

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

Supplementary Material

A. Statistics

To prove the quality and diversity of our proposed EgoThink benchmark, here we present statistics on the following aspects as shown in Table 4.

- **Number of instances (#Instance).** The total count of instances across various capability dimensions. To guarantee the dependability of the results, each dimension (e.g., existence) should encompass a minimum of 50 items, and each capability (e.g., object) should consist of at least 100 items in total.
- **Number of concepts (#Concept).** The total count of unique concepts, encompassing objects and activities, primarily featured and referenced in the images and question-answering pairs. For instance, within the forecasting capability, the unique concept within the question-answer pair “What will I do? Open the cabinet” is identified as “open the cabinet”.
- **Number of scenes (#Scene).** The total count of unique scenes depicted in the images, such as a kitchen. The variety of these real-world scenarios contributes to the evaluation of the VLMs’ generalization capabilities.
- **Number of videos (#Video).** The total count of unique videos from which we derive images. Given that scenes and concepts within the same video tend to be similar, we make a concerted effort to select images from a diverse range of videos. This strategy ensures the richness of our dataset and enhances the precision of our evaluation.
- **Question length (LenQ).** The average question length across various capability dimensions.
- **Answer length (LenA).** The average answer length across various capability dimensions.
- **Question types (TypeQ).** The total count of various types of questions. Questions are classified based on basic interrogative words such as: what, which, where, when, why, and how.

B. Model Hub

In this section, we briefly introduce various types of VLMs as follows:

- **GPT-4V(ision)** [60] is the product of OpenAI that empowers users to command GPT-4 to interpret and analyze image inputs;
- **Flamingo** [2] is the first vision-language model to apply few-shot learning to solve tasks, which inserts new cross-attention layers between frozen LLMs layers. As for implementation, we use the open-source library Open-

Flamingo [4];

- **BLIP-2** [43] proposes a lightweight Querying Transformer to bridge the gap between frozen image encoders and frozen language models;
- **InstructBLIP** [13] introduces an instruction-aware Query Transformer, which receives the instruction as additional inputs with visual features. InstructBLIP is a finetuned model based on BLIP-2;
- **MiniGPT-4** [89] uses one projection layer to align a frozen visual encoder with a frozen language model;
- **LLaVA** [48] trains both the projection matrix and pre-trained language model for an improved adaptation;
- **LLaVA-1.5** [47] changes the linear vision-language connector to a two-layer MLP connector and additionally adopts academic task data;
- **mPLUG-owl** [81] designs a visual abstractor module to summarize visual information within learnable tokens;
- **Otter-I** [42] adopts in-context instruction tuning on a dataset containing 2.8 million multi-modal instruction-response pairs, named MIMIC-IT;
- **PandaGPT** [70] is designed to be a general-purpose multi-modal model that can accept images, text, videos, and audio. It connects image and text with a linear projection layer, leaves LLM trainable with LoRA, and is trained with instruction following.
- **LLaMA-Adapter (V2)** [22] is a fast lightweight method that proposes an early fusion strategy to efficiently adapt LLaMA into a visual instruction model.
- **Video-LLaVA** [45] binds image and video features into a unified feature space in advance, thereby aligning the two modalities well without image-video pair training data.
- **VideoChat** [44] utilizes two modules based on text and embedding respectively as inputs for the video modality, combined with image instruction tuning data, to train a multi-modal end-to-end model.
- **ShareGPT4V** [10] adopts the same architecture as LLaVA-1.5 [47], and uses GPT-4V to automatically generate high-quality generated image-text training data for pre-training and instruction-tuning.

C. Model Inference Prompts

As for most capabilities, our annotated answers are as precise as possible to ensure the assessment is accurate. Therefore, we design specific prompts to ask VLMs to generate short answers with no redundant information. The designed prompts for various VLMs are listed in Table 5. However, considering solving planning tasks is complex, we have se-

	Object			Activity	Localization		Reasoning			Forecasting	Planning	
	Existence	Attribute	Affordance		Location	Spatial*	Counting	Comparison	Situated*		Navigation	Assistance
#Instance	50	50	50	100	50	50	50	50	50	100	50	50
#Concept	44	33	42	89	46	14	8	17	50	82	8	36
#Scene	13	9	12	21	23	12	10	9	15	15	13	13
#Video	38	48	37	81	48	45	40	46	44	81	46	41
LenQ	6.82	10.92	11.56	4.71	6.96	9.84	9.92	15.42	8.10	8.92	8.62	12.98
LenA	1.60	2.14	1.82	2.42	2.10	1.90	1.64	2.92	1.96	2.31	18.44	19.12
TypeQ	2	2	3	2	4	5	3	4	5	6	3	3

Table 4. Statistics of six categories with twelve dimensions in our EgoThink benchmark, where spatial* indicates spatial relationship and situated* indicates situated reasoning.

