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

Sijie Cheng[1,2,5 ‡*], **Zhicheng Guo**[1,2*], **Jingwen Wu**[3*], **Kechen Fang**[4],
Peng Li[2 ✉], **Huaping Liu**[1], **Yang Liu**[1,2 ✉]
[1]Department of Computer Science and Technology, Tsinghua University
[2]Institute for AI Industry Research (AIR), Tsinghua University
[3]Department of Electrical and Computer Engineering, University of Toronto
[4]Zhili College, Tsinghua University [5]01.AI
csj23@mails.tsinghua.edu.cn

Abstract

Vision-language models (VLMs) have recently shown promising results in traditional downstream tasks. Evaluation studies have emerged to assess their abilities, with the majority focusing on the third-person perspective, and only a few addressing specific tasks from the first-person perspective. However, 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. Meanwhile, enlarging the number of trainable parameters has the most significant impact on model performance on EgoThink. In conclusion, EgoThink serves as a valuable addition to existing evaluation benchmarks for VLMs, providing an indispensable resource for future research in the realm of embodied artificial intelligence and robotics.

*Equal contribution, ‡ Project lead, ✉ Corresponding author
Project page: https://adacheng.github.io/EgoThink/
GitHub page: https://github.com/AdaCheng/EgoThink/
Dataset page: https://huggingface.co/datasets/EgoThink/EgoThink/

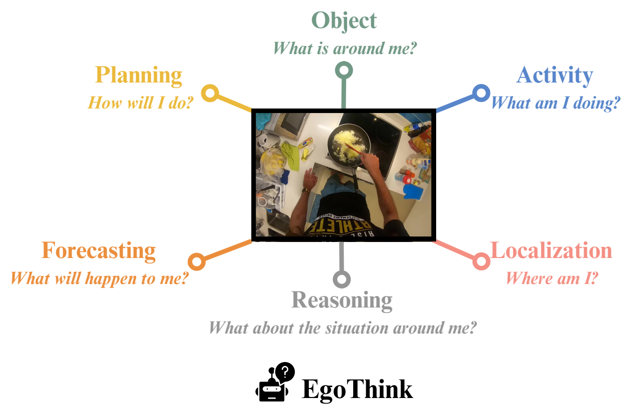


Figure 1. The main categories of our EgoThink benchmark to comprehensively assess the capability of thinking from a first-person perspective.

1. Introduction

Benefiting from the rapid development of large language models (LLMs) [8, 60, 73], vision-language models (VLMs) [2, 15, 43, 80] have shown remarkable progress in both conventional vision-language downstream tasks [2, 15, 43, 80] and following diverse human instructions [13, 42, 48, 81, 89]. Their application has expanded into broader domains such as robotics [21, 31, 40] and embodied artificial intelligence (EAI) [71, 78]. 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.

However, as shown in Table 1, the ability to think from a first-person perspective is not adequately addressed by cur-

Benchmark	Capability	Perspective	Data Source	Answer Type	Evaluator	Size
VL-CheckList [84]	Object / Attribute / Relation	Third	Datasets	PS	Accuracy	410k
LVLM-eHub [77]	General Multi-Modality	Third	Datasets	MC / OE	Metrics / LLMs / User	332k
MME [19]	General Multi-Modality	Third	Handcraft	MC	Accuracy	2,194
Tiny LVLM-eHub [68]	General Multi-Modality	Third	Datasets	OE	LLMs	2,100
MMBench [54]	General Multi-Modality	Third	Datasets / Handcraft / LLMs	MC	LLMs	2,974
PCA-EVAL [11]	Decision-Making	Third	Handcraft	OE	Accuracy / User	300
EgoTaskQA [34]	Spatial / Temporal / Causal	First	Crowdsourcing	OE	Crowdsourcing	40k
EgoVQA [16]	Object / Action / Person	Third / First	Handcraft	MC	Accuracy	520
EgoThink (Ours)	First-Person Thinking	First	Handcraft	OE	LLMs	700

Table 1. Comparison of recent comprehensive evaluation benchmarks of VLMs and our proposed benchmark EgoThink. Third and first indicate third-person and first-person perspectives. Datasets/Handcraft/LLMs denote existing datasets, manual annotation, and automatic generation by LLMs. PS/MC/OE indicate pairwise scoring, multi-choice, and open-ended question-answering, respectively.

rent evaluation benchmarks for VLMs. On one hand, most of these benchmarks (six out of nine, as listed in Table 1) focus solely on the third-person perspective. On the other hand, those benchmarks that do consider the first-person perspective only encompass a limited range of capabilities. For instance, EgoTaskQA [34] examines spatial, temporal, and causal aspects, whereas EgoVQA [16] is limited to object, action, and person aspects. 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. The initial step in developing this benchmark involves determining the necessary capabilities to assess. Humans, when interacting with the real world, consider a series of questions centered on themselves, ranging from "*What is around me?*", "*What am I doing?*", "*Where am I?*", "*What about the situation around me?*", "*What will happen to me?*" to "*How will I do?*". Drawing inspiration from this, we evaluate six core capabilities of VLMs, namely object, activity, localization, reasoning, forecasting, and planning. Each capability corresponds to one of the aforementioned questions, as illustrated in Figure 1. The next step is constructing the benchmark. We first categorize the six core capabilities into twelve detailed dimensions. We then select a minimum of 50 distinct and clear clips from egocentric videos for each dimension and manually annotate them with relevant first-person question-answer pairs. This approach ensures the quality and variety of the benchmark. The final step is evaluating VLM performance on this benchmark. Building on recent studies [7, 12, 68], we use GPT-4 [60] as an automatic evaluator. The Pearson correlation coefficient, when compared with human evaluation, shows a value of 0.68, indicating that the evaluation results are dependable.

