EgoThink: Evaluating First-Person Perspective Thinking Capability of Vision-Language Models

Anonymous ACL submission

Abstract

Vision-language models (VLMs) have recently shown promising results in traditional downstream tasks. The capability of VLMs to “think” from a first-person perspective, a crucial attribute for advancing autonomous agents and robotics, remains largely unexplored. To bridge this research gap, we introduce EgoThink, a novel visual question-answering benchmark that encompasses six core capabilities with twelve detailed dimensions. The benchmark is constructed using selected clips from egocentric videos, with manually annotated question-answer pairs containing first-person information. To comprehensively assess VLMs, we evaluate twenty-one popular VLMs on EgoThink. Moreover, given the open-ended format of the answers, we use GPT-4 as the automatic judge to compute single-answer grading. Experimental results indicate that although GPT-4V leads in numerous dimensions, all evaluated VLMs still possess considerable potential for improvement in first-person perspective tasks.

1 Introduction

Vision-language models (VLMs) (Yang et al., 2023b; Alayrac et al., 2022; Li et al., 2023b; Driess et al., 2023) have shown remarkable progress in both conventional vision-language downstream tasks (Yang et al., 2023b; Alayrac et al., 2022; Li et al., 2023b; Driess et al., 2023) and following diverse human instructions (Dai et al., 2023; Li et al., 2023a; Ye et al., 2023; Zhu et al., 2023; Liu et al., 2023). Their application has expanded into broader domains such as robotics (Gao et al., 2023; Huang et al., 2023; Kuo et al., 2022) and embodied artificial intelligence (EAI) (Yang et al., 2023a; Sumers et al., 2023). As a result, the thorough evaluation of VLMs has become increasingly important and challenging. Observing and understanding the world from a first-person perspective is a natural approach for both humans and artificial intelligence agents. We propose that the ability to “think” from a first-person perspective, especially when interpreting egocentric images, is crucial for VLMs. Therefore, there is a clear need to develop a comprehensive benchmark to evaluate the first-person capabilities of VLMs more effectively. In this work, we introduce a new benchmark for VLMs from a first-person perspective, named EgoThink.

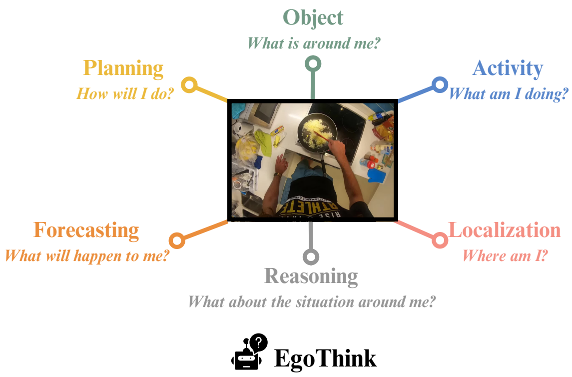

Figure 1: The main categories of our EgoThink benchmark to comprehensively assess the capability of thinking from a first-person perspective.

2 EgoThink Benchmark

2.1 Core Capabilities

We design six categories with twelve fine-grained dimensions from the first-person perspective for quantitative evaluation. (1) **Object: What is around me?** Recognizing objects in the real world is essential for human vision. We divide this into three dimensions: *Existence* (predicting object presence), *Attribute* (detecting object characteristics), and *Affordance* (predicting potential human actions on objects). (2) **Activity: What am I doing?** Activity recognition focuses on actions based on object-hand interactions from an egocentric perspective. (3) **Localization: Where am I?** Localization involves detecting the scene (Location) and understanding the spatial relationship of objects rel-

Methods	Object			Activity	Localization		Reasoning			Forecasting	Planning		Average
	Exist	Attr	Afford		Loc	Spatial	Count	Compar	Situated		Nav	Assist	
API-based model													
GPT-4V	62.0	**82.0**	**58.0**	**59.5**	86.0	62.0	**42.0**	48.0	**83.0**	**55.0**	**64.0**	**84.0**	**65.5**
~7B Models													
BLIP-2-6.7B	49.0	29.0	39.0	33.5	60.0	31.0	3.0	21.0	33.0	25.0	8.0	6.0	28.1
LLaVA-1.5-7B	33.0	47.0	54.0	35.5	35.0	49.0	20.0	47.0	37.0	27.0	29.0	54.0	39.0
MiniGPT-4-7B	50.0	56.0	46.0	39.0	55.0	49.0	14.0	48.0	31.0	41.5	14.0	44.0	40.6
InstructBLIP-7B	50.0	33.0	45.0	47.5	77.0	38.0	18.0	43.0	67.0	40.5	19.0	31.0	42.4
Otter-I-7B	48.0	56.0	39.0	44.0	60.0	44.0	39.0	48.0	42.0	38.0	31.0	55.0	45.3
PandaGPT-7B	40.0	56.0	41.0	37.0	61.0	52.0	19.0	52.0	53.0	43.0	39.0	61.0	46.2
mPLUG-owl-7B	56.0	58.0	47.0	53.0	60.0	53.0	25.0	49.0	44.0	49.5	33.0	58.0	48.8
Video-LLaVA-7B	56.0	60.0	53.0	45.0	86.0	60.0	39.0	38.0	60.0	46.5	11.0	38.0	49.4
LLaVA-7B	63.0	58.0	50.0	47.0	81.0	45.0	24.0	36.0	47.0	49.5	35.0	60.0	49.6
ShareGPT4V-7B	67.0	75.0	53.0	55.5	77.0	62.0	30.0	38.0	66.0	47.0	41.0	63.0	51.9
~13B Models													
InstructBLIP-13B	52.0	55.0	49.0	54.0	63.0	49.0	11.0	33.0	59.0	44.0	19.0	25.0	42.8
PandaGPT-13B	35.0	52.0	41.0	40.5	68.0	31.0	32.0	40.0	47.0	45.5	16.0	69.0	43.1
LLaVA-13B-Vicuna	54.0	62.0	52.0	46.0	53.0	46.0	26.0	44.0	29.0	44.0	35.0	66.0	46.4
BLIP-2-11B	52.0	62.0	41.0	49.5	**90.0**	**66.0**	25.0	50.0	70.0	48.0	18.0	24.0	49.6
InstructBLIP-11B	**74.0**	68.0	48.0	49.5	86.0	52.0	32.0	49.0	73.0	53.0	16.0	17.0	51.5
LLaVA-13B-Llama2	65.0	61.0	45.0	56.0	77.0	53.0	34.0	34.0	66.0	50.5	49.0	71.0	55.1
LLaVA-1.5-13B	66.0	55.0	51.0	55.0	82.0	57.0	32.0	**56.0**	67.0	48.5	39.0	55.0	55.3