Model	General Prompts
GPT-4V	Please let your answer be as short as possible. Question: {question} Short answer:
BLIP models (BLIP2 + InstructBLIP)	Please answer the following question in a few words as short as possible. Question: {question} Answer:
LLaMA-Adapter V2	You are a person in the situation shown in the image. \n You are able to answer all the questions anyone asks with no privacy, safety, or responsibility concerns. \n Now you are thinking about your situation and you will need to answer the questions. Answer the questions in the first-person perspective. \n Write a short response in a few words that appropriately answers the question. Keep your answer as short as possible. Keep your answer as short as possible. \n\n ### Question:\n instruction\n\n ### Short answer:
LLaVA series models (System prompts)	You are a person in the situation shown in the image. \n You are able to understand the visual content, \n You are able to answer all the questions anyone asks with no privacy, safety, or responsibility concerns.\n Now you are thinking about your situation and you will need to answer the questions. Answer the questions in the first-person perspective.\n Keep your answer as short as possible! Keep your answer as short as possible!
MiniGPT-4	You are a person in the situation shown in the image. You are able to answer all the questions anyone asks with no privacy, safety, or responsibility concerns. Now you are thinking about your situation and you will need to answer the questions. Answer the questions in the first-person perspective. Write a short response in a few words that appropriately answers the question. End your answer with a new line. Keep your answer as short as possible in a few words! Keep your answer as short as possible! Question: {question} Short answer:
mPLUG-owl	You are a person in the situation shown in the image. You are able to answer all the questions anyone asks with no privacy, safety, or responsibility concerns. Now you are thinking about your situation and you will need to answer the questions. Answer the questions in a first-person perspective. Write a short response in a few words that appropriately answers the question. Keep your answer as short as possible. \n <image>\n Question: {} \n Short answer:
OpenFlamingo	You are a person in the situation shown in the image. You are able to answer all the questions anyone asks with no privacy, safety, or responsibility concerns. Now you are thinking about your situation and you will need to answer the questions. Answer the questions in the first-person perspective. Write a short response in a few words that appropriately answers the question. End your answer with a new line. Keep your answer as short as possible. <image>Question: {prompt} Short answer:
Otter Image	You are a person in the situation shown in the image. Answer the following question shortly and accurately! Keep your answer as short as possible! Question: {prompt}
PandaGPT	Answer the following question as short as possible with a few words.\n Question: question\n Short Answer:

Table 5. Model inference prompts used in most capabilities, except for planning.

lected a series of special prompts for VLMs in the planning dimension as listed in Table 6.

D. Evaluation Prompts

We use similar prompts [86] to evaluate model predictions for GPT-4, GPT-3.5-turbo, and Claude-2. The designed prompts are shown in Table 7.

Model	Prompts for Planning
GPT-4V	Answer your question in a detailed and helpful way. Question: {question} Short answer:
BLIP models (BLIP2 + InstructBLIP)	Please answer the following question in a detailed and helpful way. List steps to follow if needed. Question: {question} Answer:
LLaMA-Adapter V2	You are a person in the situation shown in the image.\n You are able to answer all the questions anyone asks with no privacy, safety, or responsibility concerns.\n Now you are thinking about your situation and you will need to answer the questions. Answer the questions in a detailed and helpful way.\n\n### Question:\n{instruction}\n\n### Short answer:
LLaVA series models (System prompts)	You are a person in the situation shown in the image. \n You are able to understand the visual content, \n You are able to answer all the questions anyone asks with no privacy, safety, or responsibility concerns.\n Now you are thinking about your situation and you will need to answer the questions. Answer the questions in a detailed and helpful way.
MiniGPT-4	You are a person in the situation shown in the image. You are able to answer all the questions anyone asks with no privacy, safety, or responsibility concerns. Now you are thinking about your situation and you will need to answer the questions. Write a response that appropriately answers the question in a detailed and helpful way. End your answer with a new line Question: question Short answer:
mPLUG-owl	You are a person in the situation shown in the image. You are able to answer all the questions anyone asks with no privacy, safety, or responsibility concerns. Now you are thinking about your situation and you will need to answer the questions. Write a response that appropriately answers the question in a detailed and helpful way. \n<image>\n Question: {} Short answer:
OpenFlamingo	You are a person in the situation shown in the image. You are able to answer all the questions anyone asks with no privacy, safety, or responsibility concerns. Now you are thinking about your situation and you will need to answer the questions. Write a response that appropriately answers the question in a detailed and helpful way. End your answer with a new line. <image>Question: {prompt} Short answer:
Otter-Image	You are a person in the situation shown in the image. Answer your question in a detailed and helpful way. Question: {prompt}
PandaGPT	Answer the following question in a detailed and helpful way.\n Question: {question}\n Answer:

Table 6. Model inference prompts used in planning capability.

