

LifeEval: A Multimodal Benchmark for Assistive AI in Egocentric Daily Life Tasks

Supplementary Material

1. Data Curation Pipeline

Our complete data curation pipeline consists of four stages: (a) MCQ generation via direct video prompting, (b) iterative quality filtering, (c) controlled difficulty enhancement, and (d) open-ended question reformulation. At every stage, we employ Gemini 2.5 Pro [2] as the automatic question generator and rewriter. Each stage adopts a dedicated prompting scheme, further tailored to individual capability dimensions to enforce targeted reasoning and consistent task grounding. Below, we outline the prompting strategies and examples for all four stages.

(a) MCQ Generation via Direct Video Prompting.

The model is prompted to analyze the video clip and produce an initial MCQ with reasoning evidence. Prompts focus on the user’s task context, the visual scene, and the intended human–assistant collaboration scenario, resulting in grounded and contextually appropriate questions. Below is an example prompt for *Dynamic Task Reasoning*.

Example Prompt for Stage (a)

You will be shown a first-person video in which the camera wearer is performing an everyday task. Based on the visible cues in the video, design 3–5 natural language questions that the user might realistically ask an intelligent assistant. Each question must include four answer options and one correct answer. These questions are intended to evaluate the model’s ability, as an intelligent assistant, to parse user needs in real time and assist with task completion through egocentric video understanding.

Guidelines:

1. Questions should simulate real, spontaneous queries that a user might ask while performing a daily task—seeking clarification, checking progress, or requesting help. Use first-person, casual, natural language, and avoid technical terminology.
2. The evaluation focus is on dynamic task reasoning: the ability to understand ongoing task progression in continuous video, including progress assessment, state tracking, short-term outcome prediction, and causal reasoning. Examples: “How much longer do these pancakes need before I can take them off?” “Based on my assembly progress, which part should I attach next?” “With the dough I have left, about how many more small buns can I make?”
3. Each question must be strictly grounded in events or actions shown in the video. The answer must depend on visual information.
4. Each question should require nontrivial reasoning, estimation,

tion, planning, diagnosis, or external knowledge that the user might not already know but could reasonably think through. Avoid low-level perception questions such as identifying colors or counting objects.

5. Each question must have one clear correct answer and three realistic distractors (ideally within 10 words each). Avoid vague or overly open-ended questions. Binary formats (yes/no, can/can’t, need/don’t need, will/won’t) are allowed but should not exceed 30% of all questions.
6. For each question, specify a start and end timestamp (in seconds) corresponding to the video segment that best supports both the question and the answer.
7. The answer must include: (i) the correct option (A/B/C/D) and (ii) a reasoning explanation describing how the answer is derived from the video.

Output Format:

Please output each question in the following JSON format:

```
{
  "question": "The question asked by the user (first-person, natural language)",
  "options": ["A. option text", "B. option text", "C. option text", "D. option text"],
  "answer": "Correct option (A/B/C/D)",
  "reasoning": "Detailed reasoning showing how the answer is derived from the video",
  "start_time_s": <start timestamp in seconds>,
  "end_time_s": <end timestamp in seconds >
}
```

(b) **Iterative Quality Filtering.** The video clip and the generated QA pair are jointly fed into the model to refine quality. Prompts at this stage guide the model to check for clarity, correctness, grounding in video context, and adherence to the assigned capability dimension. Only questions that pass this filtering step proceed to the next stage. Below is an example prompt for *Dynamic Task Reasoning*.

Example Prompt for Stage (b)

You are an expert in question quality review and refinement. You will receive a set of multiple-choice questions grounded in first-person videos. Each question includes a prompt, four answer options, the index of the correct option, the corresponding video time interval, and a reasoning explanation. Your task is to verify, correct, and enhance the input QA pairs to improve their quality. The input is provided in JSON format, and the output must strictly follow the same JSON structure.

