Ego-Grounding for Personalized Question-Answering in Egocentric Videos

Supplementary Material

A. MyEgo Dataset

In this part, we will introduce additional details about the construction of MyEgo dataset, including the video source datasets and the multiple-choice generation pipeline. Our dataset and related resources are available at https://github.com/Ryougetsu3606/MyEgo.

A.1. Video Source Introduction

Ego4D [6] offers 3,670 hours of daily activity video spanning diverse scenarios: household, outdoor, workplace, and etc, which is captured by 931 unique camera wearers from 74 worldwide locations. We select 182 videos from QAEgo4D [3] subset, with each video involving two or more people engaging in the same scene or event.

EgoLife [14] consists of 44-hour egocentric videos recording a week of shared living experience of six young volunteers. To better raise questions related to multi-person scenarios, we further select clips containing rich group activities, including videos from all six volunteers on the same day (Day 3), as well as footage from a single volunteer (A4) spanning four days (Day 3 to Day 6). We then concatenate the sequential short clips into 257 longer ones, with each about 10 minutes.

Castle2024 [10] is a large-scale, multimodal dataset designed for advancing study in lifelogging, human activity analysis and multimodal retrieval. The entire set contains over 600 hours of videos from 15 time-aligned recorders, including 10 participants and 5 fixed cameras. We select a 30-hour subset of videos and segment them into clips of 6–20 minutes. We additionally filter out clips that lack multiple people or contain incomplete recording data, and eventually obtain 102 valid videos.

A.2. Multiple Choice Generation

Tab. 3 presents the prompt used for distractor generation. After obtaining the generated distractors, we first refine the options to produce concise expressions consistent with the style of the correct answers, preventing answer leakage due to formatting discrepancies. To further limit option bias, we manually revise QAs that both Gemini-2.5 Pro [5] and GPT-5 [8] can correctly answer even without access to the video or question. For human participation, we particularly ensure the negative options be misleading unless the models truly understand the video contents available at the question moments and correctly ground "me", "my things", and "my past" in the egocentric scene.

Table 1. Question categories and examples of MyEgo.

Category	Number/Ratio	Question Examples
Action	1,519/30.3%	What am I *holding*? Did I *give* her my power bank? Where did I *put* my pan?
Object	2,975/59.4%	Is this my rag? Where is my screwdriver? What color is my helmet?
Others	518/10.3%	Is this bowl the same as before? Did I get more than 200 scores?

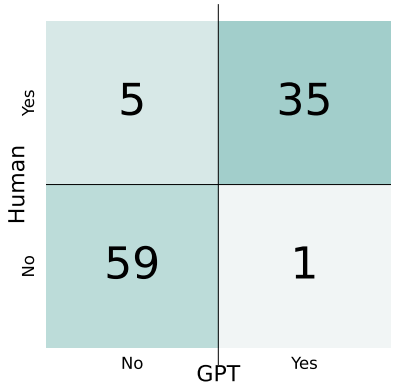


Figure 1. Agreement between human and GPT evaluation on 100 instances.

A.3. Question Categories

For better evaluation and analysis, we categorize the questions into 3 groups: 1) Action: questions focus on distinguishing my actions from those of other people nearby, 2) Object: questions focus on distinguishing my objects from other similar objects in the scene, and 3) Others: other questions that are not covered by the previous two categories. Examples of questions for each category are presented in Tab. 1. Additionally, we split the questions according to whether their answers are visible at the question moments: 1) Current: answer is visible at the question moment, and 2) Previous: answer is not visible at the question moment and demands retrieving past video contents.

B. Evaluation Details

Prompt Details. Tab. 4 shows the prompt we used to test the model performances. Following the official inference code, we add a thinking system prompt for InternVL3.5-Thinking series [13], and insert a time instruction right after the video frames for LLaVA-Video [19].

Table 2. Performances across all selected models (results based on 30% data). Dispider performs only MC-QA.

