks considers less kinds of reasoning and CogBench can be seen as the next step of these efforts.

6 Conclusion

In this paper, we incorporated the idea of Cookie Theft picture description task into the evaluation of the high-level cognitive abilities of LVLMs and a novel evaluation benchmark CogBench is designed. Images in CogBench are of high quality and contain rich cognitive reasoning, which makes it different from existing image datasets. Experiments show that there is still a large gap between the cognitive ability of LVLMs and human beings, indicating CogBench is a challenging evaluation benchmark.

Limitations

Compared to other benchmarks, CogBench is a relatively light-weight benchmark due to our strict image collection criteria. We aim to substitute quantity with high-quality images and annotations.

We will keep undating CogBench in the future to include more high-quality images and annotations.

Ethical Considerations

Most images in CogBench are manually collected from Pinterest. We followed the term of service of Pinterest to collect these images. The images are used as fair use for research purposes only. We will share the data with other researchers who will follow all the ethical considerations as established in this study.

References

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Michele Cafagna, Kees van Deemter, and Albert Gatt. 2023. Hl dataset: visually-grounded description of scenes, actions and rationales. In *Proceedings of the 16th International Natural Language Generation Conference*, pages 293–312.

Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2023a. Sharegpt4v: Improving large multimodal models with better captions. *arXiv preprint arXiv:2311.12793*.

Yangyi Chen, Karan Sikka, Michael Cogswell, Heng Ji, and Ajay Divakaran. 2023b. Measuring and improving chain-of-thought reasoning in vision-language models. *arXiv preprint arXiv:2309.04461*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*.

Louise Cummings. 2019. Describing the cookie theft picture: Sources of breakdown in alzheimer's dementia. *Pragmatics and Society*, 10:151–174.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. 2023. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*.

Xingyu Fu, Ben Zhou, Ishaan Preetam Chandratreya, Carl Vondrick, and Dan Roth. 2022. There's a Time and Place for Reasoning Beyond the Image. In *Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Harold Goodglass, Edith Kaplan, and Sandra Weintraub. 2001. *BDAE: The Boston Diagnostic Aphasia Examination*. Lippincott Williams & Wilkins, Philadelphia, PA.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Jack Hessel, Jena D Hwang, Jae Sung Park, Rowan Zellers, Chandra Bhagavatula, Anna Rohrbach, Kate Saenko, and Yejin Choi. 2022. The abduction of sherlock holmes: A dataset for visual abductive reasoning. In *European Conference on Computer Vision*, pages 558–575. Springer.

Jonathan Krause, Justin Johnson, Ranjay Krishna, and Li Fei-Fei. 2017. A hierarchical approach for generating descriptive image paragraphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 317–325.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023a. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. In *NeurIPS*.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. 2023c. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*.

OpenAI. 2023. Gpt-4 technical report.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Jae Sung Park, Chandra Bhagavatula, Roozbeh Mottaghi, Ali Farhadi, and Yejin Choi. 2020. Visualcomet: Reasoning about the dynamic context of a still image. In *In Proceedings of the European Conference on Computer Vision (ECCV)*.

Thibault Sellam, Dipanjan Das, and Ankur P Parikh. 2020. Bleurt: Learning robust metrics for text generation. In *Proceedings of ACL*.

Mashrura Tasnim, Malikeh Ehghaghi, Brian Diep, and Jekaterina Novikova. 2022. DEPAC: a corpus for depression and anxiety detection from speech. In *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*, pages 1–16, Seattle, USA. Association for Computational Linguistics.

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

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Yujia Xie, Luowei Zhou, Xiyang Dai, Lu Yuan, Nguyen Bach, Ce Liu, and Michael Zeng. 2022. Visual clues: Bridging vision and language foundations for image paragraph captioning. *Advances in Neural Information Processing Systems*, 35:17287–17300.

Peng Xu, Wenqi Shao, Kaipeng Zhang, Peng Gao, Shuo Liu, Meng Lei, Fanqing Meng, Siyuan Huang, Yu Qiao, and Ping Luo. 2023. Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models. *arXiv preprint arXiv:2306.09265*.

Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. 2023. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 9(1):1.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2023. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. *arXiv preprint arXiv:2311.04257*.

Wenpeng Yin, Dragomir Radev, and Caiming Xiong. 2021. DocNLI: A large-scale dataset for document-level natural language inference. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4913–4922, Online. Association for Computational Linguistics.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2023. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Deyao Zhu, Jun Chen, Kilichbek Haydarov, Xiaoqian Shen, Wenxuan Zhang, and Mohamed Elhoseiny. 2023a. Chatgpt asks, blip-2 answers: Automatic questioning towards enriched visual descriptions. *arXiv preprint arXiv:2303.06594*.

