EXPLORE-Bench: Egocentric Scene Prediction with Long-Horizon Reasoning

Chengjun Yu[1,*], Xuhan Zhu[2,3,*], Chaoqun Du[2], Pengfei Yu[2,†], Wei Zhai[1,†], Yang Cao[1], and Zheng-Jun Zha[1]

[1] University of Science and Technology of China, China
[2] Foundation Model Team, Li Auto Inc., China
[3] Tsinghua University, China *Equal Contribution †Corresponding Author
{yucj@mail., wzhai056@, forrest@, zhazj@}ustc.edu.cn
{zhuxuhan, duchaoqun, yupengfei1}@lixiang.com
https://github.com/JackYu6/EXPLORE-Bench/

Abstract. Multimodal large language models (MLLMs) are increasingly considered as a foundation for embodied agents, yet it remains unclear whether they can reliably reason about the long-term physical consequences of actions from an egocentric viewpoint. We study this gap through a new task, **E**gocentric **S**cene **P**rediction with **LO**ng-horizon **RE**asoning: given an initial-scene image and a sequence of atomic action descriptions, a model is asked to predict the final scene after all actions are executed. To enable systematic evaluation, we introduce **EXPLORE-Bench**, a benchmark curated from real first-person videos spanning diverse scenarios. Each instance pairs long action sequences with structured final-scene annotations, including object categories, visual attributes, and inter-object relations, which supports fine-grained, quantitative assessment. Experiments on a range of proprietary and open-source MLLMs reveal a significant performance gap to humans, indicating that long-horizon egocentric reasoning remains a major challenge. We further analyze test-time scaling via stepwise reasoning and show that decomposing long action sequences can improve performance to some extent, while incurring non-trivial computational overhead. Overall, EXPLORE-Bench provides a principled testbed for measuring and advancing long-horizon reasoning for egocentric embodied perception.

Keywords: Egocentric Vision · Long-Horizon Reasoning · Benchmark

1 Introduction

Humans perceive and act in the physical world primarily from an egocentric (first-person) view. This perspective naturally couples vision with action: what we see affects what we will do, and what we do in turn changes what we will see. Motivated by this tight perception–action loop, egocentric benchmarks [7, 22, 26, 33, 40] have recently drawn increasing attention as a promising means to

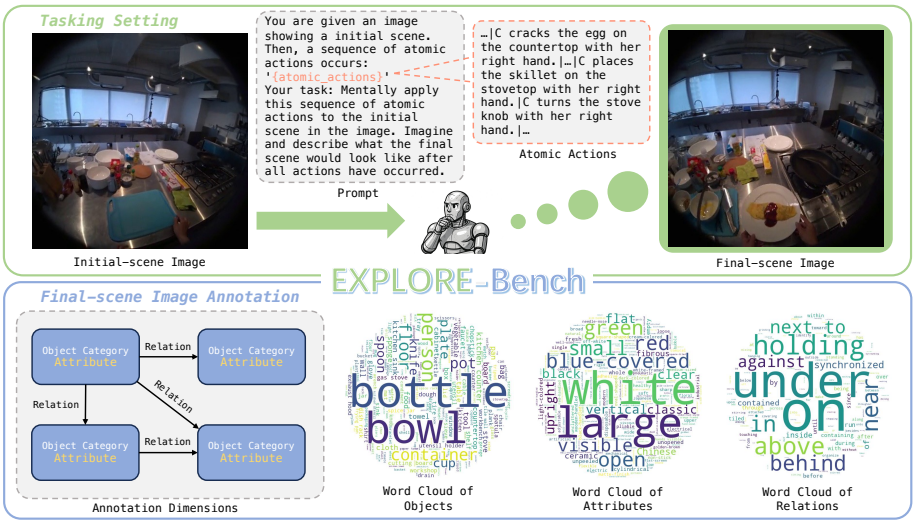


Fig. 1: Overview of EXPLORE-Bench. EXPLORE-Bench evaluates MLLMs on a new task: egocentric scene prediction with long-horizon reasoning. We annotate the final scene at the object, attribute, and relation levels to enable fine-grained scene-level evaluation. Note that the prompt is abbreviated for brevity in this figure.

evaluate the embodied potential of multimodal large language models (MLLMs) [18]. Existing egocentric benchmarks cover a broad set of capabilities. However, a fundamental ability remains under-explored: *egocentric scene prediction that requires long-horizon reasoning*—anticipating the final state of an entire scene after executing a long sequence of actions.

In this work, we formalize a new task: given an initial-scene image and a long sequence of atomic action descriptions, an MLLM is asked to predict the final scene after all actions are executed, as illustrated in Fig. 1. We refer to this as *egocentric scene prediction with long-horizon reasoning*. This capability is essential for MLLM-driven embodied agents to understand how their actions causally affect the physical world, which may support better decision making and planning. Importantly, it also helps avoid unintended negative consequences when pursuing a goal—for instance, removing a book from the bottom of a stack may destabilize the pile and cause other objects to fall. Such outcomes often depend on reasoning over atomic-action steps and maintaining a coherent representation of what changes and what remains invariant throughout the interaction.

While several egocentric benchmarks [25, 56] include evaluations related to state prediction, they typically focus on instant (or short-term) and localized state changes (*e.g.*, the state of a single object after a visual cue) and do not provide a systematic protocol to evaluate scene-level predictions under long action horizons (see Tab. 1). Moreover, without structured annotations of objects, attributes, and relations in the final scene, it is difficult to quantify what an MLLM gets right or wrong beyond coarse textual similarity. To bridge this gap,

Table 1: Comparison of widely adopted egocentric/embodied benchmarks with EXPLORE-Bench. Here, **completeness** refers to whether the atomic actions can describe a complete long-horizon task (*e.g.*, making an omelet).

Benchmarks	Atomic Actions			Scene Types		Evaluation Dimensions			
	length	influence	completeness	normal	abnormal	object	attribute	relation	scene
EgoTaskQA [22]	short	regional	✗	✓	✗	✓	✓	✓	✗
EgoMCQ [26]	short	regional	✗	✓	✗	✓	✓	✓	✗
EgoSchema [33]	short	regional	✗	✓	✗	✓	✗	✓	✗
EgoPlan-Bench [7]	short	regional	✗	✓	✗	✓	✗	✓	✗
ERQA [40]	short	regional	✗	✓	✗	✓	✗	✓	✗
EXPLORE-Bench	long	global	✓	✓	✓	✓	✓	✓	✓

we design a scalable annotation pipeline and build a new benchmark that enables fine-grained and quantitative evaluation of long-horizon egocentric scene prediction.

Specifically, we introduce **EXPLORE-Bench** (Sec. 3), a well-curated benchmark for evaluating MLLMs on egocentric scene prediction with long-horizon reasoning. Our benchmark contains 1,157 instances. Each instance consists of (i) an initial-scene image, (ii) a sequence of atomic actions with an average length of 113, and (iii) structured annotations for the final-scene image, including object categories, visual attributes, and inter-object relations. The dataset is derived from real first-person videos spanning diverse scenarios, and its structured annotations support systematic evaluation of models' ability to track and reason about long-term consequences of actions. We benchmark multiple proprietary and open-source MLLMs on EXPLORE-Bench and find a substantial gap to human performance, indicating that long-horizon egocentric scene prediction remains challenging for current models (Sec. 4). We further explore multi-step (decomposed) inference strategies and analyze when stepwise reasoning helps or hurts, providing insights for test-time scaling [61] in this setting (Sec. 5). Moreover, we specifically examine the models' performance in abnormal cases (Sec. 6). Our main contributions are summarized as follows:

- We propose a new task, *egocentric scene prediction with long-horizon reasoning*, to evaluate whether MLLM-driven agents can anticipate how their actions affect the physical world over long action horizons.
- We design a scalable data annotation pipeline and construct **EXPLORE-Bench** to support systematic, fine-grained evaluation of this task.
- We evaluate a range of proprietary and open-source MLLMs and show that, compared to humans, current models perform poorly on long-horizon egocentric scene prediction, especially in abnormal cases.
- We explore and analyze the effect of stepwise reasoning on model performance, offering practical insights for test-time scaling on egocentric scene prediction with long-horizon reasoning.