Methods	Multiple-Choice					Open-Ended					
	MC-2	MC-5	Cur.	Pre.	Avg.	Action	Object	Others	Cur.	Pre.	Avg.
Closed-source Models											
GPT-5 [8]	64.5	54.2	57.5	55.9	56.3	50.0	44.6	43.1	51.1	44.0	46.1
Gemini-2.5 Pro [5]	60.3	45.1	49.9	47.5	48.2	40.2	40.2	47.7	42.4	40.3	40.9
General Open-source Models											
Parameters > 10B											
Qwen3-VL-32B-Instruct [1]	58.3	35.4	44.1	38.4	40.0	41.4	36.4	37.8	40.0	37.3	38.1
Qwen3-VL-32B-Thinking [1]	55.2	37.9	44.3	40.0	41.3	40.1	35.5	35.8	40.6	35.4	36.9
InternVL3.5-38B-Thinking [13]	54.6	41.8	43.8	44.6	44.4	35.5	34.8	37.8	35.2	35.3	35.3
Qwen2.5-VL-32B-Instruct [2]	49.3	31.4	42.9	31.9	35.0	39.7	33.6	30.5	36.8	34.5	35.1
InternVL3.5-38B-Instruct [13]	56.3	39.0	41.9	42.7	42.5	35.0	34.1	37.8	34.3	34.9	34.7
InternVL3-38B [20]	51.3	41.9	44.1	43.7	43.8	35.7	32.5	39.1	33.3	34.5	34.1
4B <Parameters≤ 10B											
Qwen2-VL-7B-Instruct [12]	54.3	29.6	36.5	33.9	34.6	34.6	36.3	35.1	35.8	35.6	35.7
Qwen3-VL-8B-Instruct [20]	55.7	35.5	44.4	37.7	39.6	37.0	35.6	28.5	36.3	34.9	35.3
InternVL3-8B [20]	53.6	38.8	41.0	42.1	41.8	31.9	36.3	36.4	35.6	34.8	35.0
Qwen2.5-VL-7B-Instruct [2]	53.3	32.5	43.6	34.0	36.7	34.6	35.0	31.1	37.7	33.2	34.5
LLaVA-Video [19]	52.3	33.8	42.7	35.5	37.5	33.9	33.9	39.7	36.3	33.7	34.5
Qwen3-VL-8B-Thinking [1]	55.3	31.2	40.8	34.2	36.0	34.1	34.1	31.8	38.4	32.0	33.9
MiniCPM-V 4.5 [15]	50.3	35.8	41.2	37.8	38.7	35.2	32.2	37.1	34.3	33.3	33.6
LongVU [11]	56.0	31.9	40.3	35.3	36.7	31.9	33.1	36.4	30.8	34.0	33.1
InternVL3.5-8B-Thinking [13]	52.0	36.3	42.7	38.2	39.5	31.1	33.9	34.4	33.8	32.8	33.1
LLaVA-OV-7B [7]	49.7	32.7	40.8	34.3	36.1	31.1	32.4	38.4	33.8	32.1	32.6
LongVA [18]	56.3	28.6	39.6	32.1	34.2	33.9	31.0	29.8	34.7	30.5	31.7
InternVL3.5-8B-Instruct [13]	55.6	35.5	41.2	38.9	39.6	32.4	26.8	27.2	31.5	27.3	28.5
InternVL2.5-8B [4]	50.3	35.3	41.2	37.2	38.3	27.5	21.9	19.2	27.2	21.8	23.3
Parameters≤ 4B											
Qwen3-VL-4B-Instruct [1]	55.7	35.0	43.6	37.4	39.2	36.1	35.6	35.8	37.4	35.1	35.8
Qwen3-VL-4B-Thinking [1]	55.3	29.6	39.1	33.1	34.7	37.0	33.5	29.8	37.2	33.0	34.2
InternVL3.5-4B-Thinking [20]	48.3	36.8	41.7	38.1	39.1	32.8	31.0	35.8	32.4	31.8	32.0
InternVL3.5-4B-Instruct [20]	49.0	31.1	36.7	34.0	34.7	28.6	30.7	30.5	30.6	29.9	30.1
Qwen2.5-VL-3B-Instruct [2]	54.0	28.7	38.8	31.8	33.8	29.3	29.8	30.5	28.1	30.4	29.7
Memory-Enhanced Streaming QA Methods											
Flash-VStream (Qwen2-VL-7B) [16]	51.3	31.1	36.3	34.8	35.2	31.9	29.9	35.1	31.5	30.8	31.0
Dispider (Qwen2-VL-7B) [9]	45.5	31.2	31.9	33.1	32.7	-	-	-	-	-	-