General Rules:

1. The refinement focuses on verifying and improving the quality of the questions. Enhancements may include improving problem statements or increasing the plausibility of distractor options. Ensure the question has one and only one correct answer; avoid vague or ambiguous wording. Each option should ideally be no more than 10 words.
2. If two QA pairs in the input are highly similar, remove one of them. High-quality questions may be left unchanged. Low-difficulty or low-quality questions with unclear or incorrect answers must be revised, including modifying the prompt or distractor options.
3. Questions must retain natural first-person phrasing, simulating real users asking an intelligent assistant for immediate help or clarification while performing daily tasks. These questions evaluate a model’s ability to infer user intent from egocentric video and support task execution in real time.
4. The core purpose of the modified questions is to assess dynamic task reasoning: the ability to understand ongoing task progression in continuous video, including progress assessment, state tracking, short-term outcome prediction, and causal reasoning. Revised questions must remain within this evaluation scope. If the input question does not match this objective, redesign it accordingly.
5. Ensure that the question requires visual grounding in the video. If the question can be answered without watching the video, adjust and rewrite it.
6. Do not change the values of `start_time.s` or `end_time.s`. All refinements must correspond precisely to the visual content within the specified time interval.
7. Only output the refined content in JSON format, matching the input structure exactly. Do not output any additional text.

Input QA Pairs:

{ ... }

(c) Controlled Difficulty Enhancement. For the filtered MCQs, we apply prompts that instruct the model to increase question difficulty in a controlled manner. Additional constraints ensure that the revised questions remain grounded in the video, preserve the collaborative task context, and do not drift away from the designated capability dimension. Below is an example prompt for *Dynamic Task Reasoning*.

Example Prompt for Stage (c)

You are an expert in question review and refinement. You will receive a set of multiple-choice questions grounded in first-person videos. Each question includes a prompt, four answer options, the index of the correct option, the corresponding video time interval, and a reasoning explanation. Your task is to correct, revise, and enhance the input QA pairs to increase their difficulty. The input is provided in JSON format, and the output must strictly follow the same JSON structure.

General Rules:

1. The primary goal of the refinement is to increase question difficulty. You may achieve this by restructuring the question, creating more misleading distractors, designing more reasoning-intensive prompts, introducing distractors that are common-sense plausible but contradicted by video details, introducing distractors that match the video but violate common sense, or adding counter-intuitive elements so that the correct answer is not the most obvious choice.
2. Ensure that the correct option is unique and unambiguous. The answer must not be vague, and multiple correct answers are not allowed. Each option should ideally contain no more than 10 words.
3. Questions must retain natural first-person phrasing, simulating real users asking an intelligent assistant for immediate help or clarification while performing daily tasks. These questions evaluate a model’s ability to infer user intent from egocentric video and support task execution in real time.
4. The core purpose of the modified questions is to assess dynamic task reasoning: the ability to understand ongoing task progression in continuous video, including progress assessment, state tracking, short-term outcome prediction, and causal reasoning. Revised questions must remain within this evaluation scope. If an input question does not align with this dimension, redesign it accordingly.
5. Ensure that each question depends on the video content; it must not be answerable without watching the video.
6. Do not modify `start_time.s` or `end_time.s`. All revisions must remain consistent with the visual content within the specified timestamps.
7. Only output the refined content in JSON format, matching the input structure exactly. Do not include any additional text.

Input QA Pairs:

{ ... }

(d) Open-Ended Question Reformulation. Finally, the enhanced MCQs are reformulated into concise open-ended questions. The prompts ensure that the new OEQs shifts from recognition-based answering to generative reasoning, better reflecting real-world assistance scenarios. Below is an example prompt for *Dynamic Task Reasoning*.

Example Prompt for Stage (d)

You are an expert in question auditing and transformation. You will be given a set of multiple-choice questions grounded in first-person videos. Each question contains a prompt, four answer options, the index of the correct option, the corresponding video time interval, and a brief reasoning explanation. Your task is to convert each multiple-choice question into a short-answer open-ended question.

General Rules:

1. The answer must be a concise and clearly defined short

phrase or sentence, preferably within 10 words.