2 Related Work

Long-horizon Video Benchmarks. Recently, a growing number of benchmarks [12, 21, 33, 39, 45, 58, 65] for long-horizon video understanding and reasoning have emerged, providing standardized platforms for evaluating models' capabilities in temporal modeling, cross-frame relation extraction, and high-level semantic reasoning. These benchmarks emphasize sustained comprehension over long durations, requiring models to maintain long-term memory, track evolving events and object states, and integrate information across distant segments to answer queries about global narratives or long-range dependencies. Nevertheless, current long-horizon video benchmarks primarily test coarse event-level understanding, leaving the evaluation of multi-step, fine-grained state changes induced by sequences of atomic actions largely under-explored. Our EXPLORE-Bench instead assesses long-horizon egocentric scene prediction by requiring models to infer the final scene state after executing a sequence of atomic actions, with an emphasis on global scene changes involving multiple objects.

Egocentric Benchmarks. Contrary to works that focus on the exocentric view [31, 52, 59], egocentric benchmarks probe models from the first-person perspective. Existing benchmarks evaluate a broad spectrum of abilities, including recognition [36,53], memory [50,67], understanding [5,11,22,27], planning [51,63], generation [16, 54, 60], and reasoning [46, 47]. Several benchmarks [8, 9, 25, 56, 66] also emphasize broader embodied evaluation across tasks and settings. These efforts collectively demonstrate the promise of egocentric data for assessing embodied intelligence. Despite this progress, existing egocentric benchmarks do not include the task of egocentric scene prediction with long-horizon reasoning. Our task formulation differs from other egocentric reasoning benchmarks in that it requires reasoning over *action sequences* rather than videos, which better simulates an agent's reasoning after planning actions but before execution. In terms of data, we provide structured scene annotations with objects, attributes, and relations, enabling systematic evaluation of scene-level predictions.

3 EXPLORE-Bench

3.1 Overview

We introduce EXPLORE-Bench, a well-curated benchmark designed to quantitatively evaluate MLLMs' capability of egocentric scene prediction with long-horizon reasoning. EXPLORE-Bench comprises 1,157 instances derived from 1,157 real videos. These videos are sourced from two publicly available datasets: Ego4D [14] and Ego-Exo4D [15], as well as our self-recorded videos captured in diverse scenarios. The videos have an average duration of 358.28 seconds, with the longest lasting 1,524.80 seconds. Each instance in EXPLORE-Bench contains an image representing the initial scene, a sequence of atomic action descriptions, and annotations for the final-scene image. Across all instances, the number of atomic actions ranges from 11 to 694, with an average of 113 per instance. The

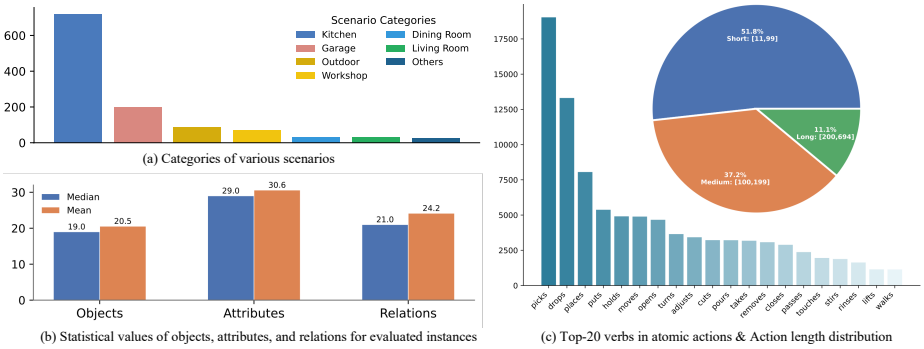


Fig. 2: Data analysis of EXPLORE-Bench reflects the rich diversity of scenarios, objects, attributes, relations, and atomic actions.

final-scene annotations include the object categories present in the scene, the visual attributes of these objects, and the relations among them. These annotations provide a solid foundation for systematically evaluating MLLMs' predictions of the final scene. We annotate 23,771 objects spanning 1,612 categories, with an average of more than 20 objects per instance. See Fig. 2 for more data analysis.

3.2 Dataset Construction

Data Collection. Ego4D [14] and Ego-Exo4D [15] cover a wide range of human activities. We first filter videos from these two datasets based on the recorded activities to select those suitable for long-horizon reasoning—for example, cooking and bicycle repair are considered valid, whereas dancing or chatting with someone are treated as invalid. For each selected video, we then extract the start and end frames of the main activity as the initial- and final-scene images of a instance, respectively. The timestamped atomic-action annotations in these datasets facilitate this process: using the timestamps of the extracted start and end frames, we retrieve the atomic action descriptions occurring in between. For our self-recorded videos, we follow the same format and annotate atomic actions (as shown in Fig. 7) at a granularity comparable to that of these two datasets.
Scene Annotation. We develop a pipeline to produce high-quality scene annotations at scale, as illustrated in Fig. 3.

Object Tagging. We first extract object tags from the final-scene images and the atomic action descriptions using the Recognize Anything Plus Model [20] and the spaCy [10] library, respectively, and merge them to maximize object coverage. Next, we remove invalid tags (*e.g.*, synonyms and human body parts) using a filtering scheme that combines an LLM-based filter with rule-based methods.

Object Grounding. The final-scene image and the filtered tags are then fed into Grounding DINO [28] to obtain bounding boxes for the objects. Each detected object instance is represented by a category label with an index (*e.g.*, `bowl.2`) to facilitate subsequent information matching and integration.

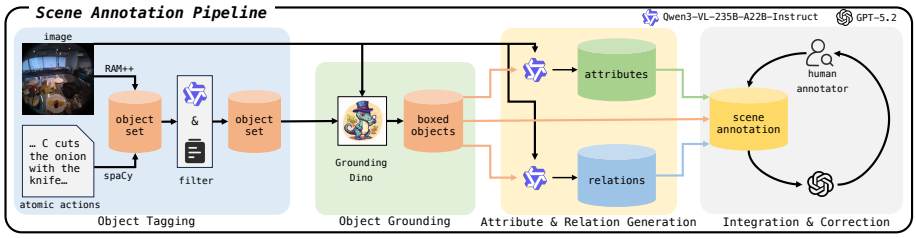


Fig. 3: Illustration of our scene annotation pipeline. Rather than having the MLLM generate annotations directly from the images, we adopt a multi-step pipeline to ensure object coverage and the accuracy of attributes and relations, greatly reducing the load of manual annotation required.

Attribute Generation. We then prompt Qwen3-VL-235B-A22B-Instruct [2] to generate descriptions of salient visual attributes for each object based on the final-scene image and the boxed objects, including shape, color, size, texture, and state, *etc.*

Relation Generation. Similarly, we prompt Qwen3-VL-235B-A22B-Instruct to produce inter-object relation triplets (`object.0`, `relation`, `object.1`), covering both spatial (*e.g.*, `under`) and interaction (*e.g.*, `holding`) relations.