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

Yan Zhuang, Qi Liu, Yuting Ning, Weizhe Huang, Rui Lv, Zhenya Huang, Guanhao Zhao, Zheng Zhang, Qingyang Mao, Shijin Wang, et al. 2023. Efficiently measuring the cognitive ability of llms: An adaptive testing perspective. *arXiv preprint arXiv:2306.10512*.

Mingchen Zhuge, Haozhe Liu, Francesco Faccio, Dylan R Ashley, Róbert Csordás, Anand Gopalakrishnan, Abdullah Hamdi, Hasan Hammoud, Vincent Herrmann, Kazuki Irie, Louis Kirsch, Bing Li, Guohao Li, Shuming Liu, Jinjie Mai, Piotr Piękos, Aditya Ramesh, Imanol Schlag, Weimin Shi, Aleksandar Stanić, Wenyi Wang, Yuhui Wang, Mengmeng Xu, Deng-Ping Fan, Bernard Ghanem, and Jürgen Schmidhuber. 2023. Mindstorms in natural language-based societies of mind. *arXiv preprint arXiv:2305.17066*.

A Image Annotation Instruction

Figure 6 shows the image annotation instruction for annotators of CogBench.

B Prompt of CoR-based GPT-assisted Question Generation

Figure 7 shows an example prompt of CoR-based GPT-assisted Question Generation for GPT-4. This prompt is used to generate questions based on [Event Reasoning] CoRs. Prompts for other CoR types are similar to this one.

C Prompt of GPT-based Description Cognition Evaluation

Figure 8 and Figure 9 shows prompts of ChatGPT or GPT-4 for cognition evaluation of Description task.

For CoR types other than [Event Relationship Reasoning], we task GPT-4 with determining whether the conclusion in each CoR is mentioned in the description. The prompt is shown in Figure 8. For [Event Relationship Reasoning], we task GPT-4 with determining whether the causal relationship between events (i.e. the whole CoR), as annotated, is present in the description. The prompt is shown in Figure 9.

D Introduction to Selected LVLMs

- **InstructBLIP** (Dai et al., 2023) builds upon BLIP-2 (Li et al., 2023b). It consists of an image encoder, a LLM, and a Query Transformer (Q-Former). During instruction tuning, only the Q-Former is updated. We use "blip2-vicuna-instruct" for testing.

- **Qwen-VL-Chat** (Bai et al., 2023) is a instruction-tuned VL chatbot based on Qwen-VL. Its training process consists of two stages of pre-training and a final stage of instruction fine-tuning. As for architecture, it consists of a vision encoder, a LLM, and position-aware vision-language adapter. We test "Qwen-VL-Chat".

- **LLaVA v1.5** (Liu et al., 2023a) is an upgraded version of LLaVA (Liu et al., 2023b). LLaVA connects a vision encoder and LLM for general-purpose visual and language understanding. It is instruction-tuned on the language-image instruction-following data generated by language-only GPT-4 (OpenAI, 2023). By using CLIP-ViT-L-336px with an MLP projection and adding academic-task-oriented VQA data with simple response formatting prompts, LLaVA v1.5 achieves better performance. "llava-v1.5-7b" and "llava-v1.5-13b" are tested.

Model	Visual Encoder	Language Model
InstructBLIP-7B	EVA-CLIP ViT-g	Vicuna-7B
Qwen-VL-Chat	OpenCLIP ViT-G	Qwen-7B
LLaVA-v1.5-7B	CLIP ViT-L	Vicuna-7B
LLaVA-v1.5-13B	CLIP ViT-L	Vicuna-13B
mPLUG-Owl-2	CLIP ViT-L	LLaMA-7B
ShareGPT4V-7B	CLIP ViT-L	Vicuna-7B
ShareGPT4V-13B	CLIP ViT-L	Vicuna-13B
GPT-4V	–	–

Table 5: LVLMs evaluated in this paper.

- **mPLUG-Owl-2** (Ye et al., 2023) mainly comprises a fundamental vision encoder, a visual abstractor, and a language decoder. It also adopts a two-stage training strategy, comprising pre-training and visual instruction tuning. We test "mplug-owl2-llama2-7b".