Automatic Evaluation. We feed the prompt in Tab. 5 to GPT-5 mini [8] to evaluate whether the responses from a tested model (*e.g.*, Gemini 2.5 Pro) match the ground-truth answers. After two rounds of manual review and refinement of the evaluation prompt, we achieve an agreement rate of 94% with human judgment on 100 instances (Fig. 1).

C. More Experiment Analyses

Tab. 2 reports the performance of all selected models on MyEgo. For efficiency, we evaluate on a randomly sampled 30% subset (1,500 instances). We further include two recent memory models Dispider [9] and Flash-VStream [16] that are specialized for long streaming VideoQA for comparison.

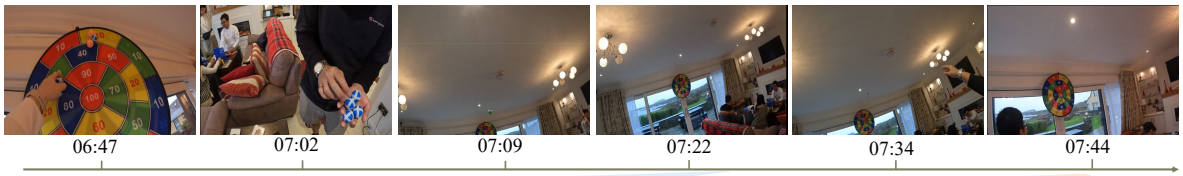
By comparing among models of different parameter scales (*e.g.*, 4B $\rightarrow$ 8B $\rightarrow$ 32B $\rightarrow$ 38B), we find that larger models do not consistently outperform the smaller ones. Even within the same model family, smaller variants (*e.g.*, 4B) can achieve competitive or even superior performance, as seen in the InternVL3 series. Additionally, the two long Video-LLMs, LongVA [17] and LongVU [11], do not exhibit outstanding results despite processing substantially more frames (128 *vs*. 32 of other general models). Similarly, the two long streaming QA methods also fail to achieve their desired level of performance. Surprisingly, the rela-

[Open-Ended] What color is my bike?
GT: Blue.

Qwen2.5-VL-7B: Your bike is black. **[Acc: No]**
LLaVA-Video: The bike is blue. **[Acc: Yes]**
Qwen3-VL-8B-Thinking: Got it … In the frames, the bike has a blue color. So the answer is blue.</think>Blue. **[Acc: Yes]**
InternVL3.5-8B-Instruct: blue **[Acc: Yes]**
LongVU: Your bike is black. **[Acc: No]**
Qwen3-VL-8B-Instruct: You bike is blue. **[Acc: Yes]**
MiniCPM-V 4.5: Blue. **[Acc: Yes]**

[Open-Ended] What color is my helmet?
GT: Black.

Qwen2.5-VL-7B: The helmet is blue. **[Acc: No]**
LLaVA-Video: Black. **[Acc: Yes]**
Qwen3-VL-8B-Thinking: Got it…a blue helmet attached to the bike frame. … So the helmet is blue…</think>Blue. **[Acc: No]**
LLaVA-OV: Red **[Acc: No]**
LongVU: Your helmet is black. **[Acc: Yes]**
Qwen3-VL-8B-Instruct: You helmet is black. **[Acc: Yes]**
MiniCPM-V 4.5: Black. **[Acc: Yes]**



[Open-Ended] Did I get more than 200 scores?
GT: Yes. You have one 100, one 20 and two 60.