Information Integration. We subsequently integrate objects in the final-scene image with their corresponding attributes and relations to form complete scene annotations. Within a single image annotation, different instances of the same category possess distinct attributes or relations.

Annotation Correction. Finally, we feed the image and the resulting annotations into GPT-5.2 [34] for annotation correction and augmentation.

Human-in-the-loop Quality Control. Human quality control is integrated throughout our dataset construction pipeline. During data collection, videos are first pre-filtered based on the scenario labels provided by the source datasets and then further screened by human reviewers. When extracting initial- and final-scene images from each video, we ensure that the frames are free of motion blur, that only the main human activity occurs between them, and that there is no drastic change in camera viewpoint across the two scenes. We also filter out ambiguous instances in which the action descriptions are insufficient to infer the final scene (*e.g.*, underspecified positions or severe occlusions). During scene annotation, our prompts and filtering rules are refined through multiple iterations with human verification. More importantly, the final scene annotations are corrected by human annotators. The annotation guidelines are updated multiple times based on annotator feedback.

3.3 Evaluation Protocol

Our task formulation requires models to predict and describe the final scene, resulting in outputs in the same format as image captioning. Inspired by CompreCap [29], we comprehensively evaluate MLLMs' predicted scene descriptions

from three aspects: object-level coverage, the accuracy of attribute descriptions, and the score of relations. The procedure is as follows:

Description Decomposition. The generated description is first split into sub-descriptions by sentence separators (*e.g.*, periods). Nouns are then extracted and lemmatized using spaCy [10] to form a set of candidate objects. This process establishes a standardized hierarchical structure, enabling accurate alignment with annotations and enhancing the stability of subsequent evaluations.

Object-level Evaluation. We employ Sentence-BERT [37] to extract word embeddings and compute a similarity matrix between the candidate objects and the annotated objects. Matched objects are identified by selecting the mutual maximum values in both rows and columns of the matrix. The object-level coverage S_{obj} is calculated from the refined matrix, quantifying the soft coverage of the annotated objects in the generated description.

Attribute-level Evaluation. In this step, we concatenate all sub-descriptions referring to the same annotated object. Then, a standardized prompt guides an LLM [48] to score attribute descriptions on a 0–5 scale based on the corresponding annotated phrases. The average score across all matched objects yields the attribute-level score S_{att}.

Relation-level Evaluation. The same scoring pipeline as in the attribute-level evaluation is applied to rate the alignment between generated and annotated relations. The resulting relation-level score S_{rel} quantitatively measures the quality of relation descriptions.

Metric Unification. S_{obj}, S_{att} and S_{rel} are linearly scaled to a 0–100 range. Then, a weighted average of the scaled scores is computed to obtain the unified score S_{uni}, providing a comprehensive quantification of the scene description:

$$S_{uni} = w_1 S_{obj} + w_2(20 \cdot S_{att}) + w_3(20 \cdot S_{rel}). \tag{1}$$

We assign weights following CompreCap [29] ($w_1 = 0.25, w_2 = 0.35, w_3 = 0.40$).

4 Evaluation on EXPLORE-Bench

4.1 Evaluation Setup

Benchmark Models. We comprehensively evaluate multimodal foundation models across diverse model families, including both proprietary MLLMs and open-source MLLMs. For proprietary MLLMs, we evaluate GPT-5.2 [34] and Gemini-3 [13]. For open-source MLLMs, we consider models from Qwen2-VL [41], Qwen2.5-VL [3], Qwen3-VL [2], InternVL3.5 [43], LLaVA-OneVision-1.5 [1], Keye-VL-1.5 [49], Ovis2.5 [30], MiMo-VL [57], MiniCPM-V-4.5 [55], GLM-4.6V [17], and Step3-VL [19], encompassing both non-thinking and thinking models. In addition, we evulate Embodied-Reasoner [62] and EgoThinker [35], which are tailored for embodied or egocentric reasoning. For models that support non-thinking and thinking modes, we report their performance in both modes. All models use the same prompt for final-scene prediction. We use greedy decoding to ensure reproducibility.

Dataset Partitioning. We partition all instances in EXPLORE-Bench into three subsets according to the length of the contained atomic-action sequence: `Short`, `Medium`, and `Long`. Specifically, the `Short` subset contains sequences of length 11–99, the `Medium` subset contains sequences of length 100–199, and the `Long` subset contains sequences ranging from 200 to 694. Note that the notion of "short" or "long" here is only relative. These three subsets contain 599, 430, and 128 instances, respectively (the distribution visualization is shown in Fig. 2). We report the performance of the models on each subset as well as on the full dataset (`Full`).

Human Performance Evaluation. We sample a subset of 100 instances from the `Short`, `Medium`, and `Long` subsets in a 52:37:11 ratio, which we refer to as EXPLORE-Bench (tiny). Human participants independently describe the final scene of each instance and their performance is evaluated using the same metrics as those used for the models. For comparison, we also report the performance of several top-tier MLLMs on EXPLORE-Bench (tiny).

Evaluation Metrics. We evaluate predicted scene descriptions using the protocol in Sec. 3.3 and report S_{obj}, S_{att}, S_{rel}, and S_{uni} on the `Short`, `Medium`, `Long`, and `Full` sets, respectively. Using an LLM as an evaluator is increasingly common [29,32,42,64], and we further verify the consistency between our evaluation protocol and human judgments. We sample 100 instances from EXPLORE-Bench and collect final-scene descriptions generated by multiple models. Human evaluators then score each description in a source-blind manner. The Spearman correlation ρ between scores produced by Qwen3-8B [48] (used as the LLM-based scorer) and human scores is 0.919. In comparison, pairwise correlations among human evaluators range from 0.912 to 0.936, indicating that the LLM scorer is largely consistent with human judgments.

4.2 Main Results

Based on the results reported in Tab. 2, we make the following key observations:

Human Performance. Human participants achieve an S_{uni} score of 59.08 on the EXPLORE-Bench (tiny) `Full` set, outperforming the best model by 7.39. Humans show a clear strength on the `Short` and `Medium` subsets. However, on the `Long` subset, the performance gap between humans and the best model narrows, suggesting that MLLMs may have a relative advantage on tasks requiring to process long action sequences, particularly in predicting inter-object relations. Although human participants outperform the top-tier MLLMs, the absolute score is not very high, highlighting the challenge posed by our benchmark.

Proprietary MLLMs. Despite a substantial gap to human performance, leading proprietary MLLMs, such as Gemini-3-Pro, deliver competitive results. In particular, the S_{uni} score of Gemini-3-Pro on the `Long` subset is close to that of humans. Gemini-3-Flash performs very similarly to Gemini-3-Pro overall, and slightly better than GPT-5.2-Chat.

Open-source MLLMs. Notably, Qwen3-VL-8B achieves higher overall performance than Gemini-3-Pro on our benchmark, although the latter is stronger on the `Long` subset. Qwen3-VL-8B-Instruct and Qwen3-VL-8B-Thinking perform

Table 2: Evaluation results on EXPLORE-Bench. `Short`, `Medium` and `Long` mean the subsets with short, medium and long atomic-action sequences, respectively. `Full` denotes the full dataset. Dark green and light green indicate the best and the second best result among all models. Orange denotes using Qwen2-VL-7B-Instruct as the base model. † denotes results on EXPLORE-Bench (tiny) set. # and * represent the non-thinking and thinking mode, respectively.