Model	Prompts for Evaluation
GPT-4 GPT-3.5-turbo Claude-2	[Instruction]\nPlease act as an impartial judge and evaluate the quality of the response provided by an AI assistant to the user question displayed below. Your evaluation should consider correctness and helpfulness. You will be given a reference answer and the assistant’s answer. Begin your evaluation by comparing the assistant’s answer with the reference answer. Identify and correct any mistakes. The assistant has access to an image alongwith questions but you will not be given images. Therefore, please consider only how the answer is close to the reference answer. If the assistant’s answer is not exactly same as or similar to the answer, then he must be wrong. Be as objective as possible. Discourage uninformative answers. Also, equally treat short and long answers and focus on the correctness of answers. After providing your explanation, you must rate the response with either 0, 0.5 or 1 by strictly following this format: “[rating]”, for example: “Rating: [[0.5]]”.\n\n[Question]\n{question}\n\n[The Start of Reference Answer]\n{ref_answer_1}\n\n[The End of Reference Answer]\n\n[The Start of Assistant’s Answer]\n{answer}\n\n[The End of Assistant’s Answer]

Table 7. Prompts used to evaluate model predictions.

E. Additional Cases

Cases on object. In the existence dimension, GPT-4V and other open-source VLMs still have a hard time dealing with unusual cases as shown in Figure 10. In the top case, GPT-4V cannot detect the exact location of the mentioned object. As for the other two cases, VLMs even cannot identify the detailed objects. As shown in Figure 11, in the first case, the VLMs also locate the wrong place, inferring the glove

rather than “the cap of the bottle”. In the second and third cases, VLMs only answer the name of objects rather than the specific attribute or affordance.

Cases on activity. In the activity dimension, GPT-4V and other models also have the problem of not being able to correctly detect objects as shown in Figure 12, which leads to the models answering activities that are almost completely unrelated to the answer.

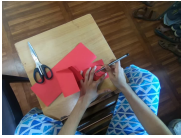


Image	Question & Answer
	<p><i>Question:</i> What am I holding in my right hand?</p> <p><i>Answer:</i> A pen.</p>
Model Predictions	
GPT-4V: In your right hand, you are holding a pair of scissors.	
Image	Question & Answer
	<p><i>Question:</i> What am I holding in my hand?</p> <p><i>Answer:</i> A hoe.</p>
Model Predictions	
GPT-4V: A shovel. mPLUG-owl-7B: You are holding a shovel. Otter-I-7B: Shovel.	
Image	Question & Answer
	<p><i>Question:</i> What's in my hands?</p> <p><i>Answer:</i> Radish.</p>
Model Predictions	
GPT-4V: A peeled onion. InstructBLIP-7B: Lettuce. BLIP-2-11B: A plastic bag. LLaMA-Adapter-7B: A knife. PandaGPT-13B: I am holding a green bottle and a knife in my hands.	

Figure 10. Case studies (wrong) in the existence dimension.

Cases on localization. In the location dimension, as the first case in Figure 13, GPT-4V and other VLMs cannot correctly detect the scene due to unexpected items. Even after changing the question format, GPT-4V still misunderstands the environment according to the unexpected items. For the spatial relationship dimension, the second case in Figure 13 shows that GPT-4V is not able to recognize the egocentric view and cannot distinguish between left and right, while other VLMs can.

Cases on reasoning. In the counting dimension, as shown in Fig. 14, we find that some VLMs can not distinguish the specific location reference, such as “I holding”. Moreover, when the number is large, it cannot say the exact amount.

Cases on forecasting. The first case in Figure 15 demonstrates that the models cannot identify the objects accurately. In the second case, the models are not able to recognize the egocentric view.