- **ShareGPT4V** (Chen et al., 2023a) follows the design of LLaVA v1.5. They incorporate a large-scale resource featuring highly descriptive captions into both the pre-training and supervised fine-tuning phases of ShareGPT4V model. We test"ShareGPT4V-7B" and "ShareGPT4V-13B".

- **GPT-4V** (OpenAI, 2023) is one of the most powerful LVLMs in the world developed by OpenAI. The version of "gpt-4-vision-preview" is tested.

Table 5 shows an overview of the design of different LVLMs.

E Evaluation Results of Description Task based on Traditional Image Captioning Metrics

Table 6 shows the performance of models on traditional image captioning evaluation metrics. Following Krause et al. (2017), we use METEOR (Banerjee and Lavie, 2005), CIDEr (Vedantam et al., 2015), BLEU-1, BLEU-2, BLEU-3, and BLEU-4 (Papineni et al., 2002) to evaluate the performance of models on Description task. Similar to the findings of Zhu et al. (2023a), it can be observed that traditional image captioning evaluation metrics are not quite suitable for evaluating the Description

You are going to see some pictures. Each picture tells a story and requires different kinds of reasoning to fully understand the story behind it. You will be first asked to identify the entities and reasoning processes in the picture. Then, you will need to describe the story of the picture based on your identified entities and reasoning processes.

First, you will be asked to identify the entities in the picture. The annotation format is [A, B, C], where A, B, C are entities.

[Entities]: Please list all entities appearing in the picture, including people, animals, objects etc. You are encouraged to list as many entities as possible. Note that these entities need to be in your picture description afterwards. For entities that are difficult to recognize, please do not list them here or describe them.

Then, you will be asked to identify different reasoning processes in the picture. The annotation format should follows the structure [A1 + A2 -> B], where A1 and A2 are premises and B is the conclusion. Note that if you write a conclusion, there must be at least one premise. Do not write a conclusion only, like [B]. Please write one conclusion at a time, do not write a reasoning process like [A1->B->C], which should be split into two. Each picture does not necessarily requires all the reasoning. Please write None, if a picture does not involve a specific kind of reasoning or it is not important in the picture.

[Special Time Reasoning] Please write your reasoning processes about the special time of the story in the picture, e.g. festivals, seasons etc. The special time is usually relevant to the story of the picture. For instance, if it is daytime in a picture, it is easily recognized and requires no reasoning and there is nothing special, you can write None. However, if there is a lamp on or a clock indicating a specific time, you can write down your reasoning about time.

[Location Reasoning] Please write your reasoning processes about the location of the story in the picture, e.g. near the school.

[Character Reasoning]: Please write your reasoning processes about the character of the subjects in the picture, e.g. teacher, doctor etc.

[Character Relationship Reasoning]: Please write your reasoning processes about relationship between characters in the picture, e.g. friendship.

[Event Reasoning]: Please write your reasoning processes about events happened in the current moment and previous moments in the picture based on clues in the picture. Note that you only need to annotate those high-level events and you can ignore those low-level events. For instance, "the woman is looking at the man" is a low level event and you can ignore its reasoning process. Differently, the reasoning process [A mother is busy cooking. + A boy is fetching cookies behind the mom. + A girl is shushing the boy. -> The boy is stealing cookies.] is a reasoning about high-level event stealing and you should write it down.

[Event Relationship Reasoning]: Please write your reasoning processes about relationship between different events in the picture. These events are usually linked through causal and temporal relations. Note that events in this part do not necessarily appears in the [Event Reasoning] part as some events here are low-level events.

[Next Moment Event Reasoning]: Please write your reasoning processes about the event that will happen in the next moment. Note that you only need to write down events that have a very high probability of happening, instead of guessing what might happen next.

[Mental State Reasoning]: Please write your reasoning processes about mental state of subjects in the picture, e.g. daydreaming, happy etc. You need to reason as best you can about the mental states of all the subjects in the picture, unless they are not showing obvious emotions.

Finally, you will be asked to describe the picture in as much detail as you can.

[Description]: Please describe all you see in the picture in a paragraph based on the entities and reasoning processes you identified above, ensuring that all of them are included in your description. Each picture has a story behind it and you need to tell that story through your description.

Figure 6: Image annotation instruction for annotators.

We have a description of an image and the description tells a detailed story unfolding in the image. In the process of describing an image, it is often necessary to engage in reasoning about events based on the clues within the image, leading to certain conclusions. For example, when we see the wind is blowing outside, and a man is reading a newspaper in the telephone booth, we can infer that he is actually hiding from the wind in the telephone booth. Therefore, in this task, in addition to the image description, the reasoning processes about event within the image description have also been extracted. For each reasoning process, we use A1+A2+...->B to represent it, where A1, A2, ... are clues we observed in the picture and B represents the conclusion about event we inferred.