Methods	Short				Medium				Long				Full			
	S_{obj}	S_{att}	S_{rel}	S_{uni}	S_{obj}	S_{att}	S_{rel}	S_{uni}	S_{obj}	S_{att}	S_{rel}	S_{uni}	S_{obj}	S_{att}	S_{rel}	S_{uni}
Performance on EXPLORE-Bench (tiny)																
Human†	72.81	2.64	3.11	61.56	70.43	2.46	2.88	57.87	67.86	2.11	2.47	51.50	71.38	2.51	2.95	59.08
GPT-5.2-Chat†	59.93	1.73	2.79	49.42	60.39	1.72	2.66	48.40	58.25	1.65	2.61	46.97	59.92	1.72	2.72	48.77
Gemini-3-Pro†	59.76	1.76	2.74	49.15	60.92	1.78	2.76	49.80	60.40	1.69	2.63	47.97	60.26	1.76	2.74	49.26
Qwen3-VL-8B-Instruct†	61.71	1.91	2.89	51.95	62.03	1.90	2.84	51.55	55.54	1.69	2.56	46.17	61.15	1.88	2.84	51.16
Qwen3-VL-8B-Thinking†	65.65	1.96	2.96	53.83	61.40	1.83	2.75	50.19	57.99	1.67	2.55	46.59	63.24	1.88	2.83	51.69
Proprietary Multimodal Foundation Models (API)																
GPT-5.2-Chat	59.91	1.74	2.70	48.71	59.88	1.67	2.65	47.85	58.06	1.65	2.61	46.91	59.69	1.70	2.67	48.19
Gemini-3-Flash	60.31	1.84	2.78	50.27	59.33	1.76	2.71	48.84	58.27	1.72	2.69	48.18	59.72	1.80	2.75	49.51
Gemini-3-Pro	61.29	1.81	2.77	50.11	60.99	1.75	2.74	49.44	59.17	1.70	2.70	48.31	60.94	1.77	2.75	49.66
Open-source Non-thinking Multimodal Foundation Models																
Qwen2.5-VL-3B-Instruct	48.17	1.36	2.27	39.70	44.19	1.22	2.17	36.92	39.92	1.09	2.08	34.28	45.78	1.28	2.21	38.07
Qwen2-VL-7B-Instruct	47.64	1.34	2.23	39.14	46.04	1.27	2.16	37.73	43.79	1.22	2.15	36.73	46.62	1.30	2.20	38.35
Keye-VL-1.5-8B#	49.70	1.34	2.24	39.78	49.49	1.34	2.24	39.69	46.15	1.25	2.17	37.68	49.23	1.33	2.23	39.51
LLaVA-OneVision-1.5-4B	51.15	1.42	2.33	41.37	49.17	1.33	2.28	39.85	46.35	1.28	2.26	38.64	49.88	1.37	2.31	40.50
Qwen3-VL-2B-Instruct	52.18	1.48	2.55	43.77	49.03	1.37	2.45	41.38	43.27	1.18	2.28	37.34	50.02	1.40	2.48	42.17
Qwen2.5-VL-7B-Instruct	52.80	1.52	2.46	43.58	51.25	1.46	2.41	42.32	47.78	1.34	2.36	40.22	51.67	1.48	2.43	42.74
LLaVA-OneVision-1.5-8B	53.25	1.54	2.51	44.13	51.21	1.47	2.44	42.63	47.62	1.38	2.41	40.83	51.87	1.49	2.47	43.21
Ovis2.5-2B#	55.65	1.60	2.56	45.64	51.70	1.44	2.43	42.49	47.96	1.33	2.35	40.13	53.33	1.51	2.49	43.86
InternVL3.5-2B	57.51	1.64	2.62	46.80	55.35	1.53	2.50	44.60	52.40	1.42	2.41	42.32	56.14	1.58	2.55	45.48
InternVL3.5-8B	57.85	1.65	2.62	47.00	55.77	1.56	2.55	45.30	53.30	1.51	2.51	43.99	56.57	1.60	2.58	46.04
MiMo-VL-7B-RL-2508#	56.13	1.67	2.60	46.53	55.82	1.64	2.59	46.14	53.42	1.58	2.54	44.68	55.71	1.65	2.59	46.18
MiniCPM-V-4.5 (8B)#	58.86	1.70	2.63	47.67	57.45	1.64	2.58	46.49	54.69	1.56	2.55	44.95	57.87	1.66	2.60	46.93
Ovis2.5-9B#	59.51	1.74	2.72	48.85	57.84	1.65	2.63	47.01	53.83	1.55	2.50	44.32	58.26	1.68	2.66	47.66
Qwen3-VL-8B-Instruct	61.34	1.88	2.84	51.23	60.78	1.85	2.81	50.64	56.83	1.73	2.71	48.00	60.63	1.85	2.82	50.65
Open-source Thinking Multimodal Foundation Models																
Ovis2.5-2B*	51.10	1.40	2.43	42.02	44.52	1.17	2.20	36.89	40.70	1.04	2.10	34.25	47.50	1.28	2.31	39.25
Qwen3-VL-2B-Thinking	59.23	1.47	2.63	46.13	57.02	1.19	2.53	42.79	48.82	1.02	2.30	37.81	57.26	1.32	2.56	43.97
Ovis2.5-9B*	55.92	1.61	2.59	45.95	53.79	1.51	2.50	44.03	49.71	1.39	2.41	41.48	54.44	1.55	2.54	44.74
Keye-VL-1.5-8B*	56.58	1.58	2.55	45.64	56.54	1.54	2.53	45.14	53.13	1.62	2.43	44.09	56.18	1.57	2.53	45.28
MiMo-VL-7B-RL-2508*	57.32	1.64	2.56	46.26	56.85	1.59	2.53	45.63	56.58	1.57	2.50	45.17	57.06	1.61	2.54	45.90
MiniCPM-V-4.5 (8B)*	58.87	1.74	2.67	48.25	57.79	1.69	2.63	47.33	54.16	1.54	2.52	44.46	57.95	1.70	2.64	47.49
GLM-4.6V-Flash (9B)	60.70	1.78	2.71	49.32	58.44	1.70	2.65	47.67	54.95	1.60	2.60	45.75	59.22	1.73	2.67	48.31
Step3-VL-10B	61.32	1.76	2.78	49.87	60.06	1.62	2.71	48.01	59.11	1.60	2.64	47.09	60.61	1.69	2.74	48.87
Qwen3-VL-8B-Thinking	63.77	1.91	2.85	52.08	62.61	1.81	2.78	50.56	58.02	1.64	2.63	47.07	62.70	1.84	2.80	50.96
Embodied/Egocentric Multimodal Models																
Embodied-Reasoner (7B)	39.74	1.00	1.96	32.65	39.69	0.99	1.96	32.56	37.22	0.97	2.02	32.23	39.44	1.00	1.97	32.57
EgoThinker (7B)	48.22	1.36	2.23	39.39	45.88	1.28	2.19	37.89	42.42	1.16	2.08	35.36	46.71	1.31	2.20	38.38

comparably overall: the former scores higher on the `Medium` and `Long` subsets, while the latter performs better on the `Short` subset. However, most open-source models still lag far behind human performance, indicating substantial limitations in their egocentric scene prediction with long-horizon reasoning. Observations on thinking models are provided in Sec. 5.1.