Based on our proposed EgoThink benchmark, we conduct comprehensive experiments to evaluate the first-person capabilities of twenty-one popular VLMs with varying model and data compositions. The findings indicate that GPT-4V stands out as the most effective model in various aspects. However, it shows less impressive results in specific capabilities such as activity and counting. Additionally, we observed that no single VLM consistently surpasses others in every aspect. For instance, GPT-4V is less effective than BLIP-2-11B for localization. Increasing the language model portion of the VLMs generally leads to better performance, but this improvement is not uniform across all models. Finally, our results highlight a significant potential for further enhancing the first-person capabilities of VLMs.

2. Related Work

Vision-Language Models. Inspired by the impressive success of LLMs [8, 61, 75], the recent popular VLMs tend to regard the powerful LLMs as the core backbone. At the beginning, VLMs usually use large-scale image-text pairwise datasets [9, 35, 46] or arbitrarily interleaved visual and textual data [2, 90] to pre-train. Furthermore, thanks to the availability of enormous image-text instruction datasets [41, 49], recent studies [13, 42, 48, 81, 89] further apply instruction tuning to help VLMs generate satisfactory answers. Benefiting from the two-stage training process, recent VLMs can achieve stunning performance on downstream vision-language tasks [3, 33, 46, 64].

Evaluations of VLMs. To evaluate the abilities of VLMs, there are diverse types of vision language downstream tasks. Conventional benchmarks, such as image caption tasks [29, 82] and visual question reasoning tasks [23, 67], mainly probe specific abilities of VLMs from the third-person perspective. Meanwhile, specialized analytical studies comprehensively evaluate the performance of VLMs from the third-person perspective, where Vlue [87] consists of five fundamental tasks and Lvlm-ehub [77] evaluates six categories of capabilities on 47 standard vision-language benchmarks. As for the first-person perspective, there are some egocentric evaluation benchmarks in the computer vision field to assess some visual capabilities [32, 53, 83, 88]. In terms of multi-modality, there are a few benchmarks, such as EgoVQA[16] and EgoTaskQA [34], where mainly specific tasks without an overall understanding. In this paper, we mainly focus on exploring the comprehensive capa-

Figure 2. Categories with fine-grained dimensions and their corresponding examples of EgoThink benchmark.

bilities of VLMs to think from a first-person perspective, as a supplement to previous evaluation benchmarks.

3. EgoThink Benchmark

In this section, we first elaborate on the core capabilities of thinking from a first-person perspective. Then, we introduce the process to manually construct our proposed benchmark EgoThink, which asks VLMs to generate open-ended answers according to first-person images and questions.

3.1. Core Capabilities

As shown in Figure 2, we specifically design six categories with twelve fine-grained dimensions from the first-person perspective for quantitative evaluation.

- **Object: What is around me?** Recognizing objects in the real world is a preliminary ability of the human visual system [50, 85, 91]. Images from a first-person or egocentric perspective [53, 65, 88] pay more attention to the objects surrounding the subject or in hands. Moreover, we further divide the object category into three fine-grained dimensions: (1) *Existence*, predicting whether there is an object as described in the images; (2) *Attribute* [17, 37], detecting properties or characteristics (e.g., color) of an object; (3) *Affordance* [28, 56], predicting potential actions that a human can apply to an object.
- **Activity: What am I doing?** Activity recognition is to automatically recognize specific human activities in video frames or still images [36, 38, 74]. From the egocentric perspective, we mainly focus on actions or activities based on object-hand interaction [6, 18, 59].
- **Localization: Where am I?** In reality, localization is a critical capability for navigation and scene understanding in the real world [55, 66]. Here we investigate the localization capability from two aspects, *Location* and *Spatial Relationship*. Location indicates detecting the scene surrounding the subject [14, 26]. Spatial reasoning contains allocentric and egocentric perspectives [24, 39, 57, 58]. We focus on the egocentric perspective, i.e., the position of the object with respect to the subject.
- **Reasoning: What about the situation around me?** During the complex decision-making process, reasoning lies everywhere in our lives. Here we mainly focus on *Counting*, *Comparison*, and *Situated Reasoning*. Due to the first-person perspective, we generally count or compare objects in our hands or surrounding ourselves. As for situated reasoning, we employ cases that cannot be answered directly from the information in the images and require further reasoning processes.
- **Forecasting: What will happen to me?** Forecasting [20, 25, 51, 52] is a critical skill in the real world. From an egocentric view, forecasting always predicts the future of

