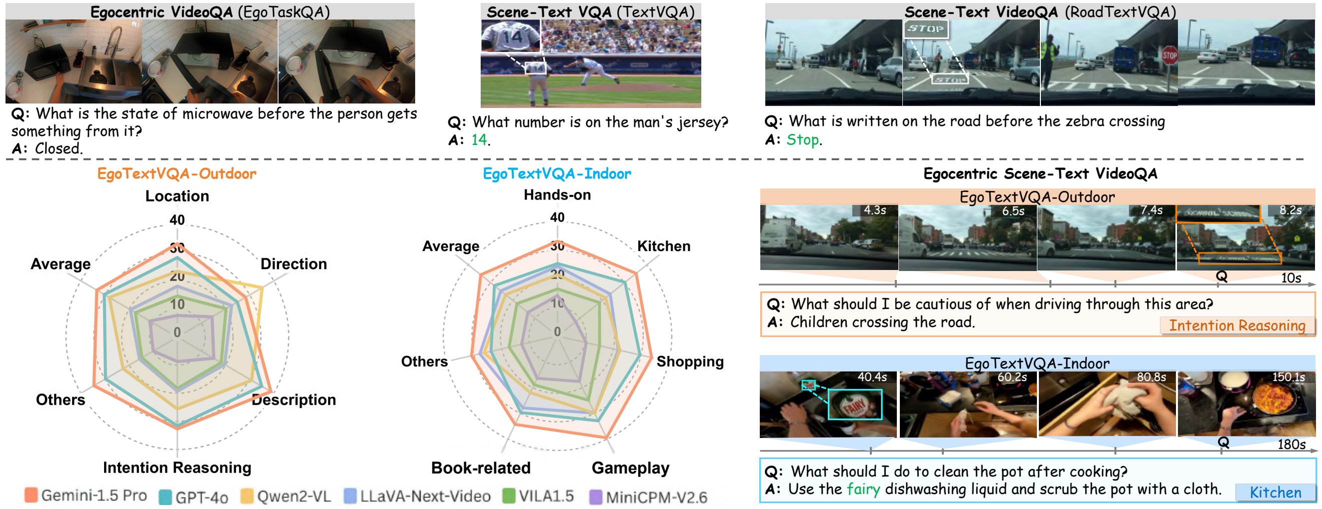
EgoTextVQA: Towards Egocentric Scene-Text Aware Video Question Answering

Sheng Zhou[1, 2*] Junbin Xiao[2†] Qingyun Li[2] Yicong Li[2] Xun Yang[3] Dan Guo[1]
Meng Wang[1] Tat-Seng Chua[2] Angela Yao[2]
[1]Hefei University of Technology [2]National University of Singapore [3]University of Science and Technology of China
hzgn97@gmail.com, {junbin, ayao}@comp.nus.edu.sg



Q: What is the state of microwave before the person gets something from it?
A: Closed.

Q: What number is on the man's jersey?
A: 14.

Q: What is written on the road before the zebra crossing
A: Stop.

Q: What should I be cautious of when driving through this area?
A: Children crossing the road.

Q: What should I do to clean the pot after cooking?
A: Use the fairy dishwashing liquid and scrub the pot with a cloth.

Figure 1. Our EgoTextVQA aims for QA assistance involving scene text from an ego-perspective mainly in outdoor driving (EgoTextVQA-Outdoor) and indoor house-keeping (EgoTextVQA-Indoor), with the questions reflecting the real user needs yet without the visual focus on scene text. Benchmarking results show that all models struggle on EgoTextVQA, highlighting continued efforts for improvements.

Abstract

We introduce EgoTextVQA, a novel and rigorously constructed benchmark for egocentric QA assistance involving scene text. EgoTextVQA contains 1.5K ego-view videos and 7K scene-text aware questions that reflect real user needs in outdoor driving and indoor house-keeping activities. The questions are designed to elicit identification and reasoning on scene text in an egocentric and dynamic environment. With EgoTextVQA, we comprehensively evaluate 10 prominent multimodal large language models. Currently, all models struggle, and the best results (Gemini 1.5 Pro) are around 33% accuracy, highlighting the severe deficiency of these techniques in egocentric QA assistance. Our further investigations suggest that precise temporal grounding and multi-frame reasoning, along with high resolution and auxiliary scene-text inputs, are key for better performance. With thorough analyses and heuristic suggestions, we hope EgoTextVQA can serve as a solid testbed for research in egocentric scene-text QA assistance. Our dataset is released at: https://github.com/zhousheng97/EgoTextVQA.

1. Introduction

The advances in wearable cameras and egocentric vision research [12, 14, 28, 39] have led to a surge of interest in developing assistance applications. A promising direction is egocentric video question answering (VideoQA) [11, 15, 26, 55, 58, 62], where AI agents offer live help by answering user questions. Despite the potential, related dataset research [9, 12, 20, 33, 60] has been designed to assess VideoQA models in their visual understanding capabilities. The questions are not curated to reflect real user needs in assistive applications (see EgoTaskQA in Figure 1). QAEgo4D [2] and AssistQ [53] are closer in aim, but these datasets are designed to assist with episodic memory and "how to" demonstration; they emphasize pure visual

---

[*] Work done when the first author was a visiting student at NUS.
[†] Corresponding author.

understanding while neglecting scene text.

Research in assistive techniques [16, 25, 45, 68] have shown that scene text, including signs, labels, and other text elements that are ubiquitous in our daily lives, is frequently involved in user questions when seeking assistance. Existing efforts [3, 45, 69] for QA on scene-text primarily aim at Optical Character Recognition (OCR) from ideal imaging conditions. They assume that people take good pictures and ask questions directly pinpointing the scene-text regions (see TextVQA [45] in Figure 1). Despite the progress (*e.g.*, leading results have accuracies of 85% on ST-VQA [3, 50]), such a simplified setting has limited application value. Consider, for example, those with a visual impairment [16], they will likely struggle to take clean and in-focus images; or can they point to scene-text regions. More recent works [48, 67] begin to study text-based VideoQA which allows QA on multiple images to reduce the proportion of unanswerable cases. Nonetheless, they still assume people know well the locations of scene text (see RoadTextVQA [48] in Figure 1). Corresponding VQA models, while achieving good results on such benchmarks, cannot reason over user intentions other than an extraction of the OCR results [19].

In light of this, we introduce EgoTextVQA, a novel and rigorously constructed Egocentric Scene-Text Aware Video Question Answering benchmark. EgoTextVQA is designed to advance research on egocentric QA assistance in real-life scenarios. We collect 1.5K ego-view videos and 7K scene-text aware questions from outdoor driving (EgoTextVQA-Outdoor) and indoor house-keeping (EgoTextVQA-Indoor) activities. By emphasizing real user needs, we allow questions and answers that do not explicitly pinpoint the scene text. However, comprehending scene text is crucial to answering the questions (see examples in Figure 1). Moreover, to simulate real-time and streaming video QA, we set a timestamp for each question and allow the models to access *only* visual contents before the question's timestamp to make related responses. Detailed visual scenarios and question types are also provided for better analysis of models.

For its realistic setting, EgoTextVQA retains all the challenges in existing scene text VQA datasets and incorporates new major difficulties: First and foremost, it is crucial to *reason about the user's intentions* with respect to the visual scene and scene text to both understand and to answer the questions. Second, the model may need to *reason across multiple frames* to either infer the user's behavior or to locate the required scene text for answering. Third, the model may need to *infer the user's current state* when posing the questions and provide actionable answers for meaningful assistance. Finally, the egocentric dynamic situation poses an additional challenge to *scene text recognition* as opposed to the well-captured images and focused scene text.

With EgoTextVQA, we comprehensively benchmark 10 prominent models that perform well on existing scene-text VQA datasets, and find they all struggle, especially the open-source ones like MiniCPM-V 2.6 [59], ShareGPT4Video [6] and CogVLM2-Video [18]. The best-performing model is Gemini 1.5 Pro [41], yet its accuracy is still lower at around 33%. Comprehensive analyses and additional heuristic explorations reveal many insights on model behaviors and possible directions for improvements:

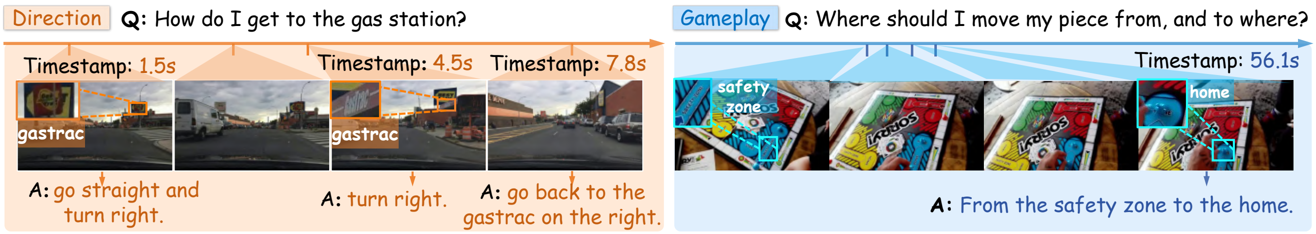
1. Temporally localizing the key frames and jointly reasoning over multiple key frames (*vs.* a single key frame) are crucial for improvements, especially for long video QA scenarios on EgoTextVQA-Indoor.
2. High-resolution image and scene text input can significantly boost models' performances. Yet, one needs to mind the compute efficiency. Also, models that inherently take fixed higher-resolution inputs are not necessarily better than those using lower-resolution inputs.
3. Scene text from additional OCR techniques, despite being auxiliary, are extremely helpful for all models.

Additionally, we find that even humans cannot perform well on this task (43% on EgoTextVQA-Outdoor and 27% on EgoTextVQA-Indoor), largely due to the difficulty of scene text recognition and the limited knowledge of humans, highlighting the significance of this research. To summarize our contributions: 1) We propose to study scene-text aware VideoQA towards egocentric assistance and construct the EgoTextVQA dataset containing both EgoTextVQA-Indoor and EgoTextVQA-Outdoor. 2) We benchmark 10 contemporary powerful models covering both open-source and closed-source ones and comprehensively analyze their limitations. 3) We explore a series of heuristic methods for improvements and share insightful findings for future work.

2. Related Work

2.1. Scene-Text VQA Benchmark

In the scene-text VQA field, various image- and video-based datasets [3, 30, 36, 45, 48, 52, 67] have been proposed. The datasets like TextVQA [45], ST-VQA [3], and ESTVQA [52] provide high-quality images and questions that clearly point to scene text, but such simple settings limit practical applications. To expand the application scope, the datasets like M4viteVQA [67] and RoadTextVQA [48] offer text-rich videos, but they design simple questions which refer to well-focused scene text. Such simplified settings make QA less practical and they challenge no more than an identification of the scene text. In addition, RoadTextVQA focuses merely on road driving. In this work, we introduce EgoTextVQA to advance research in real-life QA assistance from an egocentric perspective. Our videos cover diverse daily scenarios, with the questions reflecting real user needs yet without the visual attention on scene text.


Direction
Q: How do I get to the gas station?
Timestamp: 1.5s
Timestamp: 4.5s
Timestamp: 7.8s
gastrac
gastrac
A: go straight and turn right.
A: turn right.
A: go back to the gastrac on the right.
Gameplay
Q: Where should I move my piece from, and to where?
Timestamp: 56.1s
safety zone
home
A: From the safety zone to the home.

(a) EgoTextVQA-Outdoor

(b) EgoTextVQA-Indoor

Figure 2. Examples of EgoTextVQA. Scene text plays pivotal role in understanding and answering the questions which reflect real user needs. Yet, the videos are without the visual focus on scene text.

2.2. MLLMs for Scene-Text VQA

Recent advancements in Multimodal Large Language Models (MLLMs) [51, 63, 66] have shown significant potential in addressing scene-text VQA tasks [3, 30, 45]. Current MLLMs [23, 31, 66] achieve substantial performance gains in image scene-text QA by enhancing their ability to process higher-resolution images, leading to significant improvements and strong performance. However, these models are limited to image-level input, leaving their video comprehension capabilities largely unexplored. To address this limitation, recent models [6, 7, 18, 27, 50, 56, 57, 59, 64] are developed to support video-level or multi-frame inputs while retain their performance on image-based scene-text QA datasets. Nonetheless, these models have been evaluated separately on image-based OCR-rich benchmarks [3, 30, 34–36, 45] and video understanding benchmarks [13, 24, 33, 43, 54], lacking suitable assessment for scene-text comprehension in more realistic and dynamic environments. In this work, we will comprehensively analyze the model behaviors in egocentric QA assistance in real-life dynamic environments.