GPT-5: No. Only three balls totaling about 110, not 200+. **[Acc: No]**
LLaVA-Video: Yes, you got more than 200 scores. **[Acc: Yes]**
Qwen3-VL-8B-Thinking: … The wheel has …the maximum is 100…</think>No, … max 100 score, not 200+. **[Acc: No]**
LLaVA-OV: Yes. **[Acc: Yes]**
LongVU: You did achieve more than 200 scores. **[Acc: Yes]**
Qwen3-VL-8B-Instruct: No, the highest score is 100. **[Acc: No]**
MiniCPM-V 4.5: No, you did not get more than 200. **[Acc: No]**

[Open-Ended] Who won the Dart Ball game?
GT: You. Because the total score of your balls is the highest

Qwen2.5-VL-7B: The person in the black shirt. **[Acc: No]**
LLaVA-Video: The person in the white shirt. **[Acc: No]**
Qwen3-VL-8B-Thinking: … there's a scene with a wheel … no specific info on Dart Ball… </think> not shown. **[Acc: No]**
LLaVA-OV: The person in the blue hoodie **[Acc: No]]**
LongVU: The person wearing a black shirt. **[Acc: No]**
Qwen3-VL-8B-Instruct: Not shown. **[Acc: No]**
MiniCPM-V 4.5: Not shown. **[Acc: No]**

[Open-Ended] Where is my screwdriver?
GT: On the table.

Qwen2.5-VL-7B : On the workbench behind you. **[Acc: No]**
LLaVA-Video: The screwdriver is on the workbench. **[Acc: No]**
Qwen3-VL-8B-Thinking: … using tools.…</think>Yes, you hold it in your right hand near the handlebars. **[Acc: No]**
LLaVA-OV: The screwdriver is on the workbench. **[Acc: No]**
LongVU: Your screwdriver can be found on the table. **[Acc: Yes]**
Qwen3-VL-8B-Instruct: On the workbench, near the tools. **[Acc: No]**
MiniCPM-V 4.5: On the workbench. **[Acc: No]**

[Multiple-Choice] What am I holding?
Options:
A. Nothing **(GT)**
B. Not shown
C. A bicycle tire.
D. A cassette.
E. A wrench.

Qwen3-VL-8B-Thinking, Gemini-2.5 Pro: **E**
Qwen3-VL-8B-Instruct, LLaVA-OV, InternVL3, GPT-5: **C**

Figure 2. Result visualization on MyEgo. Models struggle to correctly answer even simple questions, indicating their severe deficiency in performing ego-grounding in videos to answer personalized questions from the camera wearers. **Top:** The camera wearer is cycling alongside another man whose bike and helmet differ in color from his. **Middle:** The camera wearer is playing a dart-ball game with others and achieves the highest score. **Bottom:** The camera wearer is repairing a bicycle with another man, repeatedly putting down and picking up his screwdriver.

Table 3. Prompts for Gemini-2.5 Pro to generate distracting answers in multi-choice QA.

You are an expert in egocentric video analysis and a creative multiple-choice question designer. Your task is to generate a complete set of multiple-choice options for each question, which includes refining the correct answer and creating four plausible distractors.
IMPORTANT INSTRUCTIONS:
1. Analyze the Video: Carefully analyze the provided video. The timestamps "question_moment" and "answer_moment" are critical.
2. Create Multiple-Choice Options: For each question, you will generate a single list of options. The **very first option** in your list MUST be the refined correct answer, followed by exactly four distractors.
3. Generate Distractors: Following the refined answer, create four distractors. Each should be formatted as '"distractor (Criterion)"' and ideally meet one of the following criteria:

Criterion A (Contextual Confusion): The item is present and visible during the 'answer_moment' but is not the correct answer.
Criterion B (Environmental Objects): The item is present in the video but is not being used or worn by the person recording.
Criterion C (Temporal Misdirection): The item appears *after* the 'question_moment' but is not the correct answer for that specific question.
Criterion D (Attentional Decoy): The item is present and visible during the 'question_moment' to distract the viewer's attention.