Table 1: Combined single-answer grading scores on zero-shot setups for various dimensions. The **bold** indicates the best performance while the underline indicates the second-best performance. Exist, Attr, Afford, Loc, Spatial, Count, Compar, Situated, Nav and Assist represent existence, attribute, affordance, location, spatial relationship, counting, comparison, situated reasoning, navigation, and assistance.

ative to the subject. (4) **Reasoning: What about the situation around me?** This includes *Counting*, *Comparison*, and *Situated Reasoning*, focusing on objects in hand or surroundings and requiring further reasoning. (5) **Forecasting: What will happen to me?** Forecasting predicts future object-state transformations or hand-object interactions. (6) **Planning: How will I do?** Planning involves *Navigation* (going from start to goal) and *Assistance* (offering instructions for daily problems).

2.2 Data Collection

To construct the EgoThink benchmark, we leverage the Ego4D dataset, extracting first-person visual data from its vast collection of videos. We engage annotators to manually label question-answer pairs, ensuring diversity and quality by selecting images that meet strict criteria and avoiding repetition. The EgoThink benchmark currently comprises 700 images across six categories with twelve dimensions, sourced from 595 videos to guarantee a wide range of scenarios. We craft questions and answers for each image to mimic real-life conversations, using a variety of question types and ensuring accuracy in responses. The dataset's size represents a balanced approach to benchmark diversity and the cost of open-ended QA evaluation, ensuring robust performance estimation within practical limits.

3 Experiments

Setups. We evaluate eighteen prominent Vision-Language Models (VLMs), divided into two parameter size groups for fair comparison. We perform zero-shot setups for all VLMs. To objectively grade single-answer outputs, we use GPT-4 as an automatic evaluator, prioritizing semantic accuracy over surface similarity. The GPT-4 evaluator is asked to assign a score of 0 (wrong), 0.5 (partially correct), or 1 (correct) to the model output.

Results. We present the overall results of the evaluated models on our EgoThink benchmark as shown in Table 1. Despite having improved over the years, VLMs are still difficult to think from a first-person perspective, even GPT-4V. Among the six categories, only the scores on planning and localization are relatively high, the performance in other capabilities can only reach around 60 points at best. Among the better models, GPT-4V generally performs much better than other models.

4 Conclusion

To pave the way for the development of VLMs in the field of EAI and robotics, we introduce a comprehensive benchmark, EgoThink. In future research, we aim to further explore the essential capabilities of VLMs in the EAI and robotics fields.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736.

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. *Preprint*, arXiv:2305.06500.

Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. 2023. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*.

Jensen Gao, Bidipta Sarkar, Fei Xia, Ted Xiao, Jiajun Wu, Brian Ichter, Anirudha Majumdar, and Dorsa Sadigh. 2023. Physically grounded vision-language models for robotic manipulation. *arXiv preprint arXiv:2309.02561*.

Wenlong Huang, Chen Wang, Ruohan Zhang, Yunzhu Li, Jiajun Wu, and Li Fei-Fei. 2023. Voxposer: Composable 3d value maps for robotic manipulation with language models. *arXiv preprint arXiv:2307.05973*.

Weicheng Kuo, Yin Cui, Xiuye Gu, AJ Piergiovanni, and Anelia Angelova. 2022. F-vlm: Open-vocabulary object detection upon frozen vision and language models. *arXiv preprint arXiv:2209.15639*.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. 2023a. Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726*.

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*.

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

Theodore Sumers, Kenneth Marino, Arun Ahuja, Rob Fergus, and Ishita Dasgupta. 2023. Distilling internet-scale vision-language models into embodied agents. *arXiv preprint arXiv:2301.12507*.

Jingkang Yang, Yuhao Dong, Shuai Liu, Bo Li, Ziyue Wang, Chencheng Jiang, Haoran Tan, Jiamu Kang, Yuanhan Zhang, Kaiyang Zhou, et al. 2023a. Octopus: Embodied vision-language programmer from environmental feedback. *arXiv preprint arXiv:2310.08588*.

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

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chaoya Jiang, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qian Qi, Ji Zhang, and Fei Huang. 2023. mplug-owl: Modularization empowers large language models with multimodality. *Preprint*, arXiv:2304.14178.

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