EgoTextVQA: Towards Egocentric Scene-Text Aware Video Question Answering

Supplementary Material

A. EgoTextVQA Dataset

A.1. Manual Participation

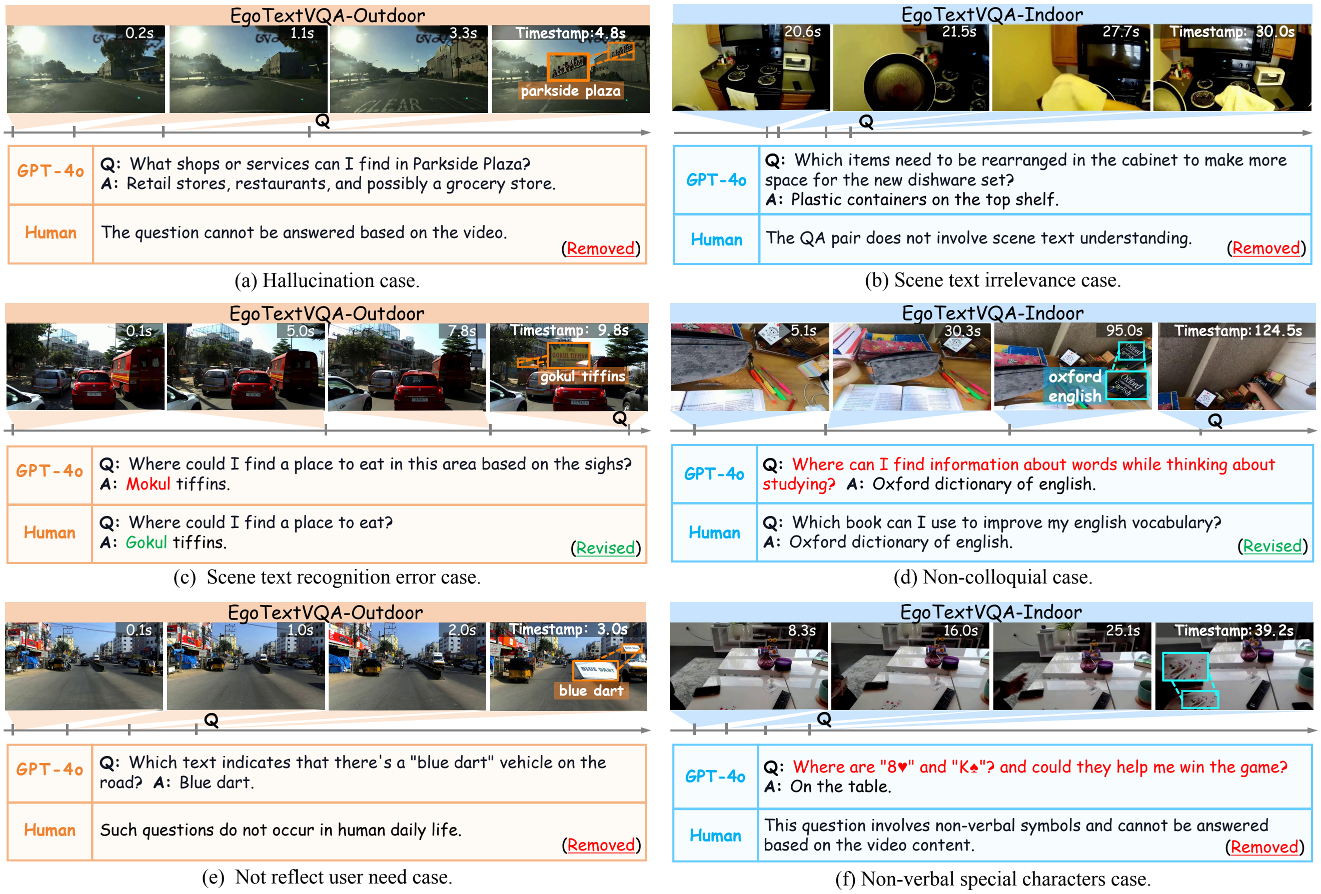
We present examples of QA pairs generated by GPT-4o in Figure 1 to highlight the issues of automatic generation and underscore the value of manual correction. The primary problems observed are as follows: **(a) Hallucinated Answers:** The generated answers are unseen from the visual environment and cannot 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** [15] advances the GPT-4 family towards more natural human-computer interactions.
- **Gemini 1.5 Pro** [18] builds on Gemini 1.0's [23] research advances and multimodal capabilities and it is optimized for a wide-range of reasoning tasks.
- **Gemini 1.5 Flash** [18] is a model from Gemini 1.5 family offering low latency and enhanced performance.
- **Qwen2-VL** [25] employs a ViT-675M [17] 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** [29] choose SigLIP-SO400M [28] as the visual encoder, Qwen2 [25] as the language model, and a two-layer MLP as the projector. It utilizes the AnyRes [13] 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** [12] integrates CLIP-L [17] as the visual encoder, LLaMA-2 [24] as the language model, and a linear projector. It fine-tunes on a mix of internal data, including OCR-VQA [14] and ST-VQA [2], 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** [6] integrates InternViT-300M [7] with InternLM2.5-7B [4] via a randomly initialized MLP projector. It is trained on OCR datasets generated by PaddleOCR [11], utilizing Chinese images from Wukong and English images from LaionCOCO [19]. Building on the strong visual representations and high-resolution image processing capabilities of InternVL1.5 [6], InternVL2 incorporates instruction tuning, enabling competitive performance in document and chart comprehension, infographics QA, scene text understanding, OCR, and multimodal reasoning tasks.
- **CogVLM2-Video** [9] utilizes the EVA-CLIP [22] as the visual encoder, LLaMA3-8B as the language model, and a 2×2 convolutional layer followed by a SwiGLU [20] as the adapter. Unlike CogVLM [26], 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 [11], (3) academic papers with extracted LaTeX via Nougat [3], 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** [27] employs SigLIP-SO400M [28] as the visual encoder, Qwen2 [25] 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 [21], OCR-VQA [14], and ST-VQA [2]. 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** [5] builts on LLaVA-Next-8B [10]. Based on the proposed ShareGPT4Video [5] dataset, the proposed captioning model ShareCaptioner-Video generates high-quality captions with detailed temporal descrip-