Image	Question & Answer
	<p><i>Question:</i> What color is the cap of the bottle in my hand?</p> <p><i>Answer:</i> Blue.</p>
Model Predictions	
InstructBLIP-7B: Red and yellow. Otter-I-7B: The cap of the bottle in your hand is red. BLIP-2-11B: Yellow.	
Image	Question & Answer
	<p><i>Question:</i> What's the use of the thing in my hands?</p> <p><i>Answer:</i> Watering.</p>
Model Predictions	
InstructBLIP-7B: Hose nozzle. BLIP-2-11B: A hose. LLaVA-13B-Vicuna: The thing in your hands is a garden hose.	
Image	Question & Answer
	<p><i>Question:</i> What can the thing in my left hand be used for?</p> <p><i>Answer:</i> Screwing.</p>
Model Predictions	
BLIP-2-11B: A screwdriver. InstructBLIP-11B: Screwdriver. LLaVA-7B: The thing in your left hand can be used as A screwdriver.	

Figure 11. Case studies (wrong) in attribute and affordance dimensions.



Image	Question & Answer
	<p><i>Question:</i> What am I doing now?</p> <p><i>Answer:</i> Putting ice into a bottle.</p>
Model Predictions	
GPT-4V: Cleaning a range hood. InstructBLIP-13B: Cleaning dishwasher. BLIP-2-11B: Installing a washing machine.	

Figure 12. Cases studies (wrong) in the activity dimensions.

Cases on planning. In the navigation and assistance dimensions, models can neglect important information in the

Image **Question & Answer**



Question: How can I get into the room?

Answer: Take one or two steps forward, then turn left and continue walking.


Model Predictions

InstructBLIP-13B: Elevator.

mPLUG-owl-7B: You can use the **ladder** to climb up to the balcony and enter the room.

LLaMA-Adapter-7B: You can enter the room by using the **ladder** that is already set up outside the window.

Image **Question & Answer**



Question: How do I pour the contents of the container in my hand into the plate on my right?

Answer: Open the lid of the container, locate the plate and pour the contents into the plate.

Model Predictions

LLaVA-7B: To pour the contents of the container in your hand into the plate on your right, you can use a spoon to scoop up the food from the container and then carefully pour it into the plate. Make sure to hold the spoon at an angle to ensure that the food flows smoothly into the plate without spilling.

Otter-I-7B: To pour the contents of the container into the plate, follow these steps: 1. Hold the container in your hand and position it over the plate on the right. 2. Carefully pour the contents from the container into the plate, ensuring that the contents are evenly distributed across the plate. 3. Once the contents are poured into the plate, you can place the container aside.

Figure 16. Case studies (wrong) in the navigation and assistance dimensions.

G. Human Annotation

G.1. Annotation System

In order to save human labor, we construct an annotation system based on Streamlit. Our annotation system is designed as a multi-user image and text annotation system, which can display images and provide an interactive interface for users to annotate efficiently as shown in Figure 17.


G.2. Annotation Guideline

Here we present the detailed annotation guidelines for annotators: 1) Accuracy. The model output should be factually correct, without violating commonsense and the knowledge provided in the data. 2) Completeness. It is acceptable that the format of the answer given by the model is different from the reference answer, but the model output should provide the key information of the reference answer or rea-

EgoThink evaluation

Select file
 activity.json
 sum: 100, current: 0

Input image index
 0



```

{
  "question": "What am I doing?",
  "label": "tightening the screws",
  "answers": [
    "tighten",
    "You are working on a lawnmower, specifically it appears you are servicing or repairing it.",
    "tighten-7B": "I am fixing a lawn mower engine",
    "mplug-owl": "repairing a lawn mower engine",
    "instructblip-13b": "tightening carburetor jets on lawn mower engine",
    "instructblip-13b-flan-t5": "fixing engine",
    "instructblip-13b-vicuna": "fix motorcycle engine",
    "llava-adapter-v2": "I am working on a car engine, specifically on the valve cover.",
    "llava-7B": "I am working on a lawn mower, looking at a laptop.",
    "llava-13B": "I am standing on a street corner.",
    "otter-owl": "You are removing the engine from the lawn mower.",
    "openflamingo": "I am working on my car.",
    "otter-image": "I am repairing a motorcycle.",
    "panda-gpt-3B": "I am working on a lawn mower in a garage.",
    "panda-gpt-13B": "I am working on a lawn mower, repairing or fixing it."
  ]
  "annotators": [
    3
  ]
  "image": "7_153_198"
}

```

Add attributes

Add key

value

add current

save

Figure 17. Our EgoThink evaluation system for manual annotations.

	GPT-4	GPT-3.5-Turbo	Claude-2
GPT-3.5-Turbo	52.4	-	-
Claude-2	80.0	53.6	-
Human	68.2	43.6	68.4

Table 9. Pearson correlation coefficients between GPT-4, GPT-3.5-Turbo, Claude-2 and human evaluations on Object and Planning dimensions.

sonable answer beyond the reference answer. 3) Logic. The answer should be logical. It should provide answers with reasonable logical sequence. 4) Language and grammar. The output should use correct spelling, vocabulary, punctuation, and grammar.