VLMs	Image Encoder	LLM	Alignment Module	TTP	ToP	Dataset Size Image-Text	Dataset Size Instruction	EgoData	Video
API-based Model									
GPT-4V [60]	Unknown								
∼7B Models									
OpenFlamingo-7B [2, 5]	CLIP-ViT-L	MPT$_{7B}$	Attention	1.4B	8.1B	2B	-	✗	✓
BLIP-2-6.7B [43]	EVA-CLIP-ViT-g	OPT$_{6.7B}$	Q-Former	108M	7.8B	129M	-	✗	✗
VideoChat-7B [44]	BLIP2-VE	Vicuna$_{7B}$	Q-Former	205M	8B	25M	18K	✗	✓
LLaVA-1.5-7B [47]	CLIP-ViT-L-336px	Llama2$_{7B}$	MLP	6.8B	7.1B	558k	665k	✗	✗
MiniGPT-4-7B [89]	BLIP2-VE	Llama2$_{7B}$	Linear	23M	7.7B	5M	3.5k	✗	✗
InstructBLIP-7B [13]	EVA-CLIP-ViT-g	Vicuna$_{7B}$	Q-Former	189M	7.9B	-	16M	✗	✗
LLaMA-Adapter-7B [22]	CLIP-ViT-L	LLaMA$_{7B}$	Early Fusion	14M	7.2B	567k	52k	✗	✗
Otter-I-7B [42]	CLIP-ViT-L	MPT$_{7B}$	Attention	1.4B	8.1B	-	2.8B	✗	✓
PandaGPT-7B [70]	ImageBind	Vicuna$_{7B}$	Linear + LLM LoRA	38M	7.9B	-	160k	✓	✓
mPLUG-owl-7B [81]	CLIP-ViT-L	LLaMA$_{7B}$	Attention	4M	7.1B	204M	158k	✗	✗
Video-LLaVA-7B [45]	LanguageBind	Vicuna$_{7B}$	Linear	6.8B	7.5B	1260k	765k	✗	✓
LLaVA-7B [48]	CLIP-ViT-L	Llama2$_{7B}$	Linear	6.7B	7.1B	595k	158k	✗	✗
ShareGPT4V-7B [10]	CLIP-ViT-L-336px	Vicuna$_{7B}$	MLP	6.7B	6.7B	1.2M	665k	✗	✗
∼13B Models									
InstructBLIP-13B [13]	EVA-CLIP-ViT-g	Vicuna$_{13B}$	Q-Former	189M	14.2B	-	16M	✗	✗
PandaGPT-13B [70]	ImageBind	Vicuna$_{13B}$	Linear+LLM LoRA	52M	13.1B	-	160k	✓	✓
LLaVA-13B-Vicuna [48]	CLIP-ViT-L-336px	Vicuna$_{13B}$	Linear	13.0B	13.3B	595k	158k	✗	✗
BLIP-2-11B [43]	EVA-CLIP-ViT-g	FlanT5$_{XXL}$	Q-Former	108M	12.2B	129M	-	✗	✗
InstructBLIP-11B [13]	EVA-CLIP-ViT-g	FlanT5$_{XXL}$	Q-Former	189M	12.3B	-	16M	✗	✗
LLaVA-13B-Llama2 [48]	CLIP-ViT-L	Llama2$_{13B}$	Linear	13.0B	13.3B	595k	158k	✗	✗
LLaVA-1.5-13B [47]	CLIP-ViT-L-336px	Llama2$_{13B}$	MLP	13.0B	13.4B	558k	665k	✗	✗

Table 2. Statistics of compared API-based and open-source VLMs, where TTP and ToP indicate Total Trainable Parameters and Total Parameters, respectively. Moreover, EgoData and Video indicate that there are egocentric visual data and video data for training, respectively.

object-state transformation or hand-object interactions.

- **Planning: How will I do?** In reality, planning [1, 30, 69] is an important capability to deal with complex problems, typically applied in *Navigation* [62, 63, 72] and *Assistance* [27, 76]. Navigation is going to a goal location from a start position, while assistance is offering instructions to solve daily problems.

3.2. Data Collection

In this section, we mainly introduce the detailed processing to construct our EgoThink benchmark.

Collecting first-person visual data. Firstly, we leverage a popular and large egocentric video dataset Ego4D [25], which is designed to advance the field of first-person perception in computer vision. To obtain a diverse representation in different scenarios, Ego4D encompasses 3,670 hours of video from 931 unique camera wearers spanning 74 global locations across 9 countries. To collect first-person visual data, we begin by extracting every frame from a subset of the Ego4D video dataset, yielding a diverse raw image dataset. Please note that our current focus is solely on images, as most VLMs today do not support video input. We intend to expand our scope to include videos in our future work. Considering the heavy human labor and the diversity of scenarios, we sample images every few dozen frames.

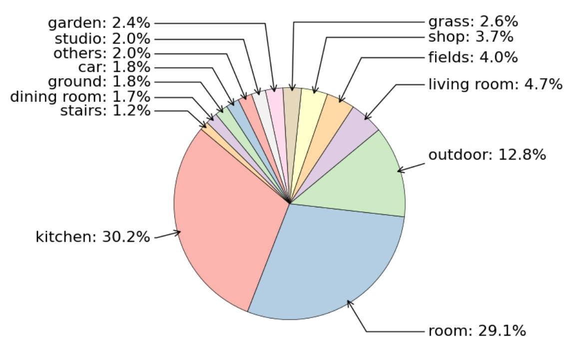


Figure 3. This chart illustrates the distribution of various scene categories within the EgoThink dataset. The 'others' category encompasses 13 different scene types, each representing less than one percent of total scenes.

To ensure high quality, we apply strict criteria for selecting the extracted frames. We first exclude images that lack clarity or fail to exhibit egocentric characteristics. Then, to obtain the high diversity within the dataset, we conduct a further screening to ensure that at most two images per video are included in the filtered image set. Finally, we obtain enormous high-quality images with exhibit egocentric characteristics as first-person image candidates.

Annotating question-answer pairs. Upon receiving a substantial collection of first-person image candidates, we

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													
OpenFlamingo-7B	16.0	55.0	37.0	15.0	34.0	34.0	21.0	40.0	21.0	31.0	11.0	11.0	27.2
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
VideoChat-7B	46.0	44.0	36.0	45.0	61.0	42.0	36.0	39.0	32.0	26.5	13.0	21.0	36.8
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
LLaMA-Adapter-7B	37.0	60.0	46.0	34.5	48.0	51.0	29.0	39.0	25.0	41.5	42.0	57.0	42.5
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 3. 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.

engage six annotators to manually label question-answer pairs. Given that the EgoThink benchmark is composed of twelve dimensions, annotators were responsible for two specific dimensions. The annotators can access all the image candidates and are asked to select appropriate images to annotate their corresponding question-answer pairs to relevant categories. Once the image is selected, it will be removed from the candidates to ensure no repetition. Moreover, to ensure the correctness of our annotations, we have three additional annotators to review the question-answer pairs after the first annotation process. The annotation will not be reserved until the three annotators all agree that the first-person visual data and the assigned question-answer pairs meet the definition of a specific dimension.