Embodied & Egocentric MLLMs. Although Embodied-Reasoner and EgoThinker exhibit strong embodied or egocentric reasoning capabilities after training on specific datasets, they perform worse than most general-purpose MLLMs on our newly proposed task. Notably, compared to their base model (Qwen2-VL-7B-Instruct), they do not exhibit a clear performance improvement on our benchmark (Embodied-Reasoner even performs worse). This suggests that egocentric scene prediction with long-horizon reasoning has been largely overlooked by the community and remains in need of further exploration and development.
Description Length. We compute the average length (in words) of the scene descriptions generated by each model (excluding chain-of-thought for thinking models) and find no clear correlation between description length and performance. The top three models—Qwen3-VL-8B-Thinking, Qwen3-VL-8B-Instruct, and Gemini-3-Pro—produce descriptions with average lengths of 545.77, 312.46, and 186.60 words, respectively, indicating that Gemini-3-Pro is both accurate and concise. Qwen2.5-VL-3B-Instruct generates descriptions with an average length of 305.58 words, which is close to Qwen3-VL-8B-Instruct, yet their scores differ substantially. Embodied-Reasoner and EgoThinker produce long descriptions (812.00 and 808.92 words on average) but deliver unsatisfactory performance. In contrast, Step3-VL-10B also generates relatively long descriptions (747.27 words on average) while achieving highly competitive results. Additionally, Ovis2.5-9B in thinking mode produces the shortest descriptions, averaging 121.69 words.

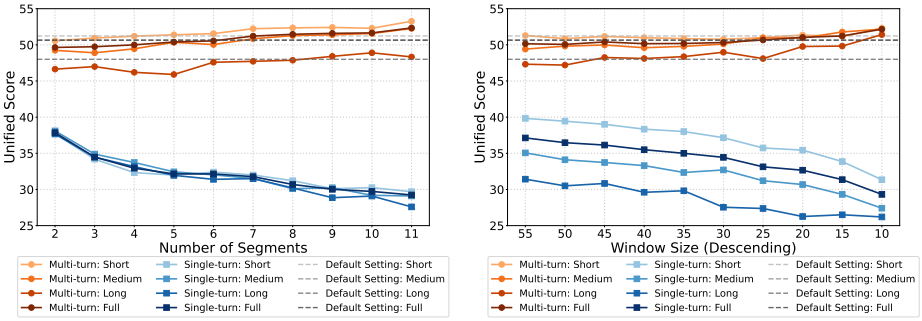
5 How Well Stepwise Reasoning Performs?

5.1 Thinking Model

As shown in Tab. 2, Keye-VL-1.5-8B [49] in thinking mode improves S_{uni} by 5.77 on the full EXPLORE-Bench compared to its non-thinking counterpart. In addition, native thinking models such as GLM-4.6V-Flash [17], Step3-VL-10B [19], and Qwen3-VL-8B-Thinking [2] deliver strong results, representing the leading performance among open-source models on our benchmark. However, for same-sized open-source MLLMs from the same model family, thinking-mode variants do not always outperform non-thinking ones, as observed for Ovis2.5-2B [30], Ovis2.5-9B, and MiMo-VL-7B-RL-2508 [57].

5.2 Non-thinking Model

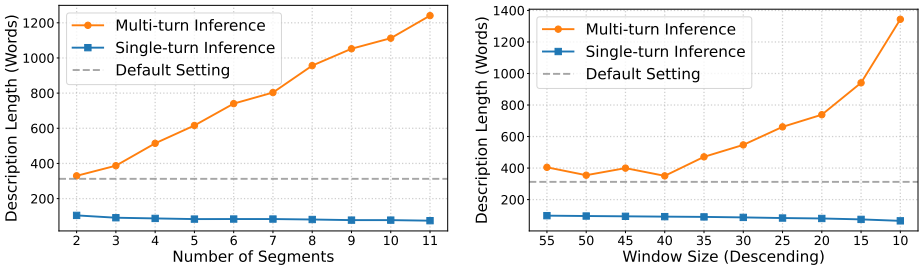
As an open-source family of foundation models with strong multimodal general-purpose capabilities, Qwen3-VL [2] has been widely adopted both in industry [6, 23, 38] and academia [24, 42, 44]. Within this family, the `Instruct` model (non-thinking) is often considered to offer greater development potential than the `Thinking` variant. Since non-thinking models cannot spontaneously reason with chain-of-thought as thinking models do, we elicit stepwise reasoning in Qwen3-VL-8B-Instruct through prompt design and explicit inference strategies.



(a) Unified score *vs.* number of segments. Note that the two dashed lines corresponding to "Default Setting: Medium" and "Default Setting: Full" overlap.

(b) Unified score *vs.* window size. Note that the two dashed lines corresponding to "Default Setting: Medium" and "Default Setting: Full" overlap.

Fig. 4: Unified score S_{uni} of Qwen3-VL-8B-Instruct across subsets under different inference strategies. `Short`, `Medium`, and `Long` denote the subsets with short, medium, and long atomic-action sequences. `Full` denotes the full dataset.



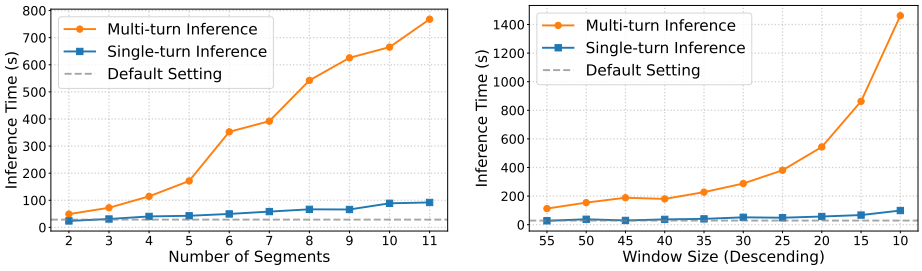
(a) Description length *vs.* number of segments

(b) Description length *vs.* window size

Fig. 5: Average single-instance final-scene description length of Qwen3-VL-8B-Instruct under different inference strategies.

Atomic Action Segments Partitioning. To enable stepwise reasoning, we first segment the atomic action sequence in two ways. The first approach evenly partitions the action sequence of each instance into a predefined number of segments (`segment_num`). This method is simple and straightforward, but does not account for variation in sequence lengths across instances. The second approach segments the sequence using a sliding window of size `window_size`. Specifically, every `window_size` atomic actions form one segment (the last segment may contain fewer actions), so longer sequences are divided into more segments. After segmenting the atomic action sequence, we apply two inference strategies to implement stepwise reasoning, which are introduced below. The experimental results of the above partitioning methods and inference strategies are provided in Fig. 4.

Single-turn Inference. Inspired by the dynamic programming [4] principle of decomposing the original problem into subproblems and performing state tran-



(a) Inference time *vs*. number of segments **(b)** Inference time *vs*. window size

Fig. 6: Average single-instance inference time of Qwen3-VL-8B-Instruct on an NVIDIA H200 GPU under different inference strategies.

sitions, we prompt Qwen3-VL-8B-Instruct to predict the scene after executing each action segment in a single-turn inference, while conditioning each prediction on the scene description after the previous segment (if available) and the initial-scene image.

We evaluate the resulting final-scene descriptions using the protocol in Sec. 3.3. However, this strategy does not yield strong gains. Instead, compared to the default setting that directly prompts the model to predict the final scene, it leads to a substantial performance drop, which becomes more severe when `segment_num` is larger or `window_size` is smaller (see Fig. 4). We find through case study that under this inference strategy, the model's generated scene descriptions mainly focus on the changes across scenes and tend to ignore the parts that remain unchanged. As a result, compared with the default setting, the final scene descriptions produced with this strategy are much shorter (see Fig. 5).