3. EgoTextVQA Dataset

3.1. Dataset Creation

We leverage existing text-rich video sequences and the multimodal understanding and generation capabilities of MLLMs to create our dataset. The generated QA candidates are carefully reviewed and refined by human annotators as necessary to ensure the dataset's high diversity and quality.

Raw Video Filtering Our videos are drawn from two public ego-view video datasets, RoadTextVQA [48] and EgoSchema [33]. They cover outdoor and indoor scenarios, centered around driving and housekeeping activities. RoadTextVQA has 3,222 ten-second videos from dashcam footage of people driving. EgoSchema features 5,063 three-minute videos sourced from Ego4D [14]; its videos are mostly about indoor activities such as cooking, game playing, manufacturing, *etc*. To find videos rich in scene text, we first apply a state-of-the-art scene-text detection system [17] to the raw videos and threshold for videos with a significant percentage of frames with scene text: exceeding 5% for Ego4D and 15% for RoadTextVQA (since it features more scene text). After automatic filtering, we obtain 700 videos from RoadTextVQA [48] and 1,800 from EgoSchema [14]. Since Ego4D videos are not collected for scene text study, we further filter the unqualified videos, mainly removing the tedious videos and videos that contain watermark or text present only on the clothes of the camera wearers. Here, we obtain 933/700 videos from Ego4D/RoadTextVQA.

Table 1. Data distribution and related examples of EgoTextVQA. The green words denote scene texts that appear in the video.

Category	#Q / #V	QA Examples
		EgoTextVQA-Outdoor
Location	1,505/553	*Q: Where should I go if I need a dentist?* A: Vijaya Dental Hospital.
Description	1,387/601	*Q: What is the license plate of the car in front of me?* A: Ts07ge5554.
Direction	952/468	*Q: Which direction do I go to get to Vapor Inn?* A: Continue straight and turn right at the next road.
Intention Reasoning	813/443	*Q: What should I be cautious of in this area?* A: Watch for pedestrians crossing at the crosswalk.
Others	191/152	*Q: What is the main theme of the event advertised on the bus and when and where is it happening?* A: The main theme is "Race Against HIV", and it is happening on Dec 1, 2018, at People's Plaza.
		EgoTextVQA-Indoor
Hands-on	698/284	*Q: How should I properly handle the AVT equipment?* A: Maintain a firm grip and use proper posture.
Shopping	358/78	*Q: Where can I find the "Milano" cookies?* A: Top shelf on the right.
Kitchen	335/149	*Q: What should I use to clean up countertop spills?* A: Ecover dish soap
Book-related	323/103	*Q: Where can I find information on learning the guitar?* A: Guitar for beginners.
Gameplay	301/124	*Q: What card did my opponent play before I placed the red 9?* A: Red 8.
Others	201/88	*Q: Where might I find the office rules?* A: On the notices on the wall near the entrance.

QA Generation Manually collecting QA pairs is labor-intensive; the QAs may also lack diversity. We use advanced MLLMs (*i.e*., GPT-4o) and generate the initial QAs with the following protocol: First, a video is decoded into 6 frames per second (fps) and evenly divided into 5 segments, wherein the frames without scene text are removed. Then, we uniformly sample 3 frames from each segment, and feed them to GPT-4o along with the prompts (in Appendix D) to generate 3 QA pairs. This leads to a maximum of 15 QA pairs per video. The prompts are carefully de-

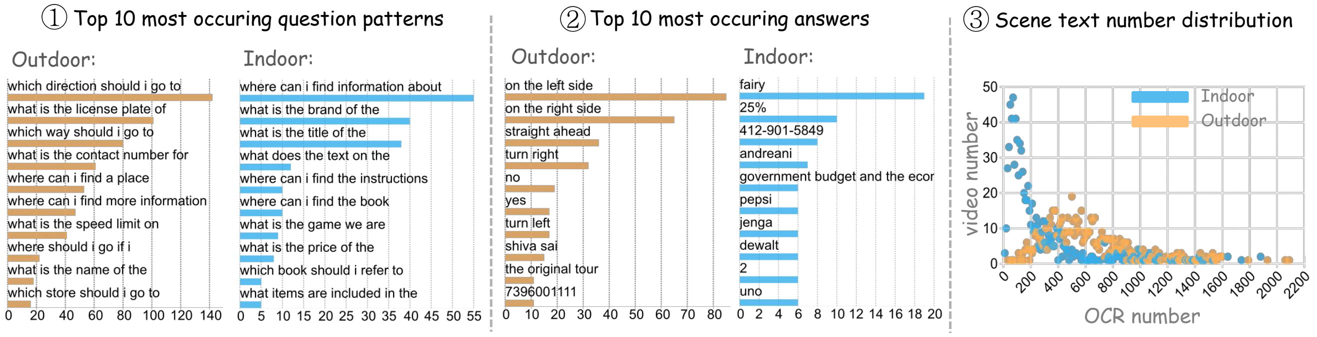


Figure 3. Distribution of QAs and OCR numbers.

Table 2. Comparison of related egocentric VQA and scene-text VQA test benchmarks. ST: Scene Text. STQ/STA: Questions/Answers contain scene text. EgoV: Egocentric Video. IntQ: Intentional Questions. QC: Question Category. OE/MC: Open-ended/Multi-choice.

Benchmark	#Que.	#Video	Ave. Len. (s)	ST	STQ (%)	STA (%)	EgoV	Indoor	Outdoor	Real Time	IntQ	QC	Task
Egocentric VQA Benchmarks													
EgoThink [10]	700	-	-	✗	-	-	✓	✓	✓	✗	✗	✓	OE
QAEgo4D [2]	1,854	166	495.1	✗	-	-	✓	✓	✓	✗	✗	✗	OE
AssistQ [53]	531	100	115.0	✗	-	-	✓	✓	✗	✗	✓	✗	MC
EgoSchema [33]	5,063	5,063	180.0	✗	-	-	✓	✓	✓	✗	✗	✗	MC
EgoMemoria [60]	7,026	629	858.5	✗	-	-	✓	✓	✓	✗	✗	✓	MC
Scene-Text VQA Benchmarks													
TextVQA [45]	5,734	-	-	✓	16	85	✗	✓	✓	✗	✗	✗	OE
ST-VQA [3]	4,163	-	-	✓	18	92	✗	✓	✓	✗	✗	✗	OE
ESTVQA [52]	5,014	-	-	✓	25	97	✗	✓	✓	✗	✗	✗	OE
M4ViteVQA [67]	2,103	680	5.7	✓	30	96	✗	✓	✓	✗	✗	✗	OE
RoadTextVQA [48]	1,052	329	10.0	✓	60	65	✓	✗	✓	✗	✗	✗	OE
EgoTextVQA (Ours)	7,064	1,507	101.7	✓	52	45	✓	✓	✓	✓	✓	✓	OE

signed to elicit questions that are: **(1)** goal-oriented, capturing real user needs by engaging with the visual scene from a first-person perspective; **(2)** aware of scene text, though not necessarily requiring the exact text transcription; **(3)** naturally expressed in a colloquial, first-person manner with referring expressions; and **(4)** challenging, as they demand an understanding of the video beyond a single image.

Manual Participation Despite using well-designed prompts, the generated QA pairs may fail to meet our strict requirements. Some generated questions are hallucinated or irrelevant to the visual content, have wrong formats, or are gibberish, *e.g.*, involving special typeset characters not part of spoken language. We first employ GPT-4o to answer the questions without videos and remove the questions that can be answered blindly ($\sim$10%). We then conduct manual filtering and correction for the remaining QAs. We provide some representative generated QA errors in Appendix A.

Specifically, we invite 9 well-trained university students for annotation. The manual check has 5 stages: (1) five annotators review for problematic QAs which may be redundant, not reflect real user needs, and irrelevant to scene text. The questions are amended where possible, or removed, leaving half of the original QAs; (2) and (3) two additional annotators re-examine and correct any mechanical issues, further reducing the QAs to 30%; (4) five annotators refine the data to improve its colloquial quality; (5) four annotators strictly refine the remaining QAs according to the criteria specified in QA generation. The participants at each stage are asked to first check the remaining QAs from the prior stage. Additionally, we manually enrich $\sim$10% time-sensitive questions to simulate real-time live QA by copying each suitable question to a different timestamp and invite five additional annotators to provide concise answers based on the visual content at the new timestamp. Ultimately, we obtain 7,064 QA pairs of 1,507 videos to form the EgoTextVQA dataset. Detailed question type and video scenario labels are provided. Figure 2 shows some representative examples; more are presented in Appendix B.5.

3.2. Dataset Analysis

Dataset Statistics EgoTextVQA characterizes assistance-seeking questions related to scene text in diverse real-life scenarios. It consists of two parts: **(1) EgoTextVQA-Outdoor** focuses on the outdoor scenarios, with 694 videos and 4,848 QA pairs that may arise when driving; **(2) EgoTextVQA-Indoor** emphasizes indoor scenarios, with 813 videos and 2,216 QA pairs that users may encounter in house-keeping activities. Distributions of videos and questions are listed in Table 1. We also analyze the most frequent question and answer patterns in Figure Fig. 3 ①, ②. The statistics reveal a clear discrepancy between QAs; the driving questions are mostly related to navigation which pure scene text cannot answer, while the house-keeping questions are information seeking, where answers can be

Table 3. Performance of SoTA MLLMs on EgoTextVQA-Outdoor. **Res.** denotes image resolution. We uniformly sample **#F** frames that the models can accept from only video content before the timestamp. In this table, we test the performance of the model on low-resolution videos (640×360 and 960×540) of the EgoTextVQA-Outdoor dataset. The human study is conducted on a random 30% of all questions, with the only constraint that answers rely on video content before the question timestamp. The **best** and second-best results are highlighted.

Method	**LLM**	**Res.**	**#F**	**Location**	**Direction**	**Description**	**Int. Reasoning**	**Others**	**Average**
Human	-	-	-	38.7 / 2.3	32.4 / 2.2	54.9 / 3.0	45.3 / 2.6	36.9 / 2.5	43.1 / 2.6
Open-source Models									
ShareGPT4Video [6]	LLaMA3-8B	-	60	11.7 / 1.5	20.8 / 1.8	5.5 / 0.7	10.7 / 1.3	10.0 / 1.2	11.5 / 1.3
CogVLM2-Video [18]	LLaMA3-8B	-	60	10.6 / 1.5	15.7 / 1.7	12.4 / 1.0	12.4 / 1.3	10.0 / 1.3	12.4 / 1.4
MiniCPM-V 2.6 [59]	Qwen2-7B	448^2	60	7.7 / 0.5	14.4 / 0.9	15.1 / 0.9	8.7 / 0.6	10.0 / 0.8	11.4 / 0.7
VILA1.5 [27]	LLaMA3-8B	384^2	32	14.6 / 1.5	20.2 / 1.6	18.2 / 1.3	18.3 / 1.5	14.7 / 1.5	17.4 / 1.5
LLaVA-NeXT-Video [64]	Qwen2-7B	384^2	60	18.1 / 1.3	22.5 / 1.5	19.6 / 1.3	19.3 / 1.4	15.7 / 1.4	19.5 / 1.4
InternVL2-8B [7]	InternLM2.5-7B	448^2	32	15.8 / 1.4	21.9 / 1.7	14.8 / 1.0	14.5 / 1.2	13.6 / 1.3	16.4 / 1.3
InternVL2-26B [7]	InternLM2-20B	448^2	16	21.5 / 1.7	27.5 / 2.0	23.5 / 1.5	22.9 / 1.7	22.5 / 1.7	23.5 / 1.7
Qwen2-VL [50]	Qwen2-7B	-	16	23.4 / 1.8	**35.2 / 2.4**	30.8 / 1.9	25.6 / 1.9	22.5 / 1.8	28.2 / **2.0**
Closed-source Models									
GPT-4o [37]	-	768^2	32	28.3 / 1.6	24.9 / 1.4	35.0 / 1.9	31.6 / 1.8	29.8 / 1.8	30.3 / 1.7
Gemini 1.5 Flash [41]	-	768^2	32	25.4 / 1.6	28.3 / 1.8	38.1 / **2.2**	28.1 / 1.8	26.6 / 2.0	30.1 / 1.9
Gemini 1.5 Pro [41]	-	768^2	32	**33.2 / 2.1**	28.3 / 1.8	**38.8 / 2.2**	**32.7 / 2.0**	**34.6 / 2.2**	**33.4 / 2.0**

extracted from scene text. Figure 3 ③ shows that the **EgoTextVQA-Indoor** videos have a long-tailed distribution in the number of OCR tokens, with the majority having less than 200. In contrast, **EgoTextVQA-Outdoor** videos are more evenly distributed with around 500.