Statistics. The EgoThink benchmark comprises a collection of 700 images across six categories with twelve fine-grained dimensions. These images are extracted from 595 videos, ensuring a broad representation of scenarios. To guarantee diversity, a wide range of scenes and concepts has been deliberately selected. As depicted in Figure 3, the dataset encompasses a diverse range of scenes, covering key scenarios relevant to EAI. Furthermore, we have meticulously crafted question and answer for each image in the EgoThink benchmark, aiming to closely replicate real-life conversations. This involves employing different question types, varying questions in length and complexity, paired with well-reasoned and accurate answers. In addition, the size of the dataset is a well-balanced trade-off between benchmark diversity and high cost of open-ended QA evaluation. We ensure our 700 examples were diverse in concepts, scenes, and videos to provide a robust performance estimation within practical limits. Detailed statistics of the EgoThink benchmark are presented in Appendix A.

4. Experiments

4.1. Experimental Setups

Vision-Language Models. We collect the most popular eighteen types of representative VLMs to assess as shown in Table 2. Due to the possible effects of model parameters, we divide models into ∼7B and ∼13B for a fair comparison. Detailed information about VLMs can be found in Appendix B. We conduct zero-shot setups for all VLMs across our EgoThink benchmark. The prompts used for each VLM are shown in Appendix C.

Single-answer grading. Considering evaluating open-ended model generations is not a trivial problem [54, 68], we propose to use GPT-4 [60] as the automatic evaluator [86] to better measure generated answers. In this protocol, we want to measure how close one model output

Image	**Original Question & Answer**
	*Question:* Am I holding chopsticks in my right hand? *Answer:* Yes.

Model Predictions

`GPT-4V:` No, (you are not holding chopsticks in your right hand. It looks like you are holding a pair of tongs.)
`InstructBLIP-11B:` Yes.
`LLaVA-7B:` Yes, you are holding chopsticks in your right hand.

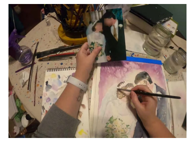
Changed Question Format & Answer

Question: What am I holding in my right hand?
Answer: Chopsticks.

Model Predictions

`GPT-4V:` In your right hand, you're holding what appears to be a pan with a handle, and it looks like you're cooking some food, possibly a stir-fry or sauteed vegetables.

Figure 4. Case study (wrong) of GPT-4V in the existence dimension of the object ability. In both the question formats of yes/no and what, GPT-4V can not correctly detect the chopsticks in my right hand, while InstructBLIP-11B and LLaVA-7B can.

Image	**Question & Answer**
	*Question:* What is the shape of the object I am holding in my left hand? *Answer:* Rectangular.

Model Predictions

`GPT-4V:` Rectangular.
`InstructBLIP-11B:` Pen.
`BLIP-2-11B:` A paint brush.
`PandaGPT-13B:` I am holding a paintbrush in my left hand.

Image	**Question & Answer**
	*Question:*What am I doing? *Answer:* Peeling the avocado.

Model Predictions

`GPT-4V:`Peeling vegetables.
`BLIP-2-11B:` Slicing an avocado.
`InstructBLIP-11B:` Chopping.
`LLaVA-13B-Vicuna:` You are cutting up an avocado in a kitchen sink.

Figure 5. Case studies (wrong) in the attribute, affordance, and activity dimensions. The top case demonstrates some VLMs locate the wrong place in the attribute and affordance dimension. The bottom case illustrates that in the activity dimension, some VLMs fail to detect the specific action.

is to the reference. Different from traditional similarity-based methods, GPT-4 pays more attention to semantics. In the detailed implementation, we format the question, the model output, and the reference in a prompt as shown in Appendix D and feed it into the GPT-4 evaluator. The GPT-4 evaluator is asked to assign a score of 0 (wrong), 0.5 (partially correct), or 1 (correct) to the model output. Additionally, we further discuss to use of GPT-3.5-Turbo, Claude-2, and humans as evaluators in Section 5.2.

4.2. Results

We first present the overall results of the evaluated models on our EgoThink benchmark as shown in Table 3. 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, only falling short in localization dimension compared to BLIP-2-11B. We will further introduce the detailed scores across different dimensions. More case studies can be found in Appendix E.

Results on object. In detail, we evaluate through three dimensions, including existence, attribute, and affordance. For existence, InstructBLIP-11B and ShareGPT4V-7B achieve the top-2 performance, indicating that they can predict the object precisely from the first-person perspective. As for GPT-4V, as illustrated in Figure 4, we observe that its performance in handed object detection leaves room for improvement. As for both attribute and affordance, the GPT-4V model has demonstrated superior performance, especially in the attribute dimension. In both dimensions, some open-source models as shown at the top of Figure 5 locate the wrong place or only answer the type of the object rather than its attribute or affordance.

Results on activity. The performance of GPT-4V outperforms all open-source VLMs in the activity dimension. Among the ∼7B models, ShareGPT4V-7B and mPLUG-owl-7B significantly outperform other VLMs and even achieve superior or comparable performance to ∼13B models. Overall, ∼13B models tend to perform better than ∼7B model in the activity dimension, but their scores are just below the passing line. The most possible reason is that detecting the specific action is difficult for VLMs as shown at the bottom of Figure 5.

Results on localization. In general, BLIP-2-11B has shown obvious advantages among all VLMs, even surpassing GPT-4V in both location and spatial relationship dimensions. In the location dimension, BLIP-2-11B, GPT-4V, and InstructBLIP-11B demonstrate superior ability to achieve

Image	Question & Answer
	*Question:* How many plates are there on my left? *Answer:* One.

Model Predictions

GPT-4V: One plate.
BLIP-2-6.7B: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15
BLIP-2-11B: 2
InstructBLIP-11B: 2
LLaVA-7B: There are two plates on my left.
PandaGPT-13B: There are two plates on my left.