2. The answer must be semantically unique and not overly ambiguous. If uniqueness cannot be guaranteed or the question is not suitable for conversion, delete the question instead.
3. The converted question should maintain a reasonable level of difficulty. Avoid overly simple yes/no responses. If the question becomes too easy, rewrite it to introduce more reasoning. If the original question is already of high quality, it may be preserved with minimal changes.
4. Questions must retain natural first-person phrasing, simulating real users asking an intelligent assistant for immediate help or clarification while performing daily tasks. These questions evaluate a model’s ability to infer user intent from egocentric video and support task execution in real time.
5. The core purpose of the modified questions is to assess dynamic task reasoning: the ability to understand ongoing task progression in continuous video, including progress assessment, state tracking, short-term outcome prediction, and causal reasoning. Revised questions must remain within this evaluation scope.
6. Ensure that each question relies on the video content; it must not be answerable without watching the video. If a question can be answered without video context, revise it accordingly.
7. Do not modify `start_time_s` or `end_time_s`. All changes must remain consistent with the visual content within the specified timestamps.

Output Format:

Please output the transformed QA pairs in the following JSON format:

```
{
  "question": "the rewritten short-answer question",
  "answer": "the concise and unambiguous answer",
  "reasoning": "the reasoning process for obtaining the answer",
  "start_time_s": <start timestamp in seconds >,
  "end_time_s": <end timestamp in seconds >
}
```

2. Benchmark Statistics

To provide a clear organizational structure for analysis, we group the collected videos into five high-level task domains: *Cooking*, *Maintenance*, *Household*, *Handcraft*, and *Shopping*. These domains capture diverse daily activities and interaction contexts rather than narrowly defined visual scenes. The distribution of questions across capability dimensions and task domains is presented in Fig. 1. As shown, the number of questions is well balanced across domains, ensuring comprehensive and diverse coverage throughout the benchmark.

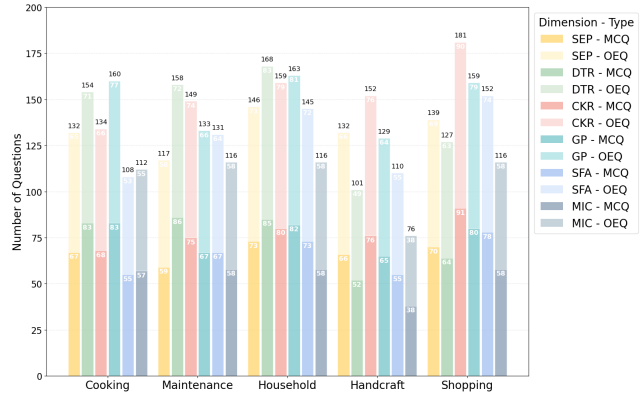


Figure 1. Distribution of question counts across task domains and capability dimensions.

3. More Experimental Results

Tab. 1 reports the evaluation results across different task domains. Overall, GPT-5 and Gemini 2.5 Pro continue to lead by a substantial margin. Among open-source models, Qwen3-VL-30B-A3B(thinking) and LLaVA-OneVision-72B achieve the strongest performance. Model performance remains relatively consistent across different task domains, with minor variations revealing interesting patterns. For multiple-choice questions, models achieve their highest scores in handcraft-related tasks, likely because these tasks involve more structured visual recognition. For open-ended questions, performance peaks in household-related domains, which appear to provide more practical and reasoning-friendly contexts for generative responses.

Fig. 2 further visualizes the performance distribution of representative MLLM series and families across task domains, capability dimensions, and question types. As also reflected in Fig. 2a, proprietary MLLMs exhibit the most uniform and comprehensive capability coverage, performing consistently well across all dimensions. In contrast, most open-source families, except the Qwen3-VL series in Fig. 2d, show a pronounced gap between MCQ and OEQ performance, indicating persistent limitations in generative and reasoning-heavy tasks. Across model families, the performance distribution is more balanced across task domains than across capability dimensions. This suggests that, while current MLLMs can generally handle a wide range of scenarios, they struggle disproportionately with certain cognitive capabilities required for effective assistance, particularly in dynamic reasoning and open-ended generation. Collectively, these findings highlight that developing balanced capabilities, rather than domain-specific proficiency, remains a key challenge for MLLMs aspiring to achieve truly collaborative intelligence.