Dataset Comparison We compare EgoTextVQA with several related benchmarks in Table 2. EgoTextVQA has several unique features. Specifically, EgoTextVQA is the first VideoQA testbed designed for *egocentric scene-text aware QA assistance* in real-world scenarios, containing 7K QA pairs across 1.5K egocentric visual scenarios, covering both indoor and outdoor activities. The QA pairs focus on scene text comprehension, with about 45% of answers and 52% of questions referencing the exact scene text. Notably, EgoTextVQA introduces *real-time QA*, providing detailed timestamps for each question. While EgoMemoria [60] also includes timestamps, its answers are based on the entire video, whereas EgoTextVQA answers are derived from video content before the question timestamp. Additionally, the questions in EgoTextVQA focus on inferring user intentions rather than pure visual understanding, with question categories provided for model analysis.

4. Experiments

4.1. Experimental Setups

Evaluation Metric We use GPT-4o mini [38] to evaluate the semantic similarity between the generated and the ground-truth (GT) answers, thus aligning closely with human scoring. Following [32], we assess the generated answers with two metrics: Accuracy (0-100%, the percentage of "yes" answers from the evaluator) and Score (0-5, with 5 being the highest match). Concretely, we format the question, predicted answer, and GT answer along with our customized evaluation prompts (in Appendix D) for GPT-4o mini to determine the "yes/no" and score.

Model Evaluation We evaluate three closed-source API-based models (GPT-4o [37], Gemini 1.5 Flash [41], Gemini 1.5 Pro [41]) and seven advanced open-source MLLMs (MiniCPM-V 2.6 [59], ShareGPT4Video [6], InternVL2 [7], VILA1.5 [27], LLaVA-NeXT-Video [64], CogVLM2-Video [18], Qwen2-VL [50]) for evaluation. The models have reported the state-of-the-art (SOTA) on existing scene-text VQA benchmarks [3, 16, 45], *e.g.*, Qwen2-VL achieves QA accuracy ∼85% on ST-VQA [3]. More details about these MLLMs and their specific prompts are presented in Appendix B.1 and D.

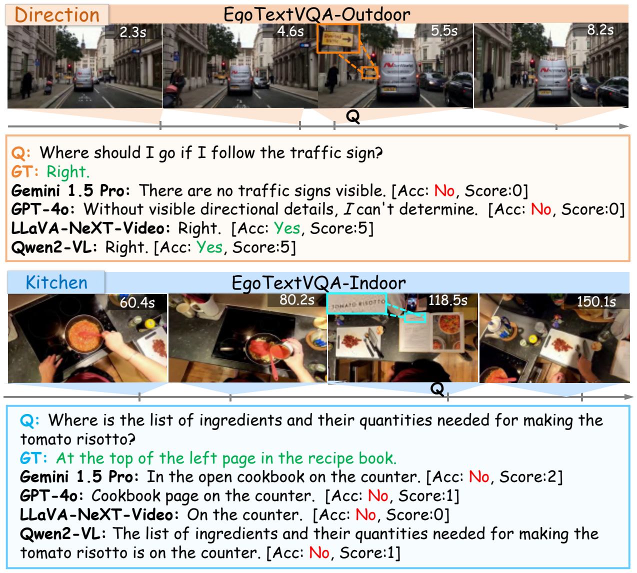
4.2. Result Analysis

Tables 3 and 4 show that all models, especially the open-source ones, struggle to perform well on EgoTextVQA. The best results, achieved by Gemini Pro 1.5 [41], are around 33∼34% on both EgoTextVQA-Outdoor and EgoTextVQA-Indoor, exceeding the best-performing open-source models by approximately ∼5% and ∼9% on EgoTextVQA-Outdoor and EgoTextVQA-Indoor, respectively. The performances vary significantly among open-source models, with Qwen2-VL [50] and LLaVA-NeXT-Video [64] performing the best, while MiniCPM-V 2.6 [59] and ShareGPT4Video [6] perform the worst.

EgoTextVQA-Outdoor Table 3 shows that MLLMs struggle across different question categories. The open-source models perform particularly poorly on "Location" and "Intention Reasoning" questions, while closed-source models, GPT-4o and Gemini 1.5 Pro/Flash, often fail on "Direction" questions. "Location" questions require understanding spatial relationships, while "Intention Reasoning" involves inferring user intent from egocentric visual context. These challenges highlight the limitations of open-source models in fine-grained visual reasoning. For "Direction" questions, closed-source models frequently respond with phrases like "unknown scene text" suggesting difficulties in recognizing or locating specific scene text (see Figure 4 Top). Overall, Gemini 1.5 Pro [41] performs best, achieving 33.4%

Table 4. Performance of SoTA MLLMs on EgoTextVQA-Indoor. **Res.** denotes image resolution. We uniformly sample the **#F** frames that the model can accept from all video frames before the timestamp as input. In this table, we test the performance of the model on the videos (the resolution mainly includes 480×360 and 640×360) of the EgoTextVQA-Indoor dataset. The human study is conducted on a randomly sampled 30% of all questions. The best and second-best Accuracy / Score results are **bolded** and <u>underlined</u> respectively.

Method	**LLM**	**Res.**	**#F**	**Hands-on**	**Kitchen**	**Shopping**	**Gameplay**	**Book-Related**	**Others**	**Average**
Human	-	-	-	26.6 / 1.9	31.3 / 2.1	18.2 / 1.4	34.8 / 2.2	27.2 / 1.8	32.8 / 2.1	27.7 / 1.9
Open-source Models										
ShareGPT4Video [6]	LLaMA3-8B	-	128	7.0 / 1.0	5.1 / 1.0	5.9 / 1.3	12.3 / 1.2	7.1 / 1.1	4.0 / 0.9	7.0 / 1.1
CogVLM2-Video [18]	LLaMA3-8B	-	128	12.3 / 1.4	11.6 / 1.4	10.9 / 1.4	23.6 / 1.8	11.2 / 1.2	13.4 / 1.4	13.5 / 1.4
MiniCPM-V 2.6 [59]	Qwen2-7B	448^2	120	14.2 / 0.8	8.1 / 0.5	10.3 / 0.6	17.6 / 1.1	16.4 / 1.0	12.9 / 0.8	13.3 / 0.8
VILA1.5 [27]	LLaMA3-8B	384^2	32	16.5 / 1.5	18.8 / 1.5	15.1 / 1.5	24.9 / 1.9	18.9 / 1.6	17.4 / 1.5	18.2 / 1.6
InternVL2-8B [7]	InternLM2.5-7B	448^2	32	13.5 / 1.2	12.5 / 1.1	14.5 / 1.3	13.3 / 1.2	14.9 / 1.3	12.4 / 1.1	13.6 / 1.2
InternVL2-26B [7]	InternLM2-20B	448^2	16	20.3 / 1.6	22.4 / 1.8	18.2 / 1.6	25.6 / 1.8	19.2 / 1.6	21.9 / 1.7	21.0 / 1.7
Qwen2-VL [50]	Qwen2-7B	448^2	48	21.5 / 1.8	21.5 / 1.7	22.5 / 1.8	29.9 / 2.1	22.0 / 1.8	26.4 / 1.8	23.3 / 1.8
LLaVA-NeXT-Video [64]	Qwen2-7B	384^2	128	24.4 / 1.7	23.6 / 1.7	21.8 / 1.7	29.9 / 1.9	27.9 / 1.9	27.8 / 1.8	25.4 / 1.8
Closed-source Models										
GPT-4o [37]	-	768^2	60	25.4 / 1.6	30.2 / 1.9	29.6 / 1.9	32.9 / 1.9	29.7 / 1.9	23.9 / 1.5	28.3 / 1.8
Gemini 1.5 Flash [41]	-	768^2	60	<u>29.7</u> / <u>2.0</u>	<u>33.7</u> / <u>2.1</u>	<u>32.3</u> / <u>2.1</u>	<u>34.5</u> / <u>2.3</u>	<u>33.6</u> / **2.2**	<u>30.1</u> / <u>2.0</u>	<u>32.0</u> / <u>2.1</u>
Gemini 1.5 Pro [41]	-	768^2	60	**33.2 / 2.1**	**35.2 / 2.1**	**33.8 / 2.1**	**39.5 / 2.4**	**34.3** / <u>2.1</u>	**30.9 / 2.0**	**34.4 / 2.1**



Q: Where should I go if I follow the traffic sign?
GT: Right.
Gemini 1.5 Pro: There are no traffic signs visible. [Acc: No, Score:0]
GPT-4o: Without visible directional details, I can't determine. [Acc: No, Score:0]
LLaVA-NeXT-Video: Right. [Acc: Yes, Score:5]
Qwen2-VL: Right. [Acc: Yes, Score:5]

Q: Where is the list of ingredients and their quantities needed for making the tomato risotto?
GT: At the top of the left page in the recipe book.
Gemini 1.5 Pro: In the open cookbook on the counter. [Acc: No, Score:2]
GPT-4o: Cookbook page on the counter. [Acc: No, Score:1]
LLaVA-NeXT-Video: On the counter. [Acc: No, Score:0]
Qwen2-VL: The list of ingredients and their quantities needed for making the tomato risotto is on the counter. [Acc: No, Score:1]

Figure 4. Result Visualization.

accuracy and a score of 2.0, making it the SOTA model on EgoTextVQA-Outdoor. However, it still trails human performance by ∼10%, indicating that even advanced MLLMs struggle with scene text perception in complex, dynamic, egocentric environments.

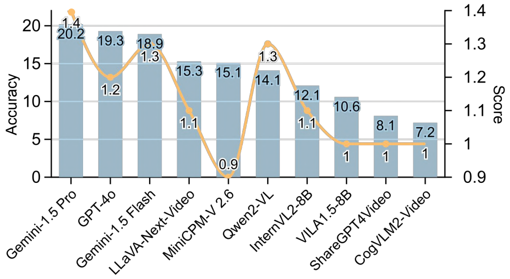


Figure 5. Performance of MLLMs on the real-time QA subset of EgoTextVQA-Outdoor (∼623 QA pairs).

In Figure 5, we further analyze model behavior on the real-time QA subset of EgoTextVQA-Outdoor. The results show that all models perform significantly worse on this subset, with the highest accuracy reaching only 20.2% (*vs.* 33.4% on the full set), underscoring the substantial challenge of egocentric live QA. Interestingly, MiniCPM-V 2.6 [59] ranks among the top two open-source models for real-time QA, despite being the worst on the full set. However, a closer examination of its predictions reveals that MiniCPM-V 2.6 primarily excels at answering unanswerable questions with responses like "I did not find it" whereas other models tend to hallucinate answers. This finding suggests that current models remain fundamentally weak in generating meaningful answers that satisfy user needs.

EgoTextVQA-Indoor Table 4 shows that MLLMs struggle with the questions across indoor scenarios, especially in "Kitchen" and "Shopping" activities. Our qualitative analysis (see Figure 4 Bottom) shows that people often ask about the location of ingredients and utensils in the kitchen, where the models tend to provide vague positional information (*e.g.*, "`on the left`" rather than "`on the top shelf on the left`"). In shopping, it is challenging for the models to locate a user-specified item among large and densely packed products. ShareGPT4Video [6] performs the worst despite training on Ego4D videos, suggesting limited scene-text awareness. Meanwhile, closed-source models perform relatively better across scenarios.

Intriguingly, we find that humans perform even worse than the closed-source models. Feedback from our human annotators and a careful analysis of the answers suggest that the primary reasons are as follows: (1) scene text recognition is challenging for humans in Ego4D videos due to lower resolution, motion blur, occlusion, and long video; (2) answering the questions needs external knowledge beyond simple scene text recognition. EgoTextVQA-

Table 5. Effects of taking as input an image (“TS Frame”) and three images (“QA Frames” analogous to QA generation) at the question timestamp.