Multi-turn Inference. To encourage the model to generate detailed and complete scene descriptions after executing each action segment, we propose a new inference strategy. This strategy follows the same underlying idea as the single-turn strategy above, but implements it via multi-turn inference. The key difference from the above strategy is that, in each inference round, the model is exposed to only one action segment and is asked to output the scene description after executing that segment, analogous to the final-scene prediction. As the number of rounds increases, the generated description becomes significantly longer (see Fig. 5).

As shown in Fig. 4, in contrast to the single-turn strategy, the multi-turn inference yields better performance when `segment_num` is larger or `window_size` is smaller. This suggests that a finer decomposition into subtasks benefits scene prediction with long-horizon reasoning, aligning with intuition. We also find that a large `segment_num` (*e.g.*, 11) and a small `window_size` (*e.g.*, 10) achieve similar S_{uni} scores on the full dataset (52.34 *vs*. 52.16); however, the former performs better on the `Short` subset (53.26 *vs*. 52.31), whereas the latter is noticeably stronger on the `Long` subset (48.33 *vs*. 51.41). Additionally, multi-turn inference with `window_size=10` improves S_{uni} by 3.41 on the `Long` subset compared to the default setting. These observations suggest that adaptively segmenting action se-

Table 3: Evaluation results of abnormal cases. Dark green and light green indicate the best and the second best result among all models.

Methods	S_{obj}	S_{att}	S_{rel}	S_{uni}	S_{key}	S_{abn}
Human	95.19	3.64	3.88	80.32	4.65	91.64
Embodied-Reasoner	63.75	1.62	2.75	49.28	1.45	30.95
Qwen2-VL-7B-Instruct	68.84	2.13	2.98	55.98	2.20	45.20
EgoThinker	72.33	2.21	2.92	56.92	2.38	48.44
Qwen3-VL-8B-Instruct	83.83	2.82	3.51	68.75	2.54	52.63
Gemini-3-Pro	80.98	2.70	3.41	66.47	2.74	56.00
Qwen3-VL-8B-Thinking	86.03	2.89	3.62	70.71	2.84	58.22
GPT-5.2-Chat	87.32	2.82	3.55	69.94	3.10	62.79

quences based on their length may offer advantages for long-horizon reasoning tasks. However, although larger `segment_num` and smaller `window_size` can improve performance over the default setting, model performance drops below the default when `segment_num` is less than 7 or `window_size` exceeds 25. Moreover, multi-turn inference incurs multiplicative computational overhead (see Fig. 6) while providing relatively limited gains.

6 Can MLLMs Describe Abnormal States?

We record and annotate 20 videos in which executing a sequence of actions leads to abnormal states in the final scene. The anomalies in these videos fall into two main categories: environmental damage (*e.g.*, objects falling) and safety hazards (*e.g.*, a faucet left running). In addition to annotating atomic actions and the objects, attributes, and relations in the final scene, we also annotate the key object states that reflect the scene anomalies (as shown in Fig. 7), enabling us to investigate whether MLLMs can infer the anomalies and describe these key states. We evaluate descriptions of key states in the same way as we evaluate object attributes, yielding a key-state score S_{key} (0-5) . We then compute the score for abnormal-scene descriptions as

$$S_{abn} = (1 - \lambda)S_{uni} + \lambda(20 \cdot S_{key}). \tag{2}$$

We report results of these cases with $\lambda = 0.9$ in Tab. 3.

Compared with the full EXPLORE-Bench, these cases have much shorter action sequences (17 on average) and involve fewer objects (9 on average) per scene, leading to relatively higher S_{uni} scores for the models. However, their S_{abn} scores still lag far behind human performance. Benefiting from everyday experience, humans can easily detect abnormal states and achieve an S_{key} score of 4.65. Relative to the base model (Qwen2-VL-7B-Instruct [41]), Embodied-Reasoner [62] performs worse overall, while EgoThinker [35] shows only marginal improvements. Qwen3-VL-8B-Thinking [2] and GPT-5.2-Chat [34] perform well among all models, yet our case study (shown in Fig. 7) demonstrates that they still fail to accurately describe abnormal states that humans can readily perceive.

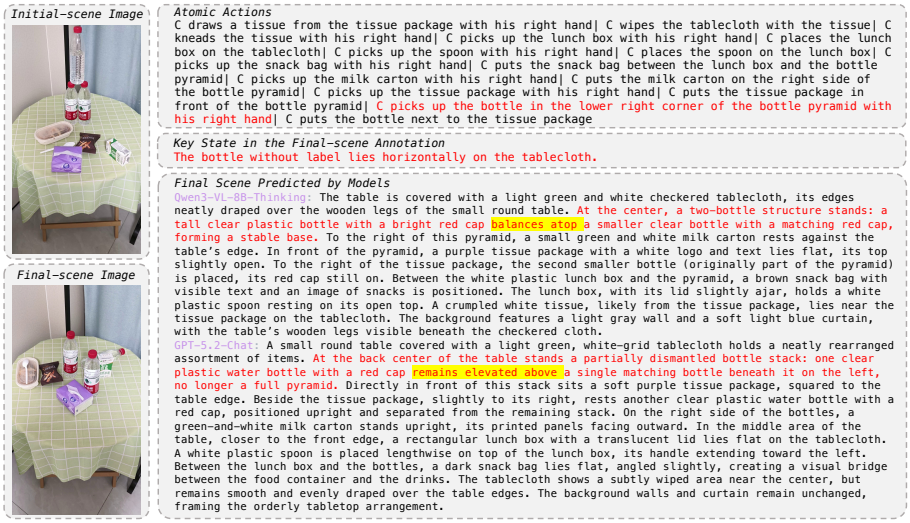


Fig. 7: Case study of an abnormal situation. "C" refers to the camera wearer. Although both models recognize the changes, neither accurately describes the key state. Instead, they generate descriptions that violate physical commonsense.

7 Conclusion

In this paper, we introduce **EXPLORE-Bench**, a benchmark for *egocentric scene prediction with long-horizon reasoning*, where models are asked to anticipate the final state of a scene given an initial image and a sequence of atomic actions. EXPLORE-Bench is built from real first-person videos across diverse scenarios and provides structured final-scene annotations—object categories, attributes, and relations—enabling fine-grained, quantitative evaluation of holistic scene-level prediction. Through extensive evaluation of both proprietary and open-source MLLMs, we find that current models still struggle to reliably track long-term action consequences and remain substantially behind human performance, especially in abnormal cases. We further investigated stepwise reasoning at test time, showing that decomposing long action sequences can improve performance in challenging long-horizon settings, while also revealing non-trivial trade-offs between accuracy gains and computational overhead.

Limitations and Future Work. Currently, our exploration of test-time scaling is limited to a small set of inference strategies. Developing more effective *and* more efficient test-time scaling methods for long-horizon reasoning remains open. While EXPLORE-Bench covers diverse everyday activities, rare or abnormal cases (*e.g.*, unexpected interactions) are still under-represented. Collecting and annotating more such instances is an important next step. Moreover, as more data become available, it would be valuable to construct a dedicated training set to further advance models' long-horizon egocentric reasoning.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (NSFC) under Grants 62225207, 62436008, 62306295, 62576328 and 625B2175. We sincerely thank the Foundation Model Team at Li Auto Inc. for providing computational resources and data support.