Table 1. Main evaluation results of representative MLLMs on LifeEval across different task domains and question types. **Bold** and underline denote the best and second-best results, respectively.

Model	Multiple-Choice Questions						Open-Ended Questions					
	Cooking	Maintenance	Household	Handcraft	Shopping	Overall	Cooking	Maintenance	Household	Handcraft	Shopping	Overall
<i>Proprietary MLLMs</i>												
GPT-5 [6]	79.20	83.11	81.67	86.62	82.48	82.48	70.20	73.30	74.26	72.53	71.48	72.39
GPT-4o [3]	71.99	75.24	76.83	77.49	70.14	74.23	52.52	55.15	57.47	54.30	55.96	55.19
GPT-5-mini [6]	78.73	79.71	78.68	84.38	78.59	79.85	66.02	68.41	68.06	69.15	65.80	67.44
Gemini-2.5-Flash [2]	73.77	78.54	75.44	81.44	76.55	76.98	62.37	68.46	65.49	62.25	61.50	64.05
Gemini-2.5-Pro [2]	84.02	83.13	84.92	91.12	85.71	85.61	71.04	<u>68.55</u>	<u>69.42</u>	<u>70.48</u>	<u>67.44</u>	<u>69.32</u>
Grok-4 [10]	74.31	73.93	74.70	80.09	74.28	75.30	58.64	61.24	61.30	59.46	59.52	60.07
<i>Open-Source MLLMs</i>												
Qwen3-VL-30B-A3B [7]	68.23	70.03	69.48	78.33	65.33	70.01	<u>45.42</u>	46.19	52.73	48.85	<u>48.36</u>	<u>48.61</u>
Qwen3-VL-30B-A3B(thinking) [7]	69.21	<u>70.69</u>	69.75	76.68	70.14	<u>71.09</u>	47.72	48.79	<u>51.54</u>	48.92	48.93	49.24
Qwen3-VL-8B [7]	59.56	63.33	58.24	67.83	54.40	60.33	40.11	37.85	46.08	37.88	40.14	40.61
Qwen3-VL-8B(thinking) [7]	65.29	66.46	69.12	75.52	65.57	68.16	43.89	42.36	48.08	43.31	46.41	44.96
InternVL3.5-38B [8]	56.07	58.01	58.96	67.68	58.99	59.69	31.49	28.88	33.66	30.68	34.59	31.99
InternVL3.5-14B [8]	54.70	56.49	57.30	60.75	58.13	57.38	28.04	25.87	30.62	26.25	31.61	28.65
InternVL3.5-8B [8]	55.08	60.48	56.54	59.47	58.65	57.98	25.45	23.49	26.43	23.47	27.55	25.40
InternVL3.5-4B [8]	52.06	55.10	50.67	59.66	50.08	53.23	26.60	21.91	23.88	20.86	27.48	24.27
LLaVA-OneVision-72B [4]	<u>68.91</u>	72.33	74.63	<u>77.72</u>	<u>68.22</u>	72.19	40.15	36.69	39.68	38.24	42.22	39.48
LLaVA-OneVision-1.5-8B [1]	55.28	60.47	58.17	63.61	51.19	57.47	21.82	22.02	25.57	24.65	24.49	23.83
LLaVA-OneVision-7B [4]	52.72	58.13	55.69	58.05	56.58	56.17	26.11	27.94	31.82	29.48	33.02	29.81
mPLUG-Owl3-7B [11]	51.27	56.17	51.90	62.00	52.87	54.55	25.31	25.68	28.09	23.83	24.94	25.66
mPLUG-Owl3-2B [11]	39.73	42.25	43.77	42.71	37.85	41.22	12.28	10.83	10.18	12.71	13.07	11.78
LLaVA-NeXT-110B [5]	62.11	66.18	66.44	71.52	62.06	65.45	32.12	33.84	37.09	36.40	35.21	34.97
LLaVA-NeXT-72B [5]	64.31	63.77	<u>71.18</u>	69.13	63.10	66.26	32.34	37.34	43.13	37.90	39.54	38.24
LLaVA-NeXT-7B [5]	46.63	42.86	45.77	44.08	46.18	45.16	10.09	12.01	12.17	10.58	12.85	11.61
<i>Open-Source Video-Specialized LMs</i>												
VideoLLaMA3-7B [12]	53.41	54.89	54.90	58.29	49.98	54.13	24.74	22.79	26.17	20.83	23.32	23.69
LLaVA-Video-7B [14]	55.29	58.70	61.27	<u>61.25</u>	55.91	58.42	<u>28.03</u>	27.51	31.53	26.50	<u>28.82</u>	28.61
LongVA-7B [13]	52.54	60.36	<u>58.30</u>	64.11	51.38	<u>57.07</u>	24.42	25.37	<u>29.13</u>	28.35	<u>25.02</u>	26.47
InternVideo2.5_Chat_8B [9]	<u>54.60</u>	<u>59.75</u>	56.37	60.70	<u>54.33</u>	56.99	32.31	<u>26.30</u>	28.96	<u>27.29</u>	29.47	28.91