Additionally, the phrasing of the distractors should be as similar as possible to the correct answer.
4. Fallback Rule: If a distractor cannot meet the above criteria, it must at least be an object or action present in the video and be a plausible, yet incorrect, answer.
5. Special Case for "Yes/No": If the answer is purely "yes" or "no" (i.e., without any additional explanation), you MUST return a list with **only two options**: the correct answer formatted as "Yes (Refined Answer)" or "No (Refined Answer)", and the opposite option formatted as "No (Distractor)" or "Yes (Distractor)". Otherwise, you should return distractor answers with explanation.
6. Input Format: You will receive the list of question-answer pairs in the following JSON format:
question: {QUESTION}
answer: {ANSWER}
question_moment: {QUESTION_MOMENT}
answer_moment: {ANSWER_MOMENT}
7. REQUIRED Output Format: You MUST return your response as a single, valid JSON array (a list of lists). The first item in each inner list must be the refined answer.
Example Output:

```
[
  "ground truth (Refined Answer)",
  "distractor A (Contextual Confusion)",
  "distractor B (Environmental Objects)",
  "distractor C (Temporal Misdirection)",
  "distractor D (Attentional Decoy)"
]
```

Please now generate the refined answers and distractors for the provided video and QA pairs.

tively old model Qwen2-VL leads the models of 7B sizes in open-ended QA, surpassing many recent systems. The above analyses collectively suggest that the explored problem of ego-grounding and personalized understanding, despite with significant practical value for personal assistance, is largely overlooked in existing technique evolution, and highlight the importance of our benchmark and analyses towards advancements in these fields.

By further analyzing performance across different question categories. We find that most models perform better on "Action" (*vs*. "Object") questions. This indicates that identifying "what I am doing" is easier than disambiguat-

Table 4. Prompts for models to answer questions.

Task	General Prompt
Open-ended	You are a first-person AI assistant integrated into a head-mounted camera. Your primary mission is to answer questions from the user (the camera wearer) about their own actions, objects, and environment as seen through your lens. You should answer directly to the user. The question is asked at the moment of the last frame in the video. Key Instructions: 1. First-Person Context: All references to 'I', 'me', or 'my' are about the user. You must distinguish between the user's actions/objects and those of other people visible in the video. 2. Grounding: Base your answers primarily on the visual evidence in the video. If the information is not present or cannot be reasonably inferred from the video, state that it is not shown. 3. For any question that requires a 'Yes' or 'No' response, you MUST follow it with a brief explanation for your reasoning. 4. All responses must be direct and clear. {VIDEO_CONTENT}. Question: {QUESTION}.
Multiple-choice	You are a first-person AI assistant integrated into a head-mounted camera. Your primary mission is to answer questions from the user (the camera wearer) about their own actions, objects, and environment as seen through your lens. You should answer directly to the user. The question is asked at the moment of the last frame in the video. Key Instructions: 1. All references to 'I', 'me', or 'my' are about the user. You must distinguish between the user's actions/objects and those of other people visible in the video. 2. Base your answers primarily on the visual evidence in the video. {VIDEO_CONTENT}. Question: {QUESTION}. Options: {2/5 OPTIONS}. There is only one correct option. Please only response with the letter of the correct option.
Removed personalized cues	You are a first-person AI assistant integrated into a head-mounted camera. Your primary mission is to answer questions from the user (the camera wearer) about their own actions, objects, and environment. You should answer directly to the user. The question is asked at the moment of the last frame in the video. Key Instructions: 1. Grounding: Base your answers primarily on the visual evidence in the video. If the information is not present or cannot be reasonably inferred from the video, state that it is not shown. 2. For any question that requires a 'Yes' or 'No' response, you MUST follow it with a brief explanation for your reasoning. 3. All responses must be direct and clear. {VIDEO_CONTENT}. Question: {QUESTION}.