Thus, given an image description and the reasoning processes about event, our task is:
1) Generate a question based on reasoning processes about event.
2) Generate four options: A, B, C, and D. There is only one correct answer among the four options, which is consistent with the description and reasoning processes provided. The correct answer option should be randomly chosen from A, B, C, and D. For those incorrect options (distractors), you are encouraged to hallucinate some clues that are highly relevant to the question and the description but do not actually consistent with the description. That is, you can distort the facts in the description and reasoning processes using elements related to the question to generate some easily selectable distractors. It would be better if you can generate some distractors that are similar to but different from the correct option. Please avoid situations where the correct option is significantly longer or shorter than the distractors.
—
For example, if the description is "There are some snow on the ground and it is windy, ... We can see it is cold. Inside a phone booth, a man is smiling while looking at a newspaper. He is sheltering from the cold wind in the phone booth..." and the question is "Why can we tell that the man is seeking shelter for warmth?", you can use "newsstand", which is related to "seeking shelter for warmth" in the question, to distort the fact in description "in a phone booth." Then you can get "the man is in the newsstand." Similarly, you can hallucinate a question related distractor "it is raining and a man is smiling and reading a newspaper in a phone booth," which is similar to the correct option "it is windy and a man is smiling and reading a newspaper in a phone booth," but different from it and inconsistent with the description.
—
3) Generate the the letter corresponding to the correct answer, that is A, B, C, or D.

Here are some examples:
———-

[Description]:
There are some snow on the ground and it is windy, indicating it is winter. There are two men and two women standing on the roadside. There is a sign that says "NO STANDING BUS STOP", indicating it is near a bus stop. A man is standing on the road side, wrapping his coat tightly around himself, and peering out onto the road. They are probably waiting for a bus here. We can see it is cold. Inside a phone booth, a man is smiling while looking at a newspaper. He is sheltering from the cold wind in the phone booth. He looks happy, because it is warm there. Two women are also wrapping their coats tightly and looking at the man in the phone booth. They are probably friends and standing together. They are unhappy with the man. There are some buildings by the road.

[Event Reasoning]:
It is windy and cold. + A man is standing in a phone booth reading newspaper. -> The man is sheltering from the cold wind in the phone booth.

[Generated Multiple-Choice Questions]:
What is the man doing in the phone booth?
A. Making a phone call.
B. Reading a book.
C. Avoiding someone he doesn't want to see.
D. Sheltering from the cold wind.
Correct Answer: [D]

Why can we tell that the man is seeking shelter for warmth?
A. It is windy and a man is smiling and reading a newspaper in a newsstand.
B. It is raining and a man is smiling and reading a newspaper in a newsstand.
C. It is windy and a man is smiling and reading a newspaper in a phone booth.
D. It is raining and a man is smiling and reading a newspaper in a phone booth.
Correct Answer: [C]
———-

Please:
1). Generate at least one question for each reasoning process.
2). Generate more diverse questions, try to generat questions from different perspectives or angles and don't limit yourself to the question templates provided in the examples.
3). Avoid generating repetitive questions with similar meanings.

Figure 7: An example prompt of CoR-based GPT-assisted Question Generation for GPT-4 to generate questions based on [Event Reasoning].

Given a <DESCRIPTION> and some <KEY POINT>s, please tell me if the <DESCRIPTION> explicitly presents the exact or similar semantics of each <KEY POINT>. The following points are required:

1) Instead of reasoning about whether each <KEY POINT> is possibly correct based on the <DESCRIPTION>, you only need to determine whether the <DESCRIPTION> mentions the semantics in the <KEY POINT>.
2) Do not overlook the semantics in the <DESCRIPTION> that are semantically equivalent to the <KEY POINT> but expressed in different ways. For instance, if the <DESCRIPTION> mentions "The woman is playing with her son...", we can tell it successfully includes semantics in the <KEY POINT> "The woman is the mother of the boy."
3) If several possible scenarios are listed using 'or' at a <KEY POINT>, you only need to determine whether one of these scenarios is mentioned in the <DESCRIPTION>.

Assign a score of 0 or 1 to each <KEY POINT>, where 0 represents NO and 1 represents YES.