Image	Question & Answer
	*Question:* Why am I putting my hand there? *Answer:* To feel the temperature of the pan.

Model Predictions

GPT-4V: To feel the pan's warmth.
InstructBLIP-7B: cleaning pots and pans.
InstructBLIP-11B: Cook.
BLIP-2-11B: I am putting my hand there to stir the pots and pans.

Figure 6. Case studies (wrong) in counting (top) and situated reasoning (bottom) dimensions. The top case shows VLMs can count but can not understand their relative position. The bottom case displays that the question requires commonsense knowledge and image understanding where only GPT-4V can answer it correctly.

around 90 points. However, perceiving the spatial relationship of an object relative to oneself is much more difficult. This phenomenon can be also reflected in the top of Figure 5 where VLMs hard to distinguish left or right hand.

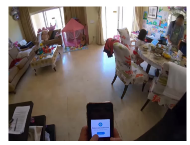
Results on reasoning. Counting is the most difficult ability [80] among all evaluated dimensions. The best-performing model, GPT-4V, only scores 42.0, far away from satisfaction. Under the first-person perspective setup, VLMs need to not only count but also understand the relative position to oneself, as shown in the top case of Figure 6. Meanwhile, the comparison dimension also reflects the high difficulty, where the best score of 56.0 is obtained by LLaVA-1.5-13B. As for situated reasoning, GPT-4V demonstrates its strong commonsense reasoning ability to answer complex questions at the bottom of Figure 6.

Results on forecasting. Achieving high performance seems to be challenging as the best score achieved by GPT-4V is only 55.0. InsturctBLIP-11B achieves a relatively high score of 53.0 which is close to that of GPT-4V. We observe that the VLMs mainly suffer from two problems: recognizing objects incorrectly or forecasting too far as shown in the top of Figure 7.

Image	Question & Answer
	*Question:* What will I do next? *Answer:* Open the car door.

Model Predictions

GPT-4V: You might open the car door.
LLaVA-7B: Next, I will wipe down the car to remove any dirt or debris.
InstructBLIP-11B: Wash car.
mPLUG-owl-7B: I will continue to clean the car's exterior and interior.

Image	Question & Answer
	*Question:* How to get outside? *Answer:* Walk forward to the end, turn right, and walk a few steps until the glass door is on the left. Turn left, open the glass door, and walk out of the room.

Model Predictions

GPT-4V: To get outside, you might want to exit through the door which is visible in the background, leading from the living space to the outdoor area. Please ensure that you have everything you need before leaving and that it's safe to go outside.
LLaMA-Adapter-7B: You can use the door in the room to exit the house.
LLaVA-7B: To get outside, the person in the image needs to open the sliding glass door.
InstructBLIP-11B: Door.

Figure 7. Case studies (wrong) in the forecast (top) and navigation dimension (bottom). The top case shows VLMs might detect the glove as a cloth, while the bottom case indicates VLMs lack navigation details and overlook image information.

Results on planning. In both navigation and assistance dimensions, the highest scores are achieved by GPT-4V with 60.0 and 84.0, respectively. LLaVA-13B-Llama2 behaves well in both dimensions with the second-best performance but its score is still 10 points lower than that of GPT-4V. The most possible reason is that answers provided by most open-source VLMs lack crucial details or overlook important information given in the images, as illustrated at the bottom of Figure 7.

5. Analysis

5.1. Effects of Components

As shown in Table 2, VLMs consist of multiple key components. In this section, we probe the influence of different components on our EgoThink benchmark.

The total parameters of LLMs. Here we compare the performance of ~7B and ~13B variants of four VLMs. Note

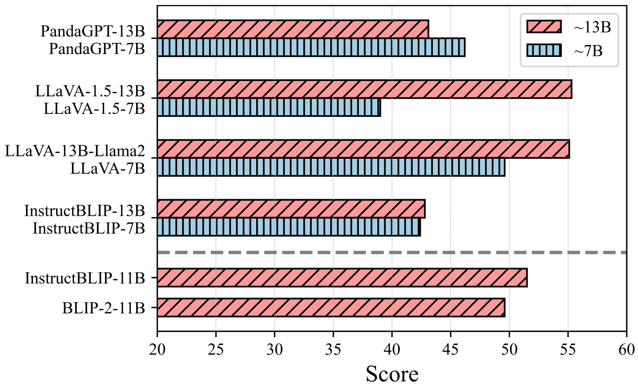


Figure 8. Impact of LLMs sizes (above the dash-line) and instruction-tuning (below the dash-line) on model performance. Average scores across all capabilities are reported.

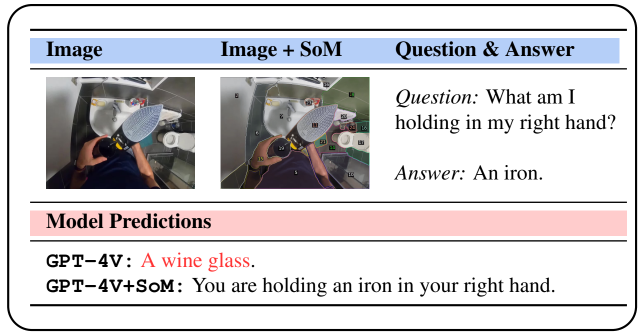


Figure 9. Case study (wrong) in the adding visual grounding information with images. The segmentation can help VLMs better locate the objects in question.

that the increase in the number of parameters mainly falls in the LLMs. Firstly, as shown in the top part of Figure 8, scaling does not lead to significant improvement for PandaGPT and InstructBLIP, while LLaVA (LLaVA-7B and LLaVA-13B-Llama2) and LLaVA-1.5 benefit a lot from scaling. We hypothesize that this is because LLaVA series models do not freeze their language models during instruction tuning, indicating that enlarging the number of trainable parameters can help improve both performance and generalization. In other words, one can see that simply scaling up language models without better alignment may not help.