Table 5. Prompts for GPT-5 mini to evaluate open-ended answers.

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 meaningful matches: Assess whether the predicted answer and the correct answer have a meaningful match, not just literal word-for-word matches.
- Criteria for Correctness: The predicted answer is considered correct if it reasonably matches the standard answer, recognizing that synonyms or varied expressions that convey the same or similar meaning are acceptable.
- If the predicted answer's yes/no conclusion conflicts with the correct answer, it is incorrect.
- The predicted answer is considered correct if it contains the core descriptive information and does not contradict the correct answer, even if some non-critical details are missing.
- Flexibility in Evaluation: Use judgment to decide if variations in the predicted answer still correctly address the question, even if they do not directly replicate the correct answer's phrasing.

Please evaluate the following question-answer pair:

Question: {QUESTION} Correct Answer: {GT-ANSWER} Predicted Answer: {MODEL-RESPONSE}

Provide your evaluation result 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 valid JSON string with keys 'pred' and 'score'.

For example: {"pred": "yes", "score": 5}, {"pred": "no", "score": 1}.

ing "my thing" from those of others in egocentric videos. The findings may depart from our common understanding about video object and action recognition. We visualize some examples in Fig. 2 for better understanding of results. Moreover, by comparing between "Current" and "Previous" questions, we observe a clear performance gap of almost all models: They perform better in answering questions whose answers are visible at the question moments. This suggests that the models are capable of better ground the camera wears and their belongings in the question moments, but such strength drops sharply when the answers are out of scene.

References

[1] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025. 2

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 2

[3] Leonard Bärmann and Alex Waibel. Where did i leave my keys? — episodic-memory-based question answering on egocentric videos. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1559–1567, 2022. 1

[4] 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. *Science China Information Sciences*, 67(12):220101, 2024. 2

[5] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 1, 2

[6] 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

[7] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 2

[8] OpenAI. Gpt-5 system card. *https://cdn.openai.com/gpt-5-system-card.pdf*, 2025. 1, 2

[9] Rui Qian, Shuangrui Ding, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Dahua Lin, and Jiaqi Wang. Dispider: Enabling video llms with active real-time interaction via disentangled perception, decision, and reaction. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24045–24055, 2025. 2

[10] Luca Rossetto, Werner Bailer, Duc-Tien Dang-Nguyen, Graham Healy, Björn Þór Jónsson, Onanong Kongmeesub, Hoang-Bao Le, Stevan Rudinac, Klaus Schöffmann, Florian Spiess, et al. The castle 2024 dataset: Advancing the art of multimodal understanding. *arXiv preprint arXiv:2503.17116*, 2025. 1

[11] Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, et al. Longvu: Spatiotemporal adaptive compression for long video-language understanding. In *Forty-second International Conference on Machine Learning*, 2025. 2

[12] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution, 2024. 2

[13] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025. 1, 2

[14] Jingkang Yang, Shuai Liu, Hongming Guo, Yuhao Dong, Xiamengwei Zhang, Sicheng Zhang, Pengyun Wang, Zitang Zhou, Binzhu Xie, Ziyue Wang, et al. Egolife: Towards egocentric life assistant. In *CVPR*, pages 28885–28900, 2025. 1

[15] Tianyu Yu, Zefan Wang, Chongyi Wang, Fuwei Huang, Wenshuo Ma, Zhihui He, Tianchi Cai, Weize Chen, Yuxiang Huang, Yuanqian Zhao, et al. Minicpm-v 4.5: Cooking efficient mllms via architecture, data, and training recipe. *arXiv preprint arXiv:2509.18154*, 2025. 2

[16] Haoji Zhang, Yiqin Wang, Yansong Tang, Yong Liu, Jiashi Feng, and Xiaojie Jin. Flash-vstream: Efficient real-time understanding for long video streams. *ICCV*, 2025. 2

[17] Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*, 2024. 2

[18] Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*, 2024. 2

[19] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024. 1, 2

[20] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025. 2