<DESCRIPTION>:
{Description generated by models.}

<KEY POINT>:
1. {Annotated key point 1.}
2. {Annotated key point 2.}
...
N. {Annotated key point N.}

Please write your answers in "[ ]" with 0 or 1 in the following format (number + square brackets):

1. [1] 2. [0]

Your answers to the {N} <KEY POINT>(s) above:
1. [ ] 2. [ ] ... N. [ ]

Figure 8: Evaluation prompt for GPT-4 of Reseasoning types other than [Event Relationship Reasoning].

Given a <DESCRIPTION> and some <EVENT RELATIONSHIP>s, please tell me whether this <DESCRIPTION> clearly depicts the cause-and-effect relationships between events.

The format of a <EVENT RELATIONSHIP> follows the structure "A1 + A2 + ... + An -> B", where A1, A2, ..., An and B are events. Events A1, A2, ..., An are the causes of event B, and event B is the result caused by events A1, A2, ..., An. The criteria for judgment lie in whether the <DESCRIPTION> mentions these events and clearly depicts the causal relationships between them.

Assign a score of 0 or 1 to each <EVENT RELATIONSHIP>, where 0 represents NO and 1 represents YES.

<DESCRIPTION>:
{Description generated by models.}

<EVENT RELATIONSHIP>:
1. {Annotated event relationship 1.}
2. {Annotated event relationship 2.}
...
N. {Annotated event relationship 3.}

Please write your answers in "[ ]" with 0 or 1 in the following format (number + square brackets):

1. [1] 2. [0]

Your answers to the {N} <EVENT RELATIONSHIP>(s) above:
1. [ ] 2. [ ] ... N. [ ]

Figure 9: Evaluation prompt for GPT-4 of [Event Relationship Reseasoning].

task. There are two possible reasons. The first possible reason is that image descriptions are longer and more flexible than traditional image captions. The second possible reason is that Description task requires evalution metrics to consider high-level semantics in description, while traditional image captioning evaluation metrics only concerns low-level information.

F Effectiveness Analysis of GPT-based Cognition Evaluation of Description Task

To prove the effectiveness of GPT-based evaluation, we manually annotated a subset by assigning 0/1 to CoRs of 20 images and use the subset to evaluate the performance of different evluation methods.

F.1 Implementation of Non-GPT-based Evaluation Methods

Apart from ChatGPT or GPT-4, some other evaluation methods are implemented to perform this task, as shown in Table 7.

For methods based on ROUGE (Lin, 2004), BERTScore (Zhang et al., 2020), and BLEURT (Sellam et al., 2020), we first split the description into sentences, then use a CoR as a reference to calculate the (recall) score for each sentence compared to the CoR. Then, the highest score among all calculated scores is taken as the score of the CoR corresponding to the description. Finally, the score is converted into 0/1 using a threshold.

We also tried Natural Language Inference (NLI) models to perform the task. First, we use DeBERTa (He et al., 2021) to perform sentence-level NLI task similar to methods mentioned above. If there is at least one "Entailment" for all the sentences, the score of the CoR will be 1. The model we adopted is *mDeBERTa-v3-base-xnli-multilingual-nli-2mil7*. The second NLI model we tried is DocNLI (Yin et al., 2021), which can directly take the description and CoR as input and do the classification task.

F.2 Result Analysis

Table 7 shows the accuracy of different evaluation methods on the subset. It can be seen that GPT-4 achieves the best performance, which indicates that GPT-based evaluation is generally consistent with human evaluation and thus effective for evaluating the performance of LVLMs on Description task.

Model	METEOR	CIDEr	BLEU-1	BLEU-2	BLEU-3	BLEU-4
InstructBLIP-7B	0.127	0.032	0.244	0.123	0.061	0.032
Qwen-VL-Chat	0.130	0.042	0.236	0.124	0.059	0.030
LLaVA-V1.5-7B	0.139	0.038	0.283	0.145	0.071	0.036
LLaVA-V1.5-13B	0.143	0.037	0.299	0.156	0.074	0.037
mPLUG-Owl-2	0.129	0.022	0.252	0.121	0.055	0.025
ShareGPT4V-7B	0.161	0.022	0.287	0.135	0.06	0.029
ShareGPT4V-13B	0.163	0.035	0.289	0.135	0.06	0.028
GPT-4V	0.207	0.008	0.240	0.118	0.052	0.025

Table 6: Model performance on Description task evaluated by traditional image captioning evaluation metrics.

Model	Accuracy
ROUGE	0.656
BERTScore	0.635
BLEURT	0.620
DeBERTa	0.693
DocNLI	0.714
GPT-3.5	0.807
GPT-4	0.849

Table 7: Accuray of different evaluation methods.