4. Failure Cases

We further present several representative failure cases to illustrate the typical error patterns exhibited by current MLLMs, as shown in Figs. 3 to 5. We observe that models are often influenced by its own prior knowledge rather than grounding its answers in the specific visual context of the task. For instance, in the last example of Fig. 3, the model relies on a prior assumption that pens are “waterproof” should not smear, while ignoring the visual evidence that the ink had not yet dried, ultimately leading to an incorrect choice. This tendency to over-rely on priors over grounded perception leads to systematic misjudgments, especially in visually subtle scenarios. For open-ended settings and interactive QA, models continue to struggle with precise, actionable, and contextually grounded assistance. Common errors include incomplete scene understanding, inaccurate or vague descriptions, and inability to reason about fine-grained visual details.

Notably, since open-ended questions require more flexible and generative reasoning, we adopt a more nuanced evaluation strategy to capture the quality of model responses. In particular, we employ GPT-5 [6] as the judge model to evaluate responses for open-ended questions. Our five-level scoring rubric provides a more nuanced assessment than binary correctness. As illustrated in Figs. 4 and 5, the rubric assigns intermediate scores for partially correct or partially grounded answers, allowing it to capture subtle

distinctions in reasoning quality. The resulting judgments are consistent and aligned with human expectations, offering a reliable measure of the degree of correctness. An example scoring prompt is provided below.

Example Scoring Prompt for Open-ended Questions

Please act as an impartial judge and evaluate the quality of the response provided by an AI assistant to the user question displayed below. Your evaluation should consider correctness and helpfulness. You will be given a reference answer and the assistant’s answer. Begin your evaluation by comparing the assistant’s answer with the reference answer. Identify and correct any mistakes. The assistant has access to video frames alongwith questions but you will not be given images. Therefore, please consider only how the answer is close to the reference answer. If the assistant’s answer is not exactly same as or similar to the answer, then he must be wrong. Be as objective as possible. Discourage uninformative answers. Also, equally treat short and long answers and focus on the correctness of answers. Ignore language differences and identical semantics count as a match. After evaluation, you must assign a rating chosen from {0, 0.25, 0.5, 0.75, 1}, where 1 indicates the highest level of meaningful match. The rating must be presented strictly in the format: “[rating]”, for example: “Rating: [[0.5]]”.