Method	**Input**	**EgoTextVQA-Outdoor**		**EgoTextVQA-Indoor**	
		Accuracy	Score	Accuracy	Score
InternVL2-8B [7]	Video	16.4	1.3	13.6	1.2
	TS Frame	15.8 ↓ 0.6	1.2 ↓ 0.1	6.5 ↓ 7.1	0.7 ↓ 0.5
	QA Frames	18.5 ↑ 2.1	1.4 ↑ 0.1	17.7 ↑ 4.1	1.4 ↑ 0.2
MiniCPM-V 2.6 [59]	Video	11.4	0.7	13.3	0.8
	TS Frame	7.9 ↓ 3.5	0.2 ↓ 0.5	3.8 ↓ 9.5	0.3 ↓ 0.4
	QA Frames	9.3 ↓ 2.1	0.6 ↓ 0.1	13.7 ↑ 0.4	0.9 ↑ 0.1
Qwen2-VL [50]	Video	28.2	2.0	23.3	1.8
	TS Frame	30.9 ↑ 2.7	2.1 ↑ 0.1	15.7 ↓ 7.6	1.4 ↓ 0.4
	QA Frames	31.9 ↑ 3.7	**2.2** ↑ 0.2	30.7 ↑ 7.4	2.2 ↑ 0.4
GPT-4o [37]	Video	30.3	1.7	28.3	1.8
	TS Frame	27.7 ↓ 2.6	1.6 ↓ 0.1	14.8 ↓ 13.5	0.9 ↓ 0.9
	QA Frames	30.4 ↑ 0.1	1.8 ↑ 0.1	**41.6** ↑ 13.3	**2.4** ↑ 0.6
Gemini 1.5 Pro [41]	Video	**33.4**	2.0	34.4	2.1
	TS Frame	30.4 ↓ 3.0	1.8 ↓ 0.2	15.8 ↓ 18.6	1.1 ↓ 1.0
	QA Frames	26.3 ↓ 7.1	1.6 ↓ 0.4	38.2 ↑ 3.8	2.4 ↑ 0.3

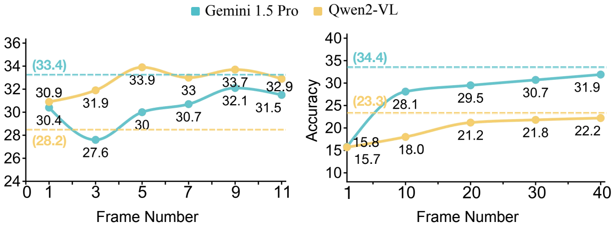
Outdoor mainly involves knowledge about store or restaurant chains (*e.g.*, “`Home Depo`” sells building materials), and EgoTextVQA-Indoor focuses on items like kitchen utensils, board games, and specialized books (*e.g.*, identifying a “`mandoline`” slices potatoes or a “`quantum`” field theory book); (3) answer diversity also results in relatively lower human performance. The first two points underscore the challenges and significance of research on egocentric scene-text QA assistance. The third point unveils a limitation of our dataset, which we wish to solve in the future by enriching the GT answers for each question. We provide more human study analysis in Appendix B.4 and C.

Outdoor Driving *vs*. Indoor House-Keeping While humans perform worse on EgoTextVQA-Indoor, the MLLMs are basically at the same performance level on both datasets. In particular, we find Qwen2-VL [50] stands out as the best open-source model on EgoTextVQA-Outdoor, whereas LLaVA-NeXT-Video [64] ranks the top on EgoTextVQA-Indoor. We speculate that LLaVA-NeXT-Video is better at coping with long videos. More model design and case analysis are provided in Appendix B.2 and B.5.

5. Heuristic Solution Investigations

To facilitate future study, we conduct extensive experiments to thoroughly analyze the challenges presented by EgoTextVQA in Tables 5 to 7 and Figures 5 to 7. Our analysis centers on the following four questions:

Q1: Can image-level understanding solve our task? Sometimes. We test the models by inputting a single frame at the question timestamp (“TS Frame”). Compared with using uniformly sampled video frames, Table 5 shows that almost all the models have a significant performance drop ranging from 0.6∼3.5% on EgoTextVQA-Outdoor and as much as 7.1∼18.6% on EgoTextVQA-Indoor. This signifies the importance of video-level reasoning, especially for EgoTextVQA-Indoor, where questions feature longer video action understanding. Notably, Qwen2-VL *improves* its performance when fed with a single frame on EgoTextVQA-Outdoor. We speculate that Qwen2-VL excels at scene text recognition at static frames but shows insufficient capability in handling information redundancy across multiple frames.



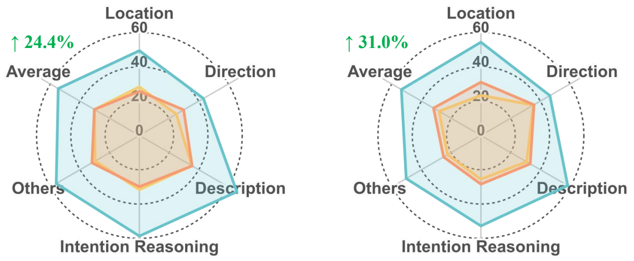
(a) EgoTextVQA-Outdoor (b) EgoTextVQA-Indoor

Figure 6. Effects with different numbers of frame inputs sampled at 1fps backward from the question timestamp.

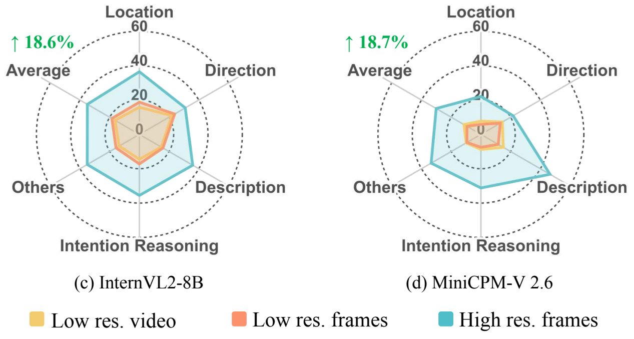
Q2: Is temporal grounding important? Sometimes. We use the three frames from question generation as model inputs (“QA Frames”), which consistently improve performance on EgoTextVQA-Indoor but have inconsistent effects on EgoTextVQA-Outdoorin Table 5. Qwen2-VL [50] and InternVL2-8B [7] increase their accuracy by 3.7% and 2.1% respectively but MiniCPM-V 2.6 [59] and Gemini 1.5 Pro decrease by 2.1% and 7.1%, respectively. Overall, the improvement suggests there is space for models to optimize their temporal grounding, *i.e.*, to localize the key frames for answering. In contrast, the decrease in performance indicates that the models may be vulnerable to seriously reduced information input. It may also reflect the great efforts made by annotators for correcting mistakes from the automatic QA generation, as also exemplified by the poor results of GPT-4o (QA generator).

We investigate a question-timestamp aware sampling strategy: starting from the question timestamp and sampling backwards at 1 fps. This strategy is inspired by intuition that people often pose questions based on their most recent visual settings. Figure 6 shows that incorporating such a sampling strategy significantly improves Qwen2-VL’s performance on EgoTextVQA-Outdoor, even with fewer frames. However, for Gemini 1.5 Pro, the opposite is true and more frames are key for better performance. On EgoTextVQA-Indoor, we find that both Qwen2-VL and Gemini 1.5 Pro shrink their performances compared with standard uniform sampling across the whole video, possibly due to a lack of necessary information for long video reasoning.

Q3: Are high-resolution frames for scene text critical? Yes. Most existing MLLMs are limited to accepting fixed, low-resolution images as input. When higher resolutions are used, the number of input frames is limited. We compare the performance of the model when using the “low-resolution video” and the three key “low-resolution frames” and “high-resolution frames” specified for question generation as input. Here we choose GPT-4o [37] and three open-



(a) GPT-4o (b) Qwen2-VL



(c) InternVL2-8B (d) MiniCPM-V 2.6

Figure 7. Higher resolution (1920×1080, 1280×720) QA Frames generally improve performance on EgoTextVQA-Outdoor.

source models that can accept frames with resolutions of up to 1920×1080 and 1280×720. Figure 7 shows that answering with three key frames ("QA Frames") at low resolution is comparable to uniform video sampling. Yet, with the increase of resolution, GPT-4o [37], Qwen2-VL [50], InternVL2-8B [7], and MiniCPM-V 2.6 [59] all show profound improvements, up to 18.6∼31.0%.

Aside from increasing the *global frame* resolution, we focus on increasing the *local scene-text* resolution while keeping the global frame resolution unchanged. We use Microsoft Azure OCR for scene text detection and apply the SOTA scene text super-resolution model DiffTSR [65] to increase the resolution of the detected regions to 380 × 128. To project the super-resolution scene text back to the original low-resolution video frames, we consider three options: resize to 1.00 ×, 1.25×, and 1.50× of the original scene text size. Table 6 shows that higher-resolution scene text improves performance on EgoTextVQA-Outdoor but yields inconsistent results on EgoTextVQA-Indoor. We speculate that scene text detection in Ego4D videos is more challenging than in RoadTextVQA due to lower resolution and complex perspectives. This is further supported by the smaller performance gains on EgoTextVQA-Indoor (*vs*. Outdoor) with additional OCR results in Table 7.

Q4: Is text alone enough to answer questions? No. We feed the best-performing open-source and closed-source models with only the scene text, but as Table 7 shows, none of these models can answer the questions reasonably. Pure scene text inputs are sufficient for answering some questions, but video input is more important for almost all the models. Furthermore, adding scene text together with video input further boosts the performances significantly, leading to the highest results in all our explorations. For instance, GPT-4o achieves 52.9% on EgoTextVQA-Outdoor and Gemini's accuracy on EgoTextVQA-Indoor reaches to 39.5%. The findings demonstrate the importance of both video and scene text inputs for better QA assistance. More heuristic solution investigations can be found in Appendix B.3.

Table 6. Effects of scene-text resolution and area ratio. *Scale* parameter is to control the magnification ratio of the scene text regions. SR: Super Resolution Scene Text. We experiment with 30% of the data for efficiency.

Method	**Scale**	**EgoTextVQA-Outdoor**		**EgoTextVQA-Indoor**	
		Accuracy	Score	Accuracy	Score
Qwen2-VL [50]	-	28.2	2.0	23.3	1.8
Qwen2-VL w/ SR	1.00 ×	31.9	2.1	24.2	1.8
	1.25 ×	33.1	2.2	23.5	1.8
	1.50 ×	34.1	2.2	22.3	1.7
Gemini 1.5 Pro [41]	-	33.4	2.0	34.4	2.1
Gemini 1.5 Pro w/ SR	1.00 ×	36.9	2.2	31.7	2.1
	1.25 ×	38.1	2.2	30.1	2.0
	1.50 ×	38.1	2.2	30.2	2.0

Table 7. Effects of different modality inputs. V: Video. Q: Question. ST: Scene Text.

Method	**Input**			**EgoTextVQA-Outdoor**		**EgoTextVQA-Indoor**	
	V	Q	ST	Accuracy	Score	Accuracy	Score
Qwen2-VL [50]	-	✓	-	2.1	0.5	2.5	0.5
	-	✓	✓	15.4	1.2	13.3	1.1
	✓	✓	-	28.2	2.0	23.3	1.8
	✓	✓	✓	39.9 ↑ 11.7	2.5 ↑ 0.5	23.6 ↑ 0.3	1.8 ↑ 0.0
LLaVA-Next-Video [64]	-	✓	-	8.5	1.1	6.0	1.0
	-	✓	✓	20.6	1.4	15.6	1.2
	✓	✓	-	19.5	1.4	22.7	1.8
	✓	✓	✓	37.4 ↑ 17.9	2.3 ↑ 0.9	30.8 ↑ 8.1	2.1 ↑ 0.3
GPT-4o [37]	-	✓	-	5.26	0.84	5.7	0.9
	-	✓	✓	24.1	1.71	17.6	1.4
	✓	✓	-	30.3	1.7	28.3	1.8
	✓	✓	✓	**52.9** ↑ 22.6	**3.0** ↑ 1.3	37.9 ↑ 9.6	2.3 ↑ 0.5
Gemini 1.5 Pro [41]	-	✓	-	4.3	0.8	4.6	0.8
	-	✓	✓	18.2	1.3	15.2	1.0
	✓	✓	-	33.4	2.0	34.4	2.1
	✓	✓	✓	49.5 ↑ 16.1	2.9 ↑ 0.9	**39.5** ↑ 5.1	**2.4** ↑ 0.3

6. Conclusion

To provide a benchmark for research on egocentric scene-text aware QA assistance, we carefully construct the EgoTextVQA dataset. EgoTextVQA highlights QA assistance involving scene text from an ego-perspective in diverse real-life scenarios, including outdoor driving and indoor housekeeping. The questions reflect the real user needs, yet the visual attention is often not focused on the scene text. With well-classified visual scenarios and question types, we comprehensively analyze models that excel on existing scene-text VQA benchmarks and find that they struggle on EgoTextVQA. We further explore heuristic solutions and provide insights for improvements, With these efforts, we hope this work can complement existing VideoQA research in advancing real-life egocentric QA assistance.

---

https://azure.microsoft.com/en-us/services/cognitive-services

7. Acknowledgements

This research was funded by the NUS Artificial Intelligence Institute (NAII) seed grant number NAII-SF-2024-003. Additionally, this work was supported by the National Key R&D Program of China (NO.2024YFB3311602), the National Natural Science Foundation of China (72188101, 62020106007, 62272435), the Major Project of Anhui Province (202203a05020011), the Anhui Provincial Natural Science Foundation (2408085J040), the Fundamental Research Funds for the Central Universities (JZ2023YQTD0072, JZ2024HGTG0309, and JZ2024AHST0337), and the New Cornerstone Science Foundation through the XPLORER PRIZE. We sincerely appreciate the dedication and meticulous efforts of our dataset annotators, whose contributions were invaluable in ensuring data quality and advancing this research.