Instruction tuning. We directly compare the performance of BLIP-2-11B and InstructBLIP-11B, because these two models differ only in instruction tuning and additional instruction-aware tokens. As presented in the bottom part of Figure 8, InstructBLIP-11B outperforms BLIP-2-11B after instruction tuning, despite an unexpectedly small margin. This may be because much of the instruction tuning data employed by InstructBLIP is collected from specific downstream tasks, whose data distributions are very different from our first-person perspective data.

The information of image encoder. Considering that there is no ablation version of VLMs for image encoder, following Set-of-Mark [79], we probe the effect of visual grounding information (i.e., a set of marks) in our setups. As presented in Figure 9, GPT-4V with additional segmentation information can correctly detect the mentioned location and objects, indicating that supplemented image information can be helpful in some situations. More discussion about quantitative experiments can be found in Appendix F.

5.2. Agreements between Human and Evaluators

In this section, we further assess the model performance on object and planning dimensions using GPT-3.5-Turbo, Claude-2, and human annotators. Due to the heavy human labor, we ask three annotators to evaluate the performance of GPT-4V, which is the overall best model. Human annotators consider the following aspects to evaluate: accuracy, completeness, logical soundness, and grammatical correctness. Our annotation system and detailed guidelines can be found in Appendix G. We further conduct GPT-3.5-Turbo and Claude-2 with the same evaluation prompt as GPT-4. The Pearson correlation coefficients between automatic evaluators (i.e., GPT-4, GPT-3.5-Turbo, Claude-2) and humans are 0.68, 0.43, and 0.68, respectively. The Cohen's Kappa coefficient among the three annotators is 0.81. This shows that evaluations made by GPT-4 and Claude-2 have a high correlation with humans. We hypothesize that recent well-performant LLMs can evaluate highly aligned with humans, given that most answers in our benchmark are relatively short and precise. Detailed scores of all evaluators and their correlations are discussed in Appendix H.

6. Conclusion

To pave the way for the development of VLMs in the field of EAI and robotics, we introduce a comprehensive benchmark, EgoThink. Designed to evaluate the capacity of VLMs to "think" from a first-person perspective, EgoThink encompasses six core capabilities across twelve detailed dimensions. We assess eighteen popular VLMs and find that even the top-performing VLMs in most dimensions achieve only around a score of 60. GPT-4V achieves the best overall performance, but can not consistently surpass other open-source VLMs across all dimensions. In the analysis, we further probe the impact of various components on model performance and find that the total number of trainable parameters in LLMs has the most significant effect. Despite the human agreement with automatic evaluators being high, the evaluation of planning is difficult due to the detailed information in the answers. In future research, we aim to improve the evaluation method and further explore the essential capabilities of VLMs in the EAI and robotics fields.

Acknowledgment: The work is supported by the National Key R&D Program of China (2022ZD0160502) and the National Natural Science Foundation of China (No.61925601).

References

[1] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, et al. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*, 2022. 4

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022. 1, 2, 4

[3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 2

[4] Anas Awadalla, Irena Gao, Joshua Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. Openflamingo, 2023. 1

[5] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*, 2023. 4

[6] Sven Bambach, Stefan Lee, David J Crandall, and Chen Yu. Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions. In *Proceedings of the IEEE international conference on computer vision*, pages 1949–1957, 2015. 3

[7] William Berrios, Gautam Mittal, Tristan Thrush, Douwe Kiela, and Amanpreet Singh. Towards language models that can see: Computer vision through the lens of natural language. *arXiv preprint arXiv:2306.16410*, 2023. 2

[8] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 1, 2

[9] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3558–3568, 2021. 2

[10] Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*, 2023. 4, 1

[11] Liang Chen, Yichi Zhang, Shuhuai Ren, Haozhe Zhao, Zefan Cai, Yuchi Wang, Peiyi Wang, Tianyu Liu, and Baobao Chang. Towards end-to-end embodied decision making via multi-modal large language model: Explorations with gpt4-vision and beyond, 2023. 2

[12] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, 2023. 2

[13] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023. 1, 2, 4

[14] Tien Do, Ondrej Miksik, Joseph DeGol, Hyun Soo Park, and Sudipta N Sinha. Learning to detect scene landmarks for camera localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11132–11142, 2022. 3

[15] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023. 1

[16] Chenyou Fan. Egovqa-an egocentric video question answering benchmark dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 2

[17] Ali Farhadi, Ian Endres, Derek Hoiem, and David Forsyth. Describing objects by their attributes. In *2009 IEEE conference on computer vision and pattern recognition*, pages 1778–1785. IEEE, 2009. 3

[18] Alireza Fathi, Ali Farhadi, and James M Rehg. Understanding egocentric activities. In *2011 international conference on computer vision*, pages 407–414. IEEE, 2011. 3

[19] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023. 2

[20] Antonino Furnari, Sebastiano Battiato, Kristen Grauman, and Giovanni Maria Farinella. Next-active-object prediction from egocentric videos. *Journal of Visual Communication and Image Representation*, 49:401–411, 2017. 3

[21] Jensen Gao, Bidipta Sarkar, Fei Xia, Ted Xiao, Jiajun Wu, Brian Ichter, Anirudha Majumdar, and Dorsa Sadigh. Physically grounded vision-language models for robotic manipulation. *arXiv preprint arXiv:2309.02561*, 2023. 1

[22] Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*, 2023. 4, 1

[23] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017. 2

[24] Stéphane Grade, Mauro Pesenti, and Martin G Edwards. Evidence for the embodiment of space perception: concurrent hand but not arm action moderates reachability and egocen-

tric distance perception. *Frontiers in Psychology*, 6:862, 2015. 3

[25] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022. 3, 4

[26] Jean-Bernard Hayet, Frédéric Lerasle, and Michel Devy. Visual landmarks detection and recognition for mobile robot navigation. In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, pages II–II. IEEE, 2003. 3