References

1. An, X., Xie, Y., Yang, K., Zhang, W., Zhao, X., Cheng, Z., Wang, Y., Xu, S., Chen, C., Zhu, D., et al.: Llava-onevision-1.5: Fully open framework for democratized multimodal training. arXiv preprint arXiv:2509.23661 (2025)
2. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
3. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
4. Bellman, R.: Dynamic programming. science **153**(3731), 34–37 (1966)
5. Chen, Q., Di, S., Xie, W.: Grounded multi-hop videoqa in long-form egocentric videos. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 2159–2167 (2025)
6. Chen, W., Du, C., Gu, F., He, W., Li, Q., Liu, Z., Pan, X., Ren, C., Rao, X., Wang, C., et al.: Mindgpt-4ov: An enhanced mllm via a multi-stage post-training paradigm. arXiv preprint arXiv:2512.02895 (2025)
7. Chen, Y., Ge, Y., Ge, Y., Ding, M., Li, B., Wang, R., Xu, R., Shan, Y., Liu, X.: Egoplan-bench: Benchmarking egocentric embodied planning with multimodal large language models. arXiv preprint arXiv:2312.06722 **3**(7) (2023)
8. Cheng, S., Guo, Z., Wu, J., Fang, K., Li, P., Liu, H., Liu, Y.: Egothink: Evaluating first-person perspective thinking capability of vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14291–14302 (2024)
9. Dang, R., Yuan, Y., Zhang, W., Xin, Y., Zhang, B., Li, L., Wang, L., Zeng, Q., Li, X., Bing, L.: Ecbench: Can multi-modal foundation models understand the egocentric world? a holistic embodied cognition benchmark. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 24593–24602 (2025)
10. Explosion: Spacy. `https://github.com/explosion/spaCy` (2015), accessed: 2026-02-19
11. Fan, C.: Egovqa-an egocentric video question answering benchmark dataset. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. pp. 0–0 (2019)
12. Fu, C., Dai, Y., Luo, Y., Li, L., Ren, S., Zhang, R., Wang, Z., Zhou, C., Shen, Y., Zhang, M., et al.: Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 24108–24118 (2025)
13. Google DeepMind: Gemini 3. `https://deepmind.google/models/gemini/` (2025), accessed: 2026-02-16

14. Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., 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. pp. 18995–19012 (2022)
15. Grauman, K., Westbury, A., Torresani, L., Kitani, K., Malik, J., Afouras, T., Ashutosh, K., Baiyya, V., Bansal, S., Boote, B., et al.: Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19383–19400 (2024)
16. Han, G., Zhai, W., Yang, Y., Cao, Y., Zha, Z.J.: Touch: Text-guided controllable generation of free-form hand-object interactions. arXiv preprint arXiv:2510.14874 (2025)
17. Hong, W., Yu, W., Gu, X., Wang, G., Gan, G., Tang, H., Cheng, J., Qi, J., Ji, J., Pan, L., et al.: Glm-4.5 v and glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. arXiv preprint arXiv:2507.01006 (2025)
18. Hou, L., Gao, L., Wu, Y., Chang, Y.: A survey on evaluation of embodied ai. Authorea Preprints (2026)
19. Huang, A., Yao, C., Han, C., Wan, F., Guo, H., Lv, H., Zhou, H., Wang, J., Zhou, J., Sun, J., et al.: Step3-vl-10b technical report. arXiv preprint arXiv:2601.09668 (2026)
20. Huang, X., Huang, Y.J., Zhang, Y., Tian, W., Feng, R., Zhang, Y., Xie, Y., Li, Y., Zhang, L.: Open-set image tagging with multi-grained text supervision. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 4117–4126 (2025)
21. Huang, Z., Li, X., Li, J., Wang, J., Zeng, X., Liang, C., Wu, T., Chen, X., Li, L., Wang, L.: Online video understanding: Ovbench and videochat-online. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 3328–3338 (2025)
22. Jia, B., Lei, T., Zhu, S.C., Huang, S.: Egotaskqa: Understanding human tasks in egocentric videos. Advances in Neural Information Processing Systems **35**, 3343–3360 (2022)
23. Li, M., Zhang, Y., Long, D., Chen, K., Song, S., Bai, S., Yang, Z., Xie, P., Yang, A., Liu, D., et al.: Qwen3-vl-embedding and qwen3-vl-reranker: A unified framework for state-of-the-art multimodal retrieval and ranking. arXiv preprint arXiv:2601.04720 (2026)
24. Li, S., Kallidromitis, K., Gokul, A., Kato, Y., Kozuka, K., Grover, A.: Mobileworldbench: Towards semantic world modeling for mobile agents. arXiv preprint arXiv:2512.14014 (2025)
25. Li, Y., Fu, Y., Qian, T., Xu, Q., Dai, S., Paudel, D.P., Van Gool, L., Wang, X.: Egocross: Benchmarking multimodal large language models for cross-domain egocentric video question answering. arXiv preprint arXiv:2508.10729 (2025)
26. Lin, K.Q., Wang, J., Soldan, M., Wray, M., Yan, R., Xu, E.Z., Gao, D., Tu, R.C., Zhao, W., Kong, W., et al.: Egocentric video-language pretraining. Advances in Neural Information Processing Systems **35**, 7575–7586 (2022)
27. Liu, S., Yu, G., Luo, X., Zheng, S., Chen, W., Liu, J., Shen, L.: Benchmarking egocentric clinical intent understanding capability for medical multimodal large language models. arXiv preprint arXiv:2601.06750 (2026)
28. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., et al.: Grounding dino: Marrying dino with grounded pre-training for

open-set object detection. In: European conference on computer vision. pp. 38–55. Springer (2024)
29. Lu, F., Wu, W., Zheng, K., Ma, S., Gong, B., Liu, J., Zhai, W., Cao, Y., Shen, Y., Zha, Z.J.: Benchmarking large vision-language models via directed scene graph for comprehensive image captioning. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 19618–19627 (2025)
30. Lu, S., Li, Y., Xia, Y., Hu, Y., Zhao, S., Ma, Y., Wei, Z., Li, Y., Duan, L., Zhao, J., et al.: Ovis2.5 technical report. arXiv preprint arXiv:2508.11737 (2025)
31. Luo, H., Zhai, W., Zhang, J., Cao, Y., Tao, D.: Learning affordance grounding from exocentric images. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2252–2261 (2022)
32. Majumdar, A., Ajay, A., Zhang, X., Putta, P., Yenamandra, S., Henaff, M., Silwal, S., Mcvay, P., Maksymets, O., Arnaud, S., et al.: Openeqa: Embodied question answering in the era of foundation models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16488–16498 (2024)
33. Mangalam, K., Akshulakov, R., Malik, J.: Egoschema: A diagnostic benchmark for very long-form video language understanding. Advances in Neural Information Processing Systems **36**, 46212–46244 (2023)
34. OpenAI: Introducing gpt-5.2. `https://openai.com/index/introducing-gpt-5-2/` (2025), accessed: 2026-02-16
35. Pei, B., Huang, Y., Xu, J., He, Y., Chen, G., Wu, F., Qiao, Y., Pang, J.: Egothinker: Unveiling egocentric reasoning with spatio-temporal cot. arXiv preprint arXiv:2510.23569 (2025)
36. Perrett, T., Darkhalil, A., Sinha, S., Emara, O., Pollard, S., Parida, K.K., Liu, K., Gatti, P., Bansal, S., Flanagan, K., et al.: Hd-epic: A highly-detailed egocentric video dataset. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 23901–23913 (2025)
37. Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bert-networks. In: Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP). pp. 3982–3992 (2019)
38. Shen, X., Chen, J., Zheng, L., Ma, H., Wei, T., Zhan, K.: Evolving from tool user to creator via training-free experience reuse in multimodal reasoning. arXiv preprint arXiv:2602.01983 (2026)
39. Song, E., Chai, W., Wang, G., Zhang, Y., Zhou, H., Wu, F., Chi, H., Guo, X., Ye, T., Zhang, Y., et al.: Moviechat: From dense token to sparse memory for long video understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18221–18232 (2024)
40. Team, G.R., Abeyruwan, S., Ainslie, J., Alayrac, J.B., Arenas, M.G., Armstrong, T., Balakrishna, A., Baruch, R., Bauza, M., Blokzijl, M., et al.: Gemini robotics: Bringing ai into the physical world. arXiv preprint arXiv:2503.20020 (2025)
41. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024)
42. Wang, S., Cui, C., Huang, Y., Chang, H.J., Cheng, Y.: Vl4gaze: Unleashing vision-language models for gaze following. arXiv preprint arXiv:2512.20735 (2025)
43. Wang, W., Gao, Z., Gu, L., Pu, H., Cui, L., Wei, X., Liu, Z., Jing, L., Ye, S., Shao, J., et al.: Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. arXiv preprint arXiv:2508.18265 (2025)