References

[1] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023. 12

[2] Leonard Bärmann and Alex Waibel. Where did i leave my keys? - episodic-memory-based question answering on egocentric videos. In *CVPR Workshops*, pages 1560–1568, 2022. 1, 4

[3] Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusinol, Ernest Valveny, CV Jawahar, and Dimosthenis Karatzas. Scene text visual question answering. In *ICCV*, pages 4291–4301, 2019. 2, 3, 4, 5, 12

[4] Lukas Blecher, Guillem Cucurull, Thomas Scialom, and Robert Stojnic. Nougat: Neural optical understanding for academic documents. *arXiv preprint arXiv:2308.13418*, 2023. 12

[5] Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, et al. Internlm2 technical report. *arXiv preprint arXiv:2403.17297*, 2024. 12

[6] Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Bin Lin, Zhenyu Tang, et al. Sharegpt4video: Improving video understanding and generation with better captions. *arXiv preprint arXiv:2406.04325*, 2024. 2, 3, 5, 6, 12

[7] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024. 3, 5, 6, 7, 8, 12

[8] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*, pages 24185–24198, 2024. 12, 13, 14

[9] Sijie Cheng, Kechen Fang, Yangyang Yu, Sicheng Zhou, Bohao Li, Ye Tian, Tingguang Li, Lei Han, and Yang Liu. Videgothink: Assessing egocentric video understanding capabilities for embodied ai. *arXiv preprint arXiv:2410.11623*, 2024. 1

[10] Sijie Cheng, Zhicheng Guo, Jingwen Wu, Kechen Fang, Peng Li, Huaping Liu, and Yang Liu. Egothink: Evaluating first-person perspective thinking capability of vision-language models. In *CVPR*, pages 14291–14302, 2024. 4, 16

[11] Shangzhe Di and Weidi Xie. Grounded question-answering in long egocentric videos. In *CVPR*, pages 12934–12943, 2024. 1

[12] Chenyou Fan. Egovqa-an egocentric video question answering benchmark dataset. In *ICCV Workshop*, 2019. 1

[13] Chaoyou Fu, Yuhan Dai, Yondong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*, 2024. 3

[14] 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 *CVPR*, pages 18995–19012, 2022. 1, 3

[15] Dan Guo, Hui Wang, and Meng Wang. Context-aware graph inference with knowledge distillation for visual dialog. *IEEE TPAMI*, 44(10):6056–6073, 2021. 1

[16] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *CVPR*, pages 3608–3617, 2018. 2, 5

[17] Haibin He, Maoyuan Ye, Jing Zhang, Juhua Liu, and Dacheng Tao. Gomatching: A simple baseline for video text spotting via long and short term matching. *arXiv preprint arXiv:2401.07080*, 2024. 3

[18] Wenyi Hong, Weihan Wang, Ming Ding, Wenmeng Yu, Qingsong Lv, Yan Wang, Yean Cheng, Shiyu Huang, Junhui Ji, Zhao Xue, et al. Cogvlm2: Visual language models for image and video understanding. *arXiv preprint arXiv:2408.16500*, 2024. 2, 3, 5, 6, 12

[19] Soumya Jahagirdar, Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Understanding video scenes through text: Insights from text-based video question answering. In *ICCV*, pages 4646–4650, 2023. 2

[20] Baoxiong Jia, Ting Lei, Song-Chun Zhu, and Siyuan Huang. Egotaskqa: Understanding human tasks in egocentric videos. In *NeurIPS*, pages 3343–3360, 2022. 1

[21] Bo Li, Kaichen Zhang, Hao Zhang, Dong Guo, Renrui Zhang, Feng Li, Yuanhan Zhang, Ziwei Liu, and Chunyuan Li. Llava-next: Stronger llms supercharge multimodal capabilities in the wild, 2024. 12

[22] Chenxia Li, Weiwei Liu, Ruoyu Guo, Xiaoting Yin, Kaitao Jiang, Yongkun Du, Yuning Du, Lingfeng Zhu, Baohua Lai, Xiaoguang Hu, et al. Pp-ocrv3: More attempts for the improvement of ultra lightweight ocr system. *arXiv preprint arXiv:2206.03001*, 2022. 12

[23] Junyan Li, Delin Chen, Tianle Cai, Peihao Chen, Yining Hong, Zhenfang Chen, Yikang Shen, and Chuang Gan. Flexattention for efficient high-resolution vision-language models. In *ECCV*, 2025. 3

[24] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *CVPR*, pages 22195–22206, 2024. 3

[25] Yicong Li, Xiang Wang, Junbin Xiao, Wei Ji, and Tat-Seng Chua. Invariant grounding for video question answering. In *CVPR*, pages 2928–2937, 2022. 2

[26] Yicong Li, Xiang Wang, Junbin Xiao, Wei Ji, and Tat-Seng Chua. Transformer-empowered invariant grounding for video question answering. *IEEE TPAMI*, 2023. 1

[27] Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. In *CVPR*, pages 26689–26699, 2024. 3, 5, 6, 12

[28] Kevin Qinghong Lin, Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Z Xu, Difei Gao, Rong-Cheng Tu, Wenzhe Zhao, Weijie Kong, et al. Egocentric video-language pretraining. *NeurIPS*, 35:7575–7586, 2022. 1

[29] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 12

[30] Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xucheng Yin, Cheng lin Liu, Lianwen Jin, and Xiang Bai. Ocrbench: On the hidden mystery of ocr in large multimodal models, 2024. 2, 3

[31] Gen Luo, Yiyi Zhou, Yuxin Zhang, Xiawu Zheng, Xiaoshuai Sun, and Rongrong Ji. Feast your eyes: Mixture-of-resolution adaptation for multimodal large language models. *arXiv preprint arXiv:2403.03003*, 2024. 3

[32] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*, 2023. 5

[33] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. In *NeurIPS*, pages 46212–46244, 2023. 1, 3, 4

[34] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *WACV*, pages 2200–2209, 2021. 3

[35] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In *WACV*, pages 1697–1706, 2022.

[36] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *ICDAR*, pages 947–952, 2019. 2, 3, 12

[37] OpenAI. Gpt-4o system card. 2024. 5, 6, 7, 8, 12, 14, 16

[38] OpenAI. Gpt-4o mini: advancing cost-efficient intelligence. 2024. 5, 16

[39] Shraman Pramanick, Yale Song, Sayan Nag, Kevin Qinghong Lin, Hardik Shah, Mike Zheng Shou, Rama Chellappa, and Pengchuan Zhang. Egovlpv2: Egocentric video-language pre-training with fusion in the backbone. In *ICCV*, pages 5285–5297, 2023. 1

[40] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021. 12

[41] Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. 2, 5, 6, 7, 8, 12, 13, 14, 16

[42] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 12

[43] Xindi Shang, Donglin Di, Junbin Xiao, Yu Cao, Xun Yang, and Tat-Seng Chua. Annotating objects and relations in user-generated videos. In *Proceedings of the 2019 on International Conference on Multimedia Retrieval*, pages 279–287, 2019. 3

[44] Noam Shazeer. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020. 12

[45] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *CVPR*, pages 8317–8326, 2019. 2, 3, 4, 5, 12

[46] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023. 12

[47] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 12

[48] George Tom, Minesh Mathew, Sergi Garcia-Bordils, Dimosthenis Karatzas, and CV Jawahar. Reading between the lanes: Text videoqa on the road. In *ICDAR*, pages 137–154, 2023. 2, 3, 4

[49] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 12

[50] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 2, 3, 5, 6, 7, 8, 12, 13, 14

[51] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*, 2023. 3, 12

[52] Xinyu Wang, Yuliang Liu, Chunhua Shen, Chun Chet Ng, Canjie Luo, Lianwen Jin, Chee Seng Chan, Anton van den

Hengel, and Liangwei Wang. On the general value of evidence, and bilingual scene-text visual question answering. In *CVPR*, pages 10126–10135, 2020. 2, 4

[53] Benita Wong, Joya Chen, You Wu, Stan Weixian Lei, Dongxing Mao, Difei Gao, and Mike Zheng Shou. Assistq: Affordance-centric question-driven task completion for egocentric assistant. In *ECCV*, pages 485–501. Springer, 2022. 1, 4

[54] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *CVPR*, pages 9777–9786, 2021. 3

[55] Junbin Xiao, Nanxin Huang, Hangyu Qin, Dongyang Li, Yicong Li, Fengbin Zhu, Zhulin Tao, Jianxing Yu, Liang Lin, Tat-Seng Chua, and Angela Yao. Videoqa in the era of llms: An empirical study. *arXiv preprint arXiv:2408.04223*, 2024. 1

[56] Xun Yang, Fuli Feng, Wei Ji, Meng Wang, and Tat-Seng Chua. Deconfounded video moment retrieval with causal intervention. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 1–10, 2021. 3

[57] Xun Yang, Shanshan Wang, Jian Dong, Jianfeng Dong, Meng Wang, and Tat-Seng Chua. Video moment retrieval with cross-modal neural architecture search. *IEEE Transactions on Image Processing*, 31:1204–1216, 2022. 3

[58] Xun Yang, Jianming Zeng, Dan Guo, Shanshan Wang, Jianfeng Dong, and Meng Wang. Robust video question answering via contrastive cross-modality representation learning. *Science China Information Sciences*, 67(10):202104, 2024. 1

[59] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024. 2, 3, 5, 6, 7, 8, 12

[60] Hanrong Ye, Haotian Zhang, Erik Daxberger, Lin Chen, Zongyu Lin, Yanghao Li, Bowen Zhang, Haoxuan You, Dan Xu, Zhe Gan, Jiasen Lu, and Yinfei Yang. Mm-ego: Towards building egocentric multimodal llms. *arXiv preprint arXiv:2410.07177*, 2024. 1, 4, 5

[61] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, pages 11975–11986, 2023. 12

[62] Haoyu Zhang, Meng Liu, Zixin Liu, Xuemeng Song, Yaowei Wang, and Liqiang Nie. Multi-factor adaptive vision selection for egocentric video question answering. In *ICML*, 2024. 1

[63] Ruiyi Zhang, Yufan Zhou, Jian Chen, Jiuxiang Gu, Changyou Chen, and Tong Sun. Llava-read: Enhancing reading ability of multimodal language models. *arXiv preprint arXiv:2407.19185*, 2024. 3

[64] Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. Llava-next: A strong zero-shot video understanding model, 2024. 3, 5, 6, 7, 8, 12, 13

[65] Yuzhe Zhang, Jiawei Zhang, Hao Li, Zhouxia Wang, Luwei Hou, Dongqing Zou, and Liheng Bian. Diffusion-based blind text image super-resolution. In *CVPR*, pages 25827–25836, 2024. 8

[66] Yi-Fan Zhang, Qingsong Wen, Chaoyou Fu, Xue Wang, Zhang Zhang, Liang Wang, and Rong Jin. Beyond llava-hd: Diving into high-resolution large multimodal models. *arXiv preprint arXiv:2406.08487*, 2024. 3

[67] Minyi Zhao, Bingjia Li, Jie Wang, Wanqing Li, Wenjing Zhou, Lan Zhang, Shijie Xuyang, Zhihang Yu, Xinkun Yu, Guangze Li, et al. Towards video text visual question answering: Benchmark and baseline. In *NeurIPS*, pages 35549–35562, 2022. 2, 4

[68] Sheng Zhou, Dan Guo, Jia Li, Xun Yang, and Meng Wang. Exploring sparse spatial relation in graph inference for text-based vqa. *IEEE TIP*, 32:5060–5074, 2023. 2

[69] Sheng Zhou, Junbin Xiao, Xun Yang, Peipei Song, Dan Guo, Angela Yao, Meng Wang, and Tat-Seng Chua. Scene-text grounding for text-based video question answering. *arXiv preprint arXiv:2409.14319*, 2024. 2

EgoTextVQA: Towards Egocentric Scene-Text Aware Video Question Answering

Supplementary Material

A. EgoTextVQA Dataset

A.1. Manual Participation

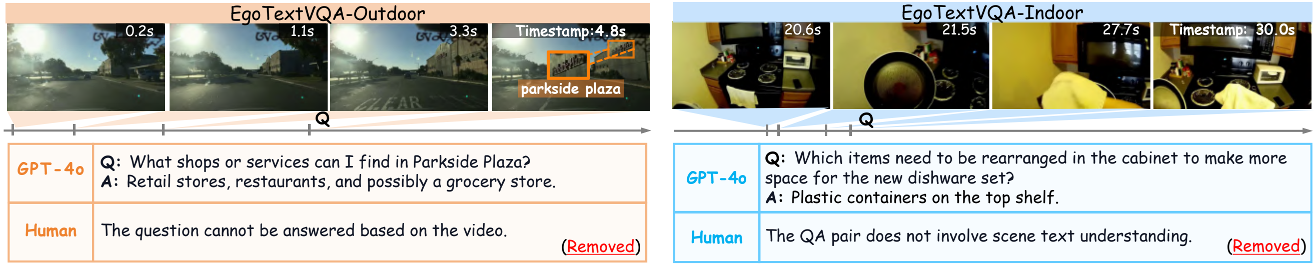
We present examples of QA pairs generated by GPT-4o in Figure 8 to highlight the issues of automatic generation and underscore the value of manual correction. The primary problems observed are as follows: **(a) Hallucinated Answers:** The generated answers are unseen from the visual environment and cannot be confirmed by the annotators. **(b) Scene Text Irrelevance:** The questions have vague references and fail to incorporate scene text understanding for answers. **(c) Scene Text Errors:** The questions or answers contain incorrect scene text. **(d) and (f) Non-colloquial Questions:** The questions are mechanical; they are phrased unnaturally and do not align well with daily spoken language. **(e) Not Reflect User Needs:** The question does not reflect real user needs and hardly occurs in human daily life. After manual participation, about 70% of the generated QAs are deleted and 30% of the remaining QAs are revised.