[27] Evert Helms, Rolf Dieter Schraft, and M Hagele. rob@work: Robot assistant in industrial environments. In *Proceedings. 11th IEEE International Workshop on Robot and Human Interactive Communication*, pages 399–404. IEEE, 2002. 4

[28] Tucker Hermans, James M Rehg, and Aaron Bobick. Affordance prediction via learned object attributes. In *IEEE international conference on robotics and automation (ICRA): Workshop on semantic perception, mapping, and exploration*, pages 181–184. Citeseer, 2011. 3

[29] MD Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. A comprehensive survey of deep learning for image captioning. *ACM Computing Surveys (CsUR)*, 51(6):1–36, 2019. 2

[30] Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, et al. Inner monologue: Embodied reasoning through planning with language models. *arXiv preprint arXiv:2207.05608*, 2022. 4

[31] Wenlong Huang, Chen Wang, Ruohan Zhang, Yunzhu Li, Jiajun Wu, and Li Fei-Fei. Voxposer: Composable 3d value maps for robotic manipulation with language models. *arXiv preprint arXiv:2307.05973*, 2023. 1

[32] Yichao Huang, Xiaorui Liu, Xin Zhang, and Lianwen Jin. A pointing gesture based egocentric interaction system: Dataset, approach and application. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 16–23, 2016. 2

[33] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019. 2

[34] Baoxiong Jia, Ting Lei, Song-Chun Zhu, and Siyuan Huang. Egotaskqa: Understanding human tasks in egocentric videos. *Advances in Neural Information Processing Systems*, 35: 3343–3360, 2022. 2

[35] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 2

[36] Charmi Jobanputra, Jatna Bavishi, and Nishant Doshi. Human activity recognition: A survey. *Procedia Computer Science*, 155:698–703, 2019. 3

[37] Nancy Kanwisher and Jon Driver. Objects, attributes, and visual attention: Which, what, and where. *Current Directions in Psychological Science*, 1(1):26–31, 1992. 3

[38] Shian-Ru Ke, Hoang Le Uyen Thuc, Yong-Jin Lee, Jenq-Neng Hwang, Jang-Hee Yoo, and Kyoung-Ho Choi. A review on video-based human activity recognition. *Computers*, 2(2):88–131, 2013. 3

[39] Roberta L Klatzky. Allocentric and egocentric spatial representations: Definitions, distinctions, and interconnections. In *Spatial cognition: An interdisciplinary approach to representing and processing spatial knowledge*, pages 1–17. Springer, 1998. 3

[40] Weicheng Kuo, Yin Cui, Xiuye Gu, AJ Piergiovanni, and Anelia Angelova. F-vlm: Open-vocabulary object detection upon frozen vision and language models. *arXiv preprint arXiv:2209.15639*, 2022. 1

[41] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Mimic-it: Multi-modal in-context instruction tuning. *arXiv preprint arXiv:2306.05425*, 2023. 2

[42] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726*, 2023. 1, 2, 4

[43] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 1, 4

[44] KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023. 4, 1

[45] Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023. 4, 1

[46] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 2

[47] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023. 4, 1

[48] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 1, 2, 4

[49] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023. 2

[50] Li Liu, Wanli Ouyang, Xiaogang Wang, Paul Fieguth, Jie Chen, Xinwang Liu, and Matti Pietikäinen. Deep learning for generic object detection: A survey. *International journal of computer vision*, 128:261–318, 2020. 3

[51] Miao Liu, Siyu Tang, Yin Li, and James M Rehg. Forecasting human-object interaction: Joint prediction of motor attention and egocentric activity. *Computer Vision–ECCV 2020*, 12346:704–721, 2020. 3

[52] Shaowei Liu, Subarna Tripathi, Somdeb Majumdar, and Xiaolong Wang. Joint hand motion and interaction hotspots prediction from egocentric videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3282–3292, 2022. 3

[53] Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, and Li Yi. Hoi4d: A 4d egocentric dataset for category-level human-object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21013–21022, 2022. 2, 3

[54] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023. 2, 5

[55] Michael Montemerlo and Sebastian Thrun. *FastSLAM: A scalable method for the simultaneous localization and mapping problem in robotics*. Springer, 2007. 3

[56] Luis Montesano, Manuel Lopes, Alexandre Bernardino, and José Santos-Victor. Learning object affordances: from sensory–motor coordination to imitation. *IEEE Transactions on Robotics*, 24(1):15–26, 2008. 3

[57] Francesca Morganti, Stefano Stefanini, and Giuseppe Riva. From allo-to egocentric spatial ability in early alzheimer's disease: a study with virtual reality spatial tasks. *Cognitive neuroscience*, 4(3-4):171–180, 2013. 3

[58] Weimin Mou, Timothy P McNamara, Christine M Valiquette, and Björn Rump. Allocentric and egocentric updating of spatial memories. *Journal of experimental psychology: Learning, Memory, and Cognition*, 30(1):142, 2004. 3

[59] Thi-Hoa-Cuc Nguyen, Jean-Christophe Nebel, and Francisco Florez-Revuelta. Recognition of activities of daily living with egocentric vision: A review. *Sensors*, 16(1):72, 2016. 3

[60] OpenAI. Gpt-4 technical report, 2023. 1, 2, 4, 5

[61] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744, 2022. 2

[62] Anish Pandey, Rakesh Kumar Sonkar, Krishna Kant Pandey, and DR Parhi. Path planning navigation of mobile robot with obstacles avoidance using fuzzy logic controller. In *2014 IEEE 8th international conference on intelligent systems and control (ISCO)*, pages 39–41. IEEE, 2014. 4