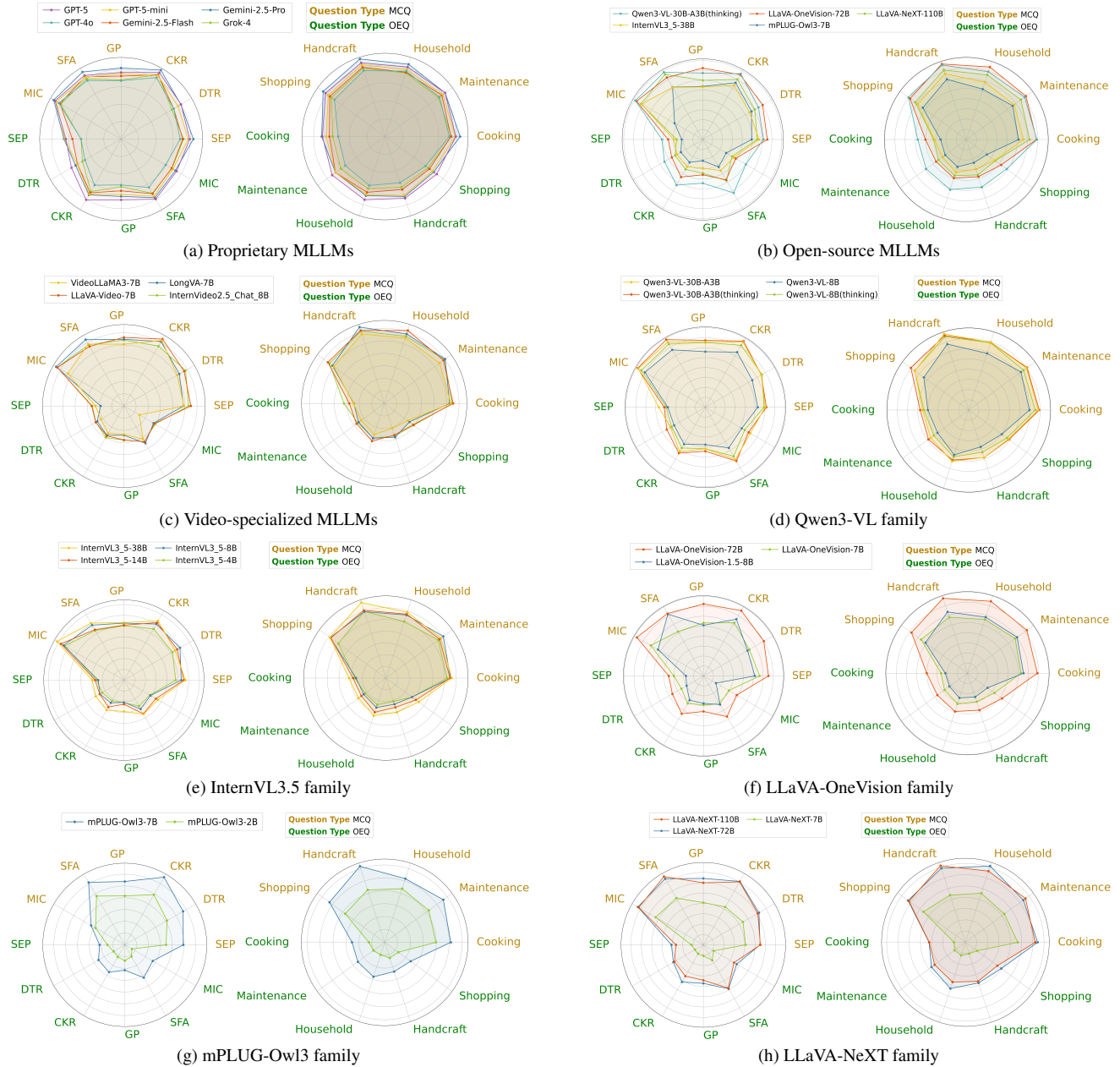


Figure 2. Performance comparison of diverse MLLM series and families on LifeEval.

5. Future Work

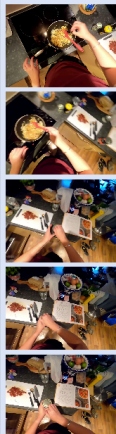
While LifeEval provides a comprehensive foundation for evaluating AI assistants, several promising directions remain for future expansion and refinement.

1. We plan to broaden the range of