B. Experiment

B.1. Model Details

We provide a concise introduction to the MLLMs evaluated in Section 4, as outlined below:

- **GPT-4o** [37] advances the GPT-4 family towards more natural human-computer interactions.
- **Gemini 1.5 Pro** [41] builds on Gemini 1.0's [47] research advances and multimodal capabilities and it is optimized for a wide-range of reasoning tasks.
- **Gemini 1.5 Flash** [41] is a model from Gemini 1.5 family offering low latency and enhanced performance.
- **Qwen2-VL** [50] employs a ViT-675M [40] as the visual encoder, Qwen2-7B as the language model, and an MLP projector. It improves upon Qwen-VL [1] with (1) naive dynamic resolution, allowing ViT to handle images of varying resolutions, and (2) multimodal rotary position embedding, which decomposes positional encoding into temporal, height, and width components. Qwen2-VL is pre-trained on diverse datasets, including image-text pairs, OCR data, interleaved articles, VQA datasets, video dialogues, and image knowledge sources, enabling a stronger multimodal understanding.
- **LLaVA-NeXT-Video** [64] choose SigLIP-SO400M [61] as the visual encoder, Qwen2 [50] as the language model, and a two-layer MLP as the projector. It utilizes the AnyRes [29] technique to segment high-resolution images for the visual encoder and extends this approach to video processing. LLaVA-NeXT-Video has excellent reasoning, OCR, and world knowledge capabilities, achieving strong performance in video-based multimodal tasks.
- **VILA1.5** [27] integrates CLIP-L [40] as the visual encoder, LLaMA-2 [49] as the language model, and a linear projector. It fine-tunes on a mix of internal data, including OCR-VQA [36] and ST-VQA [3], and improves contextual learning by unfreezing the LLM during interleaved image-text pre-training. VILA1.5 excels in video reasoning, in-context learning, visual chain-of-thought reasoning, and world knowledge.
- **InternVL2-8B** [7] integrates InternViT-300M [8] with InternLM2.5-7B [5] via a randomly initialized MLP projector. It is trained on OCR datasets generated by PaddleOCR [22], utilizing Chinese images from Wukong and English images from LaionCOCO [42]. Building on the strong visual representations and high-resolution image processing capabilities of InternVL1.5 [7], InternVL2 incorporates instruction tuning, enabling competitive performance in document and chart comprehension, infographics QA, scene text understanding, OCR, and multimodal reasoning tasks.
- **CogVLM2-Video** [18] utilizes the EVA-CLIP [46] as the visual encoder, LLaMA3-8B as the language model, and a 2×2 convolutional layer followed by a SwiGLU [44] as the adapter. Unlike CogVLM [51], CogVLM2 improves pre- and post-training data diversity and quality. The Synthetic OCR Dataset, a key pre-training resource, includes four OCR scenarios: (1) synthetic OCR images with text generated in Python, (2) real-world images with PaddleOCR [22], (3) academic papers with extracted LaTeX via Nougat [4], and (4) HTML/LaTeX-rendered tables and formulae. CogVLM2-Video adapts CogVLM2 for videos, enhancing open-domain QA with temporal localization and timestamp-aware QA.
- **MiniCPM-V 2.6** [59] employs SigLIP-SO400M [61] as the visual encoder, Qwen2 [50] as the language model, and a compression module with one-layer cross-attention and a moderate number of queries as the projector. Its training includes pre-training on English and Chinese image captioning and OCR data, followed by fine-tuning on datasets like TextVQA [45], OCR-VQA [36], and ST-VQA [3]. MiniCPM-V 2.6 excels in conversational and reasoning tasks across multiple images and videos, with high-resolution perception enabling features like table-to-markdown conversion and OCR transcription.
- **ShareGPT4Video** [6] builts on LLaVA-Next-8B [21]. Based on the proposed ShareGPT4Video [6] dataset, the proposed captioning model ShareCaptioner-Video generates high-quality captions with detailed temporal descrip-



GPT-4o Q: What shops or services can I find in Parkside Plaza? A: Retail stores, restaurants, and possibly a grocery store.

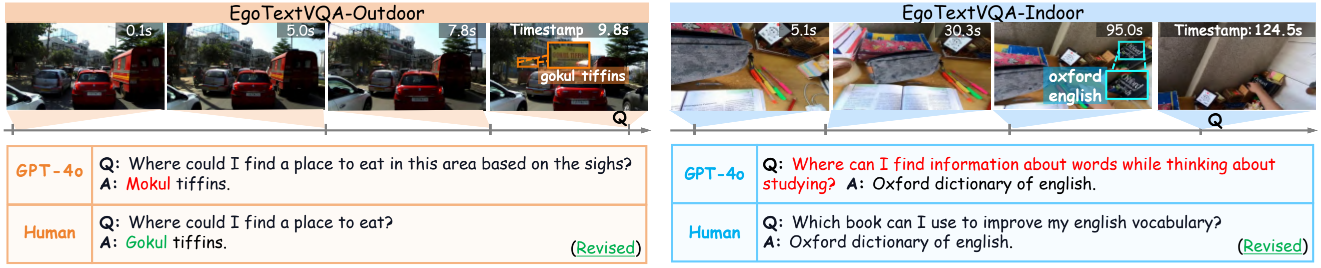
Human The question cannot be answered based on the video. (Removed)

GPT-4o Q: Which items need to be rearranged in the cabinet to make more space for the new dishware set? A: Plastic containers on the top shelf.

Human The QA pair does not involve scene text understanding. (Removed)

(a) Hallucination case.

(b) Scene text irrelevance case.



GPT-4o Q: Where could I find a place to eat in this area based on the sighs? A: Mokul tiffins.

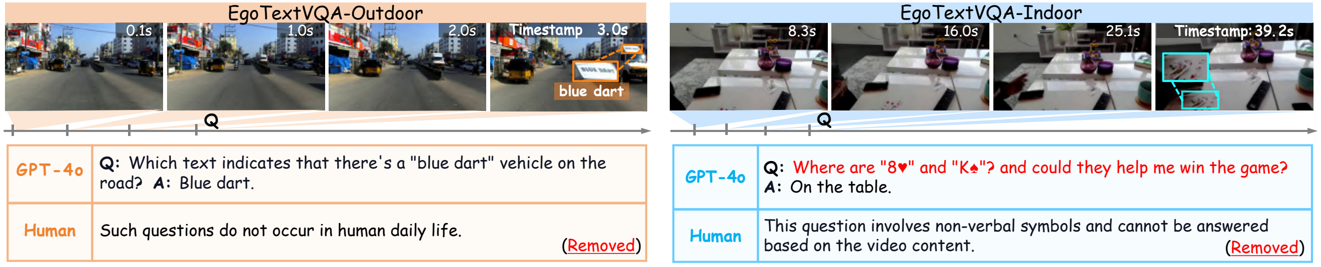
Human Q: Where could I find a place to eat? A: Gokul tiffins. (Revised)

GPT-4o Q: Where can I find information about words while thinking about studying? A: Oxford dictionary of english.

Human Q: Which book can I use to improve my english vocabulary? A: Oxford dictionary of english. (Revised)

(c) Scene text recognition error case.

(d) Non-colloquial case.



GPT-4o Q: Which text indicates that there's a "blue dart" vehicle on the road? A: Blue dart.

Human Such questions do not occur in human daily life. (Removed)

GPT-4o Q: Where are "8♥" and "K♠"? and could they help me win the game? A: On the table.

Human This question involves non-verbal symbols and cannot be answered based on the video content. (Removed)

(e) Not reflect user need case.

(f) Non-verbal special characters case.

Figure 8. Manual participation on EgoTextVQA creation.

tions for various videos. ShareCaptioner-Video is fine-tuned with the collected video caption data. For video understanding, ShareGPT4Video's training dataset combines VQA samples from various instructional video-to-text datasets with video-caption pairs.

B.2. Study of MLLM Design

We analyze key factors contributing to the superior performance of strong models (Qwen2-VL [50], LLaVA-NeXT-Video [64], and InternVL2 [8]): (1) **enhanced visual encoder** capable of handling high-resolution and long-video inputs. In Table 8, increasing the number of video frames and resolution improves Qwen2-VL's performance by 1.2% and 5.8%; (2) **more powerful LLM backbones**. Compared with InternVL2-8B, InternVL2-26B performance has a 7% increases in Table 8; and (3) **large-scale OCR training data**. Beyond the commonly used TextVQA datasets, InternVL2 leverages PaddleOCR to generate OCR samples for training. Additionally, we observe that as the number of video frames increases, Qwen2-VL's performance improves, whereas InternVL2's declines, underscoring the effectiveness of Qwen2-VL's video embedding design.

B.3. Heuristic Solution Investigations

Effect of Timestamp-Aware Sampling We further investigate an alternative question-timestamp aware sampling strategy: starting from the question timestamp and uniformly sampling within fixed durations of 4 seconds and 32 seconds. As shown in Table 9, on EgoTextVQA-Outdoor, when sampling the same number of frames (#F=16), this fixed duration sampling strategy achieves comparable or even superior performance to standard uniform sampling across the whole video for both Qwen2-VL [50] and Gemini 1.5 Pro [41]. However, on EgoTextVQA-Indoor, when sampling #F=48, we observe that while Qwen2-VL [50] maintains comparable performance to standard uniform sampling, the performance of Gemini 1.5 Pro [41] drops by by about 7%. This decline may stem from Gemini 1.5 Pro's stronger performance on questions requiring long-term video comprehension, which is less effectively captured by this fixed-duration sampling approach.

Combination of Heuristic Strategies In the main text, we have explored different heuristic strategies separately. Here, we additionally study the combinations of heuristic strate-

Table 8. Study of MLLM design on EgoTextVQA-Outdoor. The alignment module is MLP layer. VE: Visual Encoder.

Method	**VE**	**Res.**	**#F**	Accuracy	Score
Qwen2-VL [50]	ViT-675M	448^2	16	22.4	1.6
	ViT-675M	448^2	32	23.6	1.7
	ViT-675M	-	16	28.2	2.0
InternVL2-8B [8]	InternViT-300M	448^2	16	16.5	1.3
	InternViT-300M	448^2	32	16.4	1.3
InternVL2-26B [8]	InternViT-6B	448^2	16	23.5	1.7

Table 9. Effects of different numbers of video frames uniformly sampled within fix-duration before the question timestamp. We set a fixed duration of 4 seconds in EgoTextVQA-Outdoor and 32 seconds in EgoTextVQA-Indoor. S: Standard Uniformly Sampling. F: Fix-Duration Sampling. We experiment with 30% of the data for efficiency.

Method	**EgoTextVQA-Outdoor**			**EgoTextVQA-Indoor**		
	#F	Accuracy	Score	#F	Accuracy	Score
Qwen2-VL [50] w/ S	16	28.2	2.0	48	23.3	1.8
Qwen2-VL [50] w/ F	4	26.1	1.8	12	16.2	1.4
	8	29.3	2.0	24	18.3	1.5
	12	29.9	2.0	36	22.1	1.7
	16	30.9	2.1	48	22.7	1.7
Gemini 1.5 Pro [41] w/ S	32	33.4	2.0	60	34.4	2.1
Gemini 1.5 Pro [41] w/ F	4	27.9	1.7	12	22.8	1.6
	8	32.9	1.9	24	27.1	1.7
	12	33.5	2.0	36	29.0	1.8
	16	33.4	2.0	48	27.3	1.8

gies. First, for timestamp-aware video sampling, we adopt the strategy of "*fixed-duration sampling*" to EgoTextVQA-Outdoor and "*1fps-backward sampling*" to EgoTextVQA-Indoor, inspired by the results of the two different video sampling strategies on these two datasets. The results in Table 10 show that the models achieve cumulative performance improvements as heuristic strategies are progressively applied on EgoTextVQA-Outdoor. Yet, the improvements are not stable on EgoTextVQA-Indoor, suggesting the significant challenge of egocentric scene-text aware QA assistance in daily house-keeping.

B.4. Human Study

In Section 4, the human results are based on two rounds of standard human studies. Based on the reason analysis for the poor human performance in Section 4.2, we further validate the human performance by reducing the scene text recognition challenge. We sample 100 additional questions for humans to answer by providing the corresponding question frames. Table 11 shows that humans perform better without the challenge of temporal grounding but still lag behind the best closed-source model (GPT-4o [37]). This suggests advanced models may surpass humans in scene-text recognition or external knowledge, highlighting the importance of research on scene-text QA assistance.

Table 10. Effects of combining different heuristic strategies. T: Timestamp-Aware Sampling. ST: Additional Scene Text Input. HR: High-Resolution Scene Text (Scale = 1.25×). We experiment with 30% of the data for efficiency.

Method	**Input**			**EgoTextVQA-Outdoor**		**EgoTextVQA-Indoor**	
	T	ST	HR	Accuracy	Score	Accuracy	Score
Qwen2-VL [50]	-	-	-	28.2	2.0	23.3	1.8
	✓	-	-	30.9	2.1	22.6	1.7
	✓	✓	-	42.4	2.7	25.3	1.8
	✓	✓	✓	42.6	2.7	27.1	1.9
Gemini 1.5 Pro [41]	-	-	-	33.4	2.0	34.4	2.1
	✓	-	-	34.7	2.0	31.1	2.0
	✓	✓	-	49.5	2.9	**38.0**	**2.3**
	✓	✓	✓	**51.1**	**3.0**	36.5	2.2

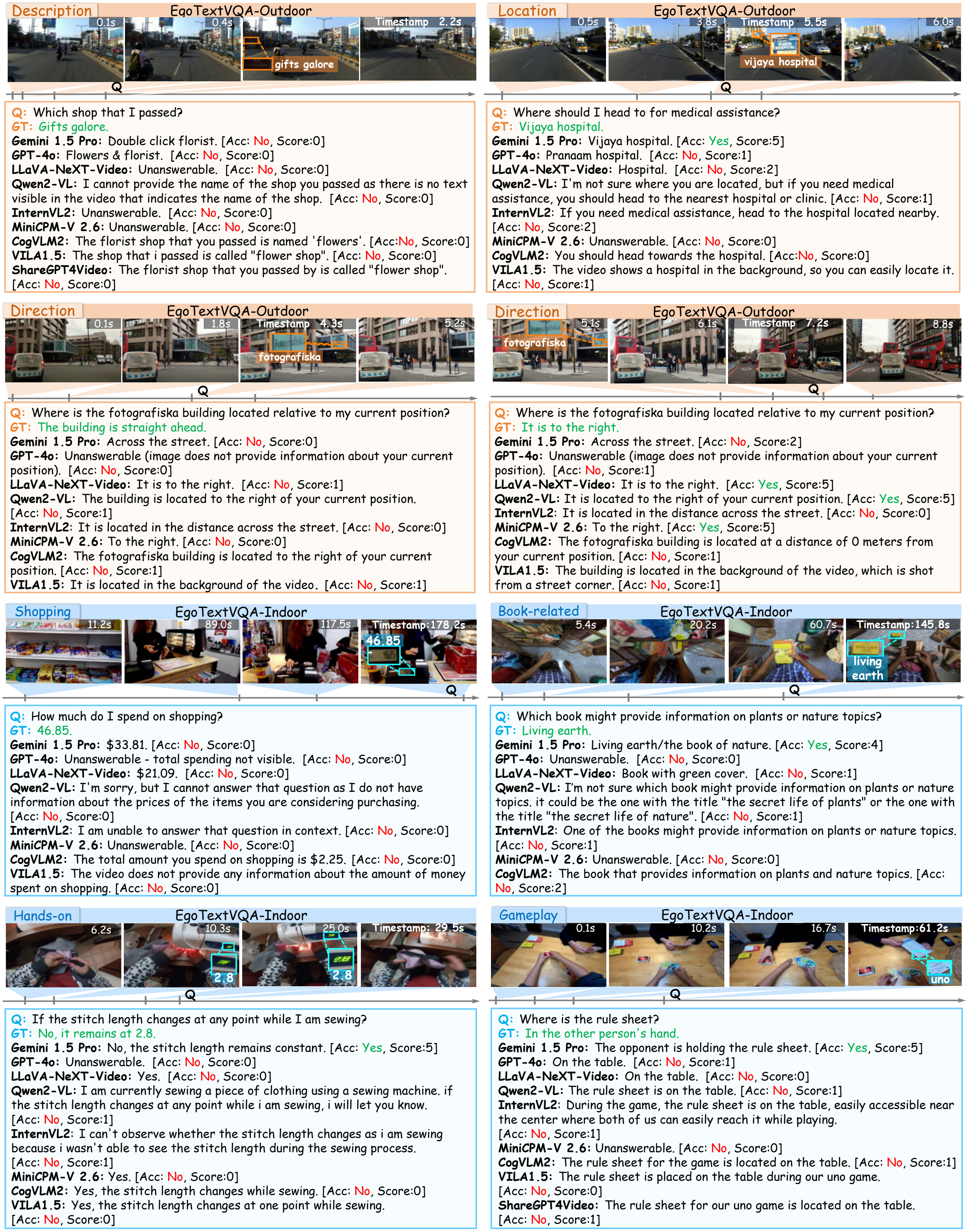
Table 11. Results of using video *vs.* QA frames (three frames for QA generation) on EgoTextVQA-Indoor.

Method	**Video**		**QA Frames**	
	Accuracy	Score	Accuracy	Score
Human	26.0	1.9	36.0	2.3
GPT-4o [27]	25.0	1.6	**39.0**	**2.4**
Gemini-1.5 Pro [29]	**33.0**	**2.0**	35.0	2.2

B.5. Case Analysis

As shown in Figure 9, we qualitatively analyze the performance of MLLMs on EgoTextVQA. For **EgoTextVQA-Outdoor**, the "Description" example shows that all models struggle to accurately identify the target referred to by the question at the queried timestamp. The "Location" example shows that only Gemini 1.5 Pro [41] correctly inferred the intention of the question and provided the precise answer. For the real-time "Direction" examples, where the same question is posed at different timestamps, which corresponds to different answers, the left example shows that the questioned building is located right in front of the user at the question timestamp (4.3s), but all models fail to provide the correct answer, likely due to poor 3D spatial relation reasoning. Also, if the user poses the same question at 7.2s when he has moved to the left of the building, all models are unresponsive to such visual changes and tend to keep their original answers. This indicates that the models struggle to provide reasonable answers based on the real-time visual context in dynamic environment.

For **EgoTextVQA-Indoor**, the "Shopping" example shows that all models fail to effectively answer the total expense of 46.85 after the checkout process, indicating their limitations in infer the total number after observing the price changes on the cashier's display during the checkout process. In the "Book" example, the models need to identify the book from a large collection of books that matches the user's needs (*e.g.*, related to topics on plants or nature). The wrong answers indicate that the corresponding models are either weak at scene-text recognition and knowledge reasoning or tend to hallucinate their responses to match some



Q: Which shop that I passed?
GT: Gifts galore.
Gemini 1.5 Pro: Double click florist. [Acc: No, Score:0]
GPT-4o: Flowers & florist. [Acc: No, Score:0]
LLaVA-NeXT-Video: Unanswerable. [Acc: No, Score:0]
Qwen2-VL: I cannot provide the name of the shop you passed as there is no text visible in the video that indicates the name of the shop. [Acc: No, Score:0]
InternVL2: Unanswerable. [Acc: No, Score:0]
MiniCPM-V 2.6: Unanswerable. [Acc: No, Score:0]
CogVLM2: The florist shop that you passed is named 'flowers'. [Acc:No, Score:0]
VILA1.5: The shop that i passed is called "flower shop". [Acc: No, Score:0]
ShareGPT4Video: The florist shop that you passed by is called "flower shop". [Acc: No, Score:0]

Q: Where should I head to for medical assistance?
GT: Vijaya hospital.
Gemini 1.5 Pro: Vijaya hospital. [Acc: Yes, Score:5]
GPT-4o: Pranaam hospital. [Acc: No, Score:1]
LLaVA-NeXT-Video: Hospital. [Acc: No, Score:2]
Qwen2-VL: I'm not sure where you are located, but if you need medical assistance, you should head to the nearest hospital or clinic. [Acc: No, Score:1]
InternVL2: If you need medical assistance, head to the hospital located nearby. [Acc: No, Score:2]
MiniCPM-V 2.6: Unanswerable. [Acc: No, Score:0]
CogVLM2: You should head towards the hospital. [Acc:No, Score:0]
VILA1.5: The video shows a hospital in the background, so you can easily locate it. [Acc: No, Score:1]

Q: Where is the fotografiska building located relative to my current position?
GT: The building is straight ahead.
Gemini 1.5 Pro: Across the street. [Acc: No, Score:0]
GPT-4o: Unanswerable (image does not provide information about your current position). [Acc: No, Score:0]
LLaVA-NeXT-Video: It is to the right. [Acc: No, Score:1]
Qwen2-VL: The building is located to the right of your current position. [Acc: No, Score:1]
InternVL2: It is located in the distance across the street. [Acc: No, Score:0]
MiniCPM-V 2.6: To the right. [Acc: No, Score:0]
CogVLM2: The fotografiska building is located to the right of your current position. [Acc: No, Score:1]
VILA1.5: It is located in the background of the video. [Acc: No, Score:1]

Q: Where is the fotografiska building located relative to my current position?
GT: It is to the right.
Gemini 1.5 Pro: Across the street. [Acc: No, Score:2]
GPT-4o: Unanswerable (image does not provide information about your current position). [Acc: No, Score:1]
LLaVA-NeXT-Video: It is to the right. [Acc: Yes, Score:5]
Qwen2-VL: It is located to the right of your current position. [Acc: Yes, Score:5]
InternVL2: It is located in the distance across the street. [Acc: No, Score:0]
MiniCPM-V 2.6: To the right. [Acc: Yes, Score:5]
CogVLM2: The fotografiska building is located at a distance of 0 meters from your current position. [Acc: No, Score:1]
VILA1.5: The building is located in the background of the video, which is shot from a street corner. [Acc: No, Score:1]

Q: How much do I spend on shopping?
GT: 46.85.
Gemini 1.5 Pro: $33.81. [Acc: No, Score:0]
GPT-4o: Unanswerable - total spending not visible. [Acc: No, Score:0]
LLaVA-NeXT-Video: $21.09. [Acc: No, Score:0]
Qwen2-VL: I'm sorry, but I cannot answer that question as I do not have information about the prices of the items you are considering purchasing. [Acc: No, Score:0]
InternVL2: I am unable to answer that question in context. [Acc: No, Score:0]
MiniCPM-V 2.6: Unanswerable. [Acc: No, Score:0]
CogVLM2: The total amount you spend on shopping is $2.25. [Acc: No, Score:0]
VILA1.5: The video does not provide any information about the amount of money spent on shopping. [Acc: No, Score:0]

Q: Which book might provide information on plants or nature topics?
GT: Living earth.
Gemini 1.5 Pro: Living earth/the book of nature. [Acc: Yes, Score:4]
GPT-4o: Unanswerable. [Acc: No, Score:0]
LLaVA-NeXT-Video: Book with green cover. [Acc: No, Score:1]
Qwen2-VL: I'm not sure which book might provide information on plants or nature topics. it could be the one with the title "the secret life of plants" or the one with the title "the secret life of nature". [Acc: No, Score:1]
InternVL2: One of the books might provide information on plants or nature topics. [Acc: No, Score:1]
MiniCPM-V 2.6: Unanswerable. [Acc: No, Score:0]
CogVLM2: The book that provides information on plants and nature topics. [Acc: No, Score:2]

Q: If the stitch length changes at any point while I am sewing?
GT: No, it remains at 2.8.
Gemini 1.5 Pro: No, the stitch length remains constant. [Acc: Yes, Score:5]
GPT-4o: Unanswerable. [Acc: No, Score:0]
LLaVA-NeXT-Video: Yes. [Acc: No, Score:0]
Qwen2-VL: I am currently sewing a piece of clothing using a sewing machine. if the stitch length changes at any point while i am sewing, i will let you know. [Acc: No, Score:1]
InternVL2: I can't observe whether the stitch length changes as i am sewing because i wasn't able to see the stitch length during the sewing process. [Acc: No, Score:1]
MiniCPM-V 2.6: Yes. [Acc: No, Score:0]
CogVLM2: Yes, the stitch length changes while sewing. [Acc: No, Score:0]
VILA1.5: Yes, the stitch length changes at one point while sewing. [Acc: No, Score:0]

Q: Where is the rule sheet?
GT: In the other person's hand.
Gemini 1.5 Pro: The opponent is holding the rule sheet. [Acc: Yes, Score:5]
GPT-4o: On the table. [Acc: No, Score:1]
LLaVA-NeXT-Video: On the table. [Acc: No, Score:0]
Qwen2-VL: The rule sheet is on the table. [Acc: No, Score:1]
InternVL2: During the game, the rule sheet is on the table, easily accessible near the center where both of us can easily reach it while playing. [Acc: No, Score:1]
MiniCPM-V 2.6: Unanswerable. [Acc: No, Score:0]
CogVLM2: The rule sheet for the game is located on the table. [Acc: No, Score:1]
VILA1.5: The rule sheet is placed on the table during our uno game. [Acc: No, Score:0]
ShareGPT4Video: The rule sheet for our uno game is located on the table. [Acc: No, Score:1]

Figure 9. Result visualization on EgoTextVQA.

question key words. Similar issues are also observed in the "Hands-on" example. Finally, the failures in the "Gameplay" example suggest that most models are weak in reasoning the real-time object state and people's real-time actions from an ego point of view. For example, while the "rule sheet" is on the table most of the time, it is on the other game player's hand at the time of user questioning.

C. Agreement between Human and Evaluator

In this section, we evaluate the performance of models on EgoTextVQA-Outdoor using GPT-4o mini [38] and human annotators. Following [10], we invite three annotators to assess GPT-4o [37] and Gemini 1.5 Pro [41], the overall best-performing model. Human annotators maintain the same scoring principle as the model, as shown in Table 16. We randomly sample 100 QA pairs for evaluation. As shown in Table 12, GPT-4o mini and human annotators achieve similar Accuracy and Score, with Pearson correlation coefficients of 0.80 and 0.87, respectively, indicating strong consistency. The Cohen's Kappa coefficients among three volunteers are 0.77 on Accuracy, indicating a high human agreement. To ensure reproducibility, future evaluations should use the same model version (GPT-4o-mini-2024-07-18) and the prompt in Table 16.

Table 12. Judgments of human and GPT-4o mini.

Method	**GPT-4o [37]**		**Gemini 1.5 Pro [41]**	
	Accuracy	Score	Accuracy	Score
Human	36.0	1.9	47.3	2.5
GPT-4o mini [38]	34.0	1.8	42.0	2.3

D. Model Prompts

Table 13 provides the prompts used by GPT-4o for question generation and filtering. Table 14 lists the prompts employed by GPT-4o for automatic question label annotation. Table 15 details the specific prompts applied for model inference. Table 16 shows the prompts used by GPT-4o-mini for model evaluation. Table 17 includes the prompts designed for heuristic solutions with different modality inputs.

Table 13. Prompts for question-answer generation and filtering on EgoTextVQA

Question-Answer Generation Prompts
Question Prompt: Give you a first-person perspective video, which records the scene you see from the first-person perspective. Please judge from your perspective whether there are scene texts in the video. If so, please tell me what these scene texts are. Then, you have some questions about the scene texts you see, and ask three questions related to the activities you are going to carry out. These scene texts can serve as clues to help you answer your questions. Please generate three highly diverse questions based on the scene texts related to your activities in the first-person perspective video. If there is no scene text in the video, it is not necessary. Your questions should meet the following requirements: Requirement 1: The questions should involve scene text understanding in the video. Requirement 2: The questions should be goal-oriented and relevant to human daily life. Requirement 3: The questions should require understanding multiple video frames, not just a single frame. Requirement 4: The questions should be asked from a first-person perspective, expressed as colloquially as possible, and the first-person pronoun “I” should be used appropriately. Requirement 5: The questions should be of moderate length. When announcing the question please label each question as “Question 1, 2, 3: {*question*}”. Please start your questions with the question word “what”, “where”, “which”, etc. You don’t need to explain too much about what you are doing or indicate the location of the scene text in the video. Avoid the words “video” and “frame” in the questions. Remember to make sure that the correct answer to your question can be taken directly from the video and is concise enough. Examples of good questions: “Question 1: Which way is the exit?” “Question 2: Could you tell me how much this item costs?” “Question 3: What is the speed limit on this road?” Image:{*image1*} Image:{*image2*} Image:{*image3*}
Answer Prompt: I provide three questions as follows: {*question*} You need to create an exam that tests above student abilities based on the three questions I just provided. Each question should have open-ended but short correct answers. Your answers have the following requirements: Requirement 1: Your answers should be short and be closely related to the scene text in the video. Requirement 2: Your answers should not mention any particular video frame number. Requirement 3: Do not use letters for the answer choices. You must print one correct answer and four wrong answers on separate lines in the following format: ¨ Correct Answer :{*answer*}
Automatic Filtering Prompt
You are a helpful assistant. You can answer the following questions based on your general knowledge. Question: {*question*} Answer briefly with a single word, a phrase, or a short sentence.

Table 14. Prompts for GPT-4o to annotate question categories on EgoTextVQA-Outdoor.

Question Classification Prompts
Question: {*question*} Which of the following five categories does this question belong to? Please only answer the category name, such as Direction. 1. **Location**: Questions about a place or location. For example: 1) Where is the gas station? 2) Which stores can I find on the right side of the road at this intersection? 2. **Direction**: Questions related to navigation, driving direction, and turns. For example: 1) Is the next road a left or right turn? 2) If I want to go to Cava, on which side of the street should I look for it? 3) Where should trucks go according to the signs? 3. **Description**: Questions that focus on scene text such as road signs, price labels, and billboards. For example: 1) What does the sign on the side of the road say? 2) What is the name of the center on the left side of the road? 3) What is the name of the street to my right? 4. **Intention Reasoning**: Questions about behavioral activities involving drivers or passengers to solve personal needs. For example: 1) Where do I need to go to solve my financial problems? 2) Is there a place nearby where I can shop for appliances and electronics? 5. **Others**: Composite questions that involve multiple different or the same types of the above, such as asking about both description and location. For example: 1) What event is being advertised on the bus, and where is it taking place? 2) What is the contact number for the leadspace building, and what service might they provide?

Table 15. Prompts for MLLM inference on EgoTextVQA.

Model	**General Prompts**
GPT-4o	Based on the following images from a video, please briefly answer the following question with a single word, a phrase, or a short sentence. Question: {*question*}. Output the answer to the question in the following format: Answer: {*answer*}. If you cannot answer the question, please answer “Unanswerable” and briefly explain why you cannot answer.
Gemini 1.5 Flash	Based on the following images from a video, please briefly answer the following question with a single word, a phrase, or a short sentence. Question: {*question*}. Output the answer to the question in the following format: Answer: {*answer*}. If you cannot answer the question, please answer “Unanswerable” and briefly explain why you cannot answer.
Gemini 1.5 Pro	Based on the following images from a video, please briefly answer the following question with a single word, a phrase, or a short sentence. Question: {*question*}. Output the answer to the question in the following format: Answer: {*answer*}. If you cannot answer the question, please answer “Unanswerable” and briefly explain why you cannot answer.
LLaVA-Next-Video	Please answer the following questions related to this video. If you cannot answer the question, please answer “Unanswerable” and briefly explain why you cannot answer. Keep your answer as short as possible. Keep your answer as short as possible. Keep your answer as short as possible. Question: {*question*}
CogVLM2-Video	You are a person in the situation shown in the following consecutive images from a video. You can answer questions that humans ask to help them make decisions. Now you are observing your surroundings and answering questions based on the current situation. Understanding the scene text around you is important for answering questions. Answer the questions in the first-person perspective. If you cannot answer the question, please answer “Unanswerable” and briefly explain why you cannot answer. Question: {*question*}
InternVL2	You are a person in the situation shown in the following consecutive images from a video. You can answer questions that humans ask to help them make decisions. Now you are observing your surroundings and answering questions based on the current situation. Understanding the scene text around you is important for answering questions. Answer the questions in the first-person perspective. If you cannot answer the question, please answer ’Unanswerable’ and briefly explain why you cannot answer. Keep your answer as short as possible! Keep your answer as short as possible! Keep your answer as short as possible! Question: {*question*}
Qwen2-VL	You are a person in the situation shown in the following consecutive images from a video. You can answer questions that humans ask to help them make decisions. Now you are observing your surroundings and answering questions based on the current situation. Understanding the scene text around you is important for answering questions. Answer the questions in the first-person perspective. If you cannot answer the question, please answer ’Unanswerable’ and briefly explain why you cannot answer. Question: {*question*}
VILA1.5	You are a helpful language and vision assistant. You are able to understand the visual content that the user provides, and assist the user with a variety of tasks using natural language. Question: {*question*}
ShareGPT4Video	You are a person in the situation shown in the following consecutive images from a video. You can answer questions that humans ask to help them make decisions. Now you are observing your surroundings and answering questions based on the current situation. Understanding the scene text around you is important for answering questions. Answer the questions in the first-person perspective. If you cannot answer the question. please answer “Unanswerable” and briefly explain why you cannot answer. Question: {*question*}
MiniCPM-V 2.6	You are a person in the situation shown in the following consecutive images from a video. You can answer questions that humans ask to help them make decisions. Now you are observing your surroundings and answering questions based on the current situation. Understanding the scene text around you is important for answering questions. Answer the questions in the first-person perspective. If you cannot answer the question, please answer ’Unanswerable’ and briefly explain why you cannot answer. Keep your answer as short as possible! Keep your answer as short as possible! Keep your answer as short as possible! Question: {*question*}

Table 16. Prompts for GPT-4o-mini to evaluate MLLMs on EgoTextVQA.

Evaluation Prompts
You are an intelligent chatbot designed for evaluating the correctness of generative outputs for question-answer pairs. Your task is to compare the predicted answer with the correct answer and determine if they match meaningfully. Here’s how you can accomplish the task: ##INSTRUCTIONS: - Focus on the meaningful match between the predicted answer and the correct answer. Please note that not only matches of noun phrases between answers, but also matches of prepositional phrases. For example, “at the car wash on your right” does not exactly match “car wash”. “at the gas station beside the sign ’gas sale’” does not exactly match “gas station”” - Consider synonyms or paraphrases as valid matches. Note that the predicted answer must be consistent with the string type of the correct answer, which may include phone numbers, email addresses, numbers, dates, etc. For example, the string types “www.usps.com” and “visit their website” are inconsistent, the string types “9849041316” and “advertiser’s contact number” are inconsistent.” - Evaluate the correctness of the prediction compared to the answer.” Please evaluate the following video-based question-answer pair: Question: {*question*} Correct Answer: {*GT answer*} Predicted Answer: {*predicted answer*} Provide your $eval_{code}$ only as a yes/no and score where the score is an integer value between 0 and 5, with 5 indicating the highest meaningful match. Please generate the response in the form of a Python dictionary string with keys ’pred’ and ’score’, where the value of ’pred’ is a string of ’yes’ or ’no’ and the value of ’score’ is in INTEGER, not STRING. DO NOT PROVIDE ANY OTHER OUTPUT TEXT OR EXPLANATION. Only provide the Python dictionary string. For example, your response should look like this: {’pred’: ’yes’, ’score’: 5}, {’pred’: ’no’, ’score’: 1}.

Table 17. Prompts for heuristic solution study of different modality inputs on EgoTextVQA.

Model Input	Prompts
w/ Q	You are a helpful assistant. You can answer the following questions based on your general knowledge. Question: {*question*}
w/ Q & ST	You are a helpful assistant. You are provided with some important scene text information. You can answer the following questions based on your common sense or the scene text information I provide. Please answer as briefly as possible. Please note that this scenario text information is very important. You can find the scene text related to the question as the answer. Scene Text: {*OCR results*} Question: {*question*}
w/ V & Q & ST	You are a person in the situation shown in the following consecutive images from a video. You can answer questions that humans ask to help them make decisions. Now you are observing your surroundings and answering questions based on the current situation. I will provide you with the following scene text that may be included in each image. Understanding the scene text is important for answering questions. Answer the questions in the first-person perspective. If you cannot answer the question, please answer ’Unanswerable’ and briefly explain why you cannot answer. The scene texts in Frame 0 include: {*OCR results*}. The scene texts in Frame 1 include: {*OCR results*}. The scene texts in Frame 2 include: {*OCR results*}. The scene texts in Frame {*frame id*} include: {*OCR results*}. Question: {*question*}