[63] BK Patle, Anish Pandey, DRK Parhi, AJDT Jagadeesh, et al. A review: On path planning strategies for navigation of mobile robot. *Defence Technology*, 15(4):582–606, 2019. 4

[64] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015. 2

[65] Vignesh Ramanathan, Anmol Kalia, Vladan Petrovic, Yi Wen, Baixue Zheng, Baishan Guo, Rui Wang, Aaron Marquez, Rama Kovvuri, Abhishek Kadian, et al. Paco: Parts and attributes of common objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7141–7151, 2023. 3

[66] Sajad Saeedi, Michael Trentini, Mae Seto, and Howard Li. Multiple-robot simultaneous localization and mapping: A review. *Journal of Field Robotics*, 33(1):3–46, 2016. 3

[67] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VIII*, pages 146–162. Springer, 2022. 2

[68] Wenqi Shao, Yutao Hu, Peng Gao, Meng Lei, Kaipeng Zhang, Fanqing Meng, Peng Xu, Siyuan Huang, Hongsheng Li, Yu Qiao, et al. Tiny lvlm-ehub: Early multimodal experiments with bard. *arXiv preprint arXiv:2308.03729*, 2023. 2, 5

[69] Chan Hee Song, Jiaman Wu, Clayton Washington, Brian M Sadler, Wei-Lun Chao, and Yu Su. Llm-planner: Few-shot grounded planning for embodied agents with large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2998–3009, 2023. 4

[70] Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. Pandagpt: One model to instruction-follow them all. *arXiv preprint arXiv:2305.16355*, 2023. 4, 1

[71] Theodore Sumers, Kenneth Marino, Arun Ahuja, Rob Fergus, and Ishita Dasgupta. Distilling internet-scale vision-language models into embodied agents. *arXiv preprint arXiv:2301.12507*, 2023. 1

[72] Loreto Susperregi, Izaskun Fernandez, Ane Fernandez, Santiago Fernandez, Iñaki Maurtua, and Irene Lopez de Vallejo. Interacting with a robot: a guide robot understanding natural language instructions. In *International Conference on Ubiquitous Computing and Ambient Intelligence*, pages 185–192. Springer, 2012. 4

[73] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 1

[74] Michalis Vrigkas, Christophoros Nikou, and Ioannis A Kakadiaris. A review of human activity recognition methods. *Frontiers in Robotics and AI*, 2:28, 2015. 3

[75] Guan Wang, Sijie Cheng, Xianyuan Zhan, Xiangang Li, Sen Song, and Yang Liu. Openchat: Advancing open-source language models with mixed-quality data. *arXiv preprint arXiv:2309.11235*, 2023. 2

[76] Weronika Wojtak, Flora Ferreira, Paulo Vicente, Luís Louro, Estela Bicho, and Wolfram Erlhagen. A neural integrator model for planning and value-based decision making of a robotics assistant. *Neural Computing and Applications*, 33: 3737–3756, 2021. 4

[77] Peng Xu, Wenqi Shao, Kaipeng Zhang, Peng Gao, Shuo Liu, Meng Lei, Fanqing Meng, Siyuan Huang, Yu Qiao, and Ping Luo. Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models. *arXiv preprint arXiv:2306.09265*, 2023. 2

[78] Jingkang Yang, Yuhao Dong, Shuai Liu, Bo Li, Ziyue Wang, Chencheng Jiang, Haoran Tan, Jiamu Kang, Yuanhan Zhang, Kaiyang Zhou, et al. Octopus: Embodied vision-language programmer from environmental feedback. *arXiv preprint arXiv:2310.08588*, 2023. 1

[79] Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v. *arXiv preprint arXiv:2310.11441*, 2023. 8

[80] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 2023. 1, 7

[81] 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. mplug-owl: Modularization empowers large language models with multimodality, 2023. 1, 2, 4

[82] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4651–4659, 2016. 2

[83] Yifan Zhang, Congqi Cao, Jian Cheng, and Hanqing Lu. Egogesture: A new dataset and benchmark for egocentric hand gesture recognition. *IEEE Transactions on Multimedia*, 20(5):1038–1050, 2018. 2

[84] Tiancheng Zhao, Tianqi Zhang, Mingwei Zhu, Haozhan Shen, Kyusong Lee, Xiaopeng Lu, and Jianwei Yin. Vl-checklist: Evaluating pre-trained vision-language models with objects, attributes and relations. *arXiv preprint arXiv:2207.00221*, 2022. 2

[85] Zhong-Qiu Zhao, Peng Zheng, Shou-tao Xu, and Xindong Wu. Object detection with deep learning: A review. *IEEE transactions on neural networks and learning systems*, 30 (11):3212–3232, 2019. 3

[86] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*, 2023. 5, 2

[87] Wangchunshu Zhou, Yan Zeng, Shizhe Diao, and Xinsong Zhang. Vlue: A multi-task benchmark for evaluating vision-language models. *arXiv preprint arXiv:2205.15237*, 2022. 2

[88] Chenchen Zhu, Fanyi Xiao, Andres Alvarado, Yasmine Babaei, Jiabo Hu, Hichem El-Mohri, Sean Culatana, Roshan Sumbaly, and Zhicheng Yan. Egoobjects: A large-scale egocentric dataset for fine-grained object understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20110–20120, 2023. 2, 3

[89] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 1, 2, 4

[90] Wanrong Zhu, Jack Hessel, Anas Awadalla, Samir Yitzhak Gadre, Jesse Dodge, Alex Fang, Youngjae Yu, Ludwig Schmidt, William Yang Wang, and Yejin Choi. Multimodal c4: An open, billion-scale corpus of images interleaved with text. *arXiv preprint arXiv:2304.06939*, 2023. 2

[91] Zhengxia Zou, Keyan Chen, Zhenwei Shi, Yuhong Guo, and Jieping Ye. Object detection in 20 years: A survey. *Proceedings of the IEEE*, 2023. 3