44. Wang, Y., Ke, L., Zhang, B., Qu, T., Yu, H., Huang, Z., Yu, M., Xu, D., Yu, D.: N3d-vlm: Native 3d grounding enables accurate spatial reasoning in vision-language models. arXiv preprint arXiv:2512.16561 (2025)
45. Wu, H., Li, D., Chen, B., Li, J.: Longvideobench: A benchmark for long-context interleaved video-language understanding. Advances in Neural Information Processing Systems **37**, 28828–28857 (2024)
46. Wu, P., Liu, Y., Liu, M., Shen, J.: St-think: How multimodal large language models reason about 4d worlds from ego-centric videos. arXiv preprint arXiv:2503.12542 (2025)
47. Yan, J., Ren, R., Liu, J., Xu, S., Wang, L., Wang, Y., Zhong, X., Wang, Y., Zhang, L., Chen, X., et al.: Teleego: Benchmarking egocentric ai assistants in the wild. arXiv preprint arXiv:2510.23981 (2025)
48. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025)
49. Yang, B., Wen, B., Ding, B., Liu, C., Chu, C., Song, C., Rao, C., Yi, C., Li, D., Zang, D., et al.: Kwai keye-vl 1.5 technical report. arXiv preprint arXiv:2509.01563 (2025)
50. Yang, J., Liu, S., Guo, H., Dong, Y., Zhang, X., Zhang, S., Wang, P., Zhou, Z., Xie, B., Wang, Z., et al.: Egolife: Towards egocentric life assistant. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 28885–28900 (2025)
51. Yang, R., Chen, H., Zhang, J., Zhao, M., Qian, C., Wang, K., Wang, Q., Koripella, T.V., Movahedi, M., Li, M., et al.: Embodiedbench: Comprehensive benchmarking multi-modal large language models for vision-driven embodied agents. In: International Conference on Machine Learning. pp. 70576–70631. PMLR (2025)
52. Yang, Y., Zhai, W., Luo, H., Cao, Y., Luo, J., Zha, Z.J.: Grounding 3d object affordance from 2d interactions in images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10905–10915 (2023)
53. Yang, Y., Zhai, W., Wang, C., Yu, C., Cao, Y., Zha, Z.J.: Egochoir: Capturing 3d human-object interaction regions from egocentric views. Advances in Neural Information Processing Systems **37**, 54529–54557 (2024)
54. Yu, C., Zhai, W., Yang, Y., Cao, Y., Zha, Z.J.: Hero: Human reaction generation from videos. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10262–10274 (2025)
55. Yu, T., Wang, Z., Wang, C., Huang, F., Ma, W., He, Z., Cai, T., Chen, W., Huang, Y., Zhao, Y., et al.: Minicpm-v 4.5: Cooking efficient mllms via architecture, data, and training recipe. arXiv preprint arXiv:2509.18154 (2025)
56. Yuan, Y., Dang, R., Li, L., Li, W., Jiao, D., Li, X., Zhao, D., Wang, F., Zhang, W., Xiao, J., et al.: Eoc-bench: Can mllms identify, recall, and forecast objects in an egocentric world? In: The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2025)
57. Yue, Z., Lin, Z., Song, Y., Wang, W., Ren, S., Gu, S., Li, S., Li, P., Zhao, L., Li, L., et al.: Mimo-vl technical report. arXiv preprint arXiv:2506.03569 (2025)
58. Yue, Z., Zhang, Q., Hu, A., Zhang, L., Wang, Z., Jin, Q.: Movie101: A new movie understanding benchmark. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 4669–4684 (2023)
59. Zhai, W., Cao, Y., Zhang, J., Xie, H., Tao, D., Zha, Z.J.: On exploring multiplicity of primitives and attributes for texture recognition in the wild. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(1), 403–420 (2023)

60. Zhang, L., Li, Z., Li, T., Cao, Z., Xu, R., Long, X., Wang, W., Wang, J., Liu, Y., Wang, W., et al.: Egoreact: Egocentric video-driven 3d human reaction generation. arXiv preprint arXiv:2512.22808 (2025)
61. Zhang, Q., Lyu, F., Sun, Z., Wang, L., Zhang, W., Hua, W., Wu, H., Guo, Z., Wang, Y., Muennighoff, N., et al.: A survey on test-time scaling in large language models: What, how, where, and how well? arXiv preprint arXiv:2503.24235 (2025)
62. Zhang, W., Wang, M., Liu, G., Huixin, X., Jiang, Y., Shen, Y., Hou, G., Zheng, Z., Zhang, H., Li, X., et al.: Embodied-reasoner: Synergizing visual search, reasoning, and action for embodied interactive tasks. arXiv preprint arXiv:2503.21696 (2025)
63. Zhao, B., Fang, J., Dai, Z., Wang, Z., Zha, J., Zhang, W., Gao, C., Wang, Y., Cui, J., Chen, X., et al.: Urbanvideo-bench: Benchmarking vision-language models on embodied intelligence with video data in urban spaces. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 32400–32423 (2025)
64. Zheng, L., Chiang, W.L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., et al.: Judging llm-as-a-judge with mt-bench and chatbot arena. Advances in neural information processing systems **36**, 46595–46623 (2023)
65. Zhou, J., Shu, Y., Zhao, B., Wu, B., Liang, Z., Xiao, S., Qin, M., Yang, X., Xiong, Y., Zhang, B., et al.: Mlvu: Benchmarking multi-task long video understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13691–13701 (2025)
66. Zhou, S., Xiao, J., Li, Q., Li, Y., Yang, X., Guo, D., Wang, M., Chua, T.S., Yao, A.: Egotextvqa: Towards egocentric scene-text aware video question answering. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 3363–3373 (2025)
67. Zhou, W., Cao, K., Zheng, H., Liu, Y., Zheng, X., Liu, M., Kristensson, P.O., Mayol-Cuevas, W., Zhang, F., Lin, W., et al.: X-lebench: A benchmark for extremely long egocentric video understanding. arXiv preprint arXiv:2501.06835 (2025)