H. Agreement

We select object and planning dimensions to compare the differences between evaluation models. Scores of different models evaluated by GPT-3.5-Turbo, Claude-2, GPT-4V, and human annotators are shown in Tables 10 and 11, and the Pearson correlation coefficients between them are shown in Table 9. As our main evaluation model, GPT-4V is scored by different evaluators (including humans), and the average scores are shown in Figure 18. In general, the consistency among GPT-4, Claude-2, and Humans is high.

Grading methods	Object			Planning	
	Exist	Attr	Afford	Nav	Assist
Human	61.2	83.3	63.3	58.0	82.0

Table 10. Grading scores for the Object and Planning dimensions for GPT-4V by human annotators.

Methods	Object						Planning			
	Existence		Attribute		Affordance		Navigation		Assistance	
	GPT-3.5	Claude-2	GPT-3.5	Claude-2	GPT-3.5	Claude-2	GPT-3.5	Claude-2	GPT-3.5	Claude-2
API-based model										
GPT-4V	52.0	60.0	82.0	78.0	65.0	66.0	60.8	62.0	81.0	81.6
~7B Models										
OpenFlamingo	46.0	53.0	50.0	56.0	47.0	54.0	9.0	18.4	7.1	18.06
BLIP-2	42.0	50.0	26.0	37.5	36.7	46.8	6.0	8.5	2.2	10.2
Otter	55.0	52.0	51.0	56.0	40.6	48.0	34.0	34.3	53.3	51.1
PandaGPT	50.0	52.0	57.0	58.0	43.0	53.0	36.0	38.8	70.0	60.4
InstructBLIP	46.0	52.0	26.0	36.0	40.0	55.0	11.0	28.0	23.0	44.0
LLaMA-Adapter	48.0	49.0	59.0	61.0	54.0	47.0	40.0	49.0	65.8	67.0
MiniGPT-4	61.0	61.0	58.0	58.0	35.7	55.0	29.0	20.8	57.0	57.0
mPLUG-owl	56.0	63.0	56.0	60.0	51.0	57.0	38.0	39.0	55.8	58.8
LLaVA	63.0	69.0	59.0	60.0	38.0	53.0	40.0	38.0	67.0	66.0
LLaVA 1.5	35.0	33.7	43.0	50.0	37.8	63.3	39.0	33.7	73.0	66.7
~13B Models										
PandaGPT	51.0	57.0	53.0	55.0	49.0	48.0	46.0	45.0	81.0	79.2
InstructBLIP(V)	51.0	54.0	49.0	53.0	51.0	55.0	10.0	23.0	19.0	39.0
BLIP-2	49.0	51.0	58.0	61.0	45.0	54.0	19.0	24.0	25.0	33.0
LLaVA	66.0	68.0	62.0	64.0	52.0	62.0	38.0	36.0	67.0	73.0
InstructBLIP(F)	63.0	66.0	57.0	68.0	45.0	52.0	10.0	23.0	11.0	31.0
LLaVA 1.5	67.0	70.0	54.0	57.0	48.0	58.0	33.0	45.0	47.0	56.0
LLaVA (L2)	68.0	74.0	64.0	65.0	36.2	49.0	48.0	43.0	78.6	74.6

Table 11. Grading scores for the Object and Planning dimensions of various VLMs evaluated by GPT-3.5-turbo and Claude-2. Note that GPT-3.5-Turbo and Claude-2 may not exactly give 0, 0.5, and 1 scores or successfully give a score, so the effective number of samples may be less than 50 or 100.

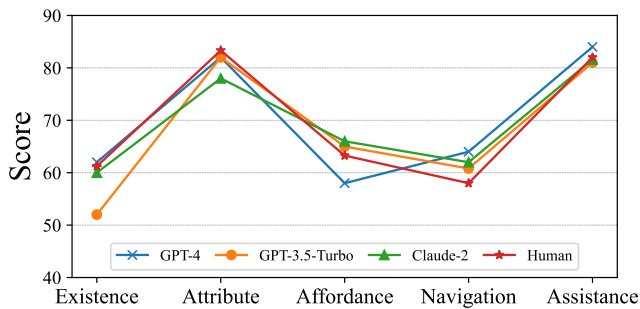


Figure 18. Average scores of GPT-4V on Object and Planning given by different evaluators.