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

The question cannot be answered based on the video.

(a) Hallucination case.

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.

The QA pair does not involve scene text understanding.

(b) Scene text irrelevance case.

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

Q: Where could I find a place to eat?
A: Gokul tiffins.

(c) Scene text recognition error case.

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

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

(d) Non-colloquial case.

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

Such questions do not occur in human daily life.

(e) Not reflect user need case.

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

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

(f) Non-verbal special characters case.

Figure 1. 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 [25], LLaVA-NeXT-Video [29], and InternVL2 [7]): (1) **enhanced visual encoder** capable of handling high-resolution and long-video inputs. In Table 1, 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 1; 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 2, 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 [25] and Gemini 1.5 Pro [18]. However, on EgoTextVQA-Indoor, when sampling #F=48, we observe that while Qwen2-VL [25] maintains comparable performance to standard uniform sampling, the performance of Gemini 1.5 Pro [18] 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 1. Study of MLLM design on EgoTextVQA-Outdoor. The alignment module is MLP layer. VE: Visual Encoder.

Method	**VE**	**Res.**	**#F**	Accuracy	Score
Qwen2-VL [25]	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 [7]	InternViT-300M	448^2	16	16.5	1.3
	InternViT-300M	448^2	32	16.4	1.3
InternVL2-26B [7]	InternViT-6B	448^2	16	23.5	1.7

Table 2. 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 [25] w/ S	16	28.2	2.0	48	23.3	1.8
Qwen2-VL [25] 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 [18] w/ S	32	33.4	2.0	60	34.4	2.1
Gemini 1.5 Pro [18] 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 3 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 4 shows that humans perform better without the challenge of temporal grounding but still lag behind the best closed-source model (GPT-4o [15]). 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 3. 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 [25]	-	-	-	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 [18]	-	-	-	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

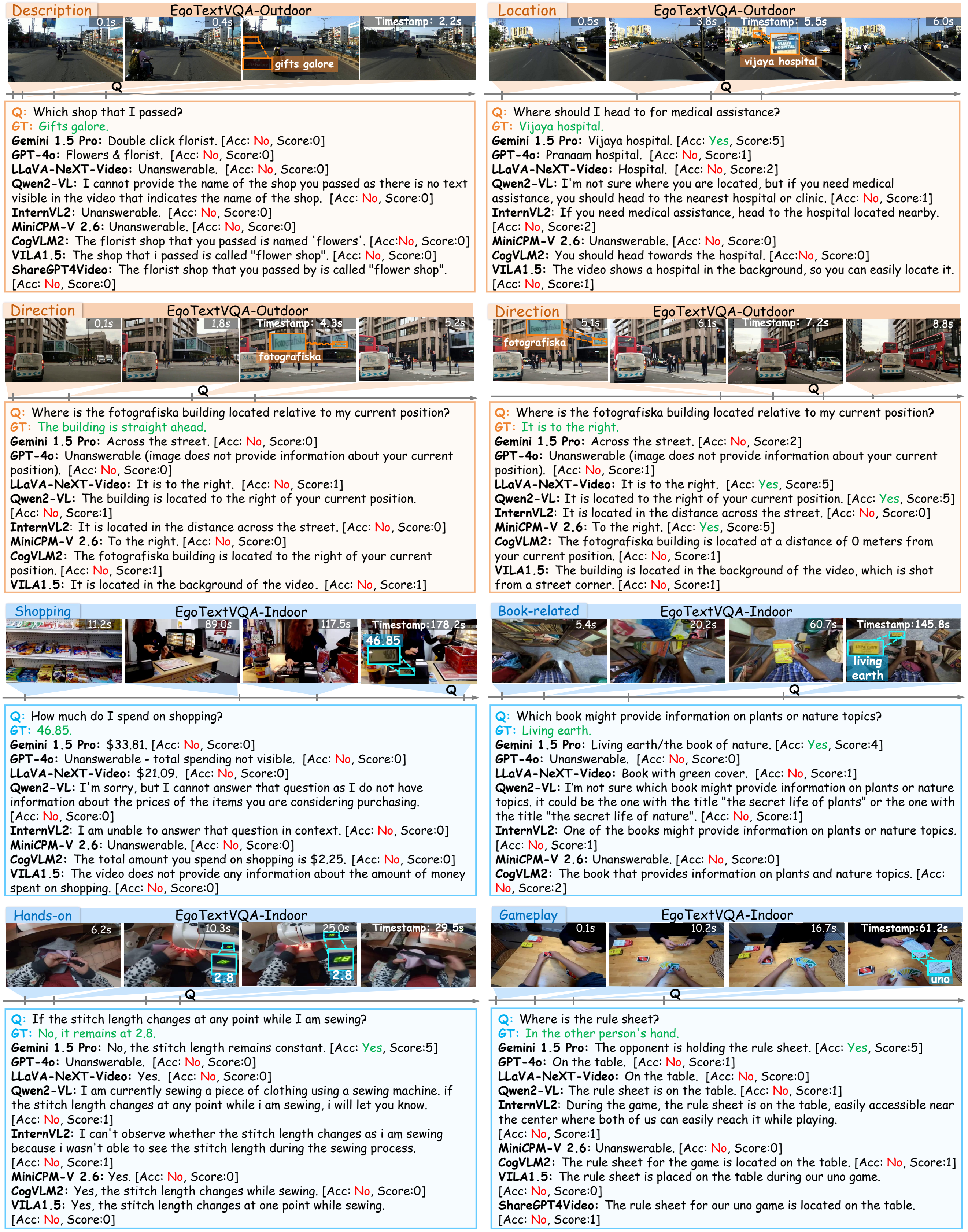
Table 4. Results of using video *vs.* QA frames (three frames for QA generation) on EgoTextVQA-Indoor.

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 2, 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 [18] 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 2. 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 [16] and human annotators. Following [8], we invite three annotators to assess GPT-4o [15] and Gemini 1.5 Pro [18], the overall best-performing model. Human annotators maintain the same scoring principle as the model, as shown in Table 9. We randomly sample 100 QA pairs for evaluation. As shown in Table 5, 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 9.

Table 5. Judgments of human and GPT-4o mini.

Method	**GPT-4o [15]**		**Gemini 1.5 Pro [18]**	
	Accuracy	Score	Accuracy	Score
Human	36.0	1.9	47.3	2.5
GPT-4o mini [16]	34.0	1.8	42.0	2.3

D. Model Prompts

Table 6 provides the prompts used by GPT-4o for question generation and filtering. Table 7 lists the prompts employed by GPT-4o for automatic question label annotation. Table 8 details the specific prompts applied for model inference. Table 9 shows the prompts used by GPT-4o-mini for model evaluation. Table 10 includes the prompts designed for heuristic solutions with different modality inputs.

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] 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. 12

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

[4] 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

[5] 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. 12

[6] 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. 12

[7] 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

[8] 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. 16

[9] 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. 12

[10] 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

[11] 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

[12] 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. 12

[13] 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

[14] 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. 12

[15] OpenAI. Gpt-4o system card. 2024. 12, 14, 16

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

[17] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry,

Table 6. 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 7. 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 8. 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 9. 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 10. 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*}

Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021. 12

[18] 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. 12, 13, 14, 16

[19] 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

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

[21] 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. 12

[22] 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

[23] 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

[24] 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

[25] 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. 12, 13, 14

[26] 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. 12

[27] 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. 12

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

[29] 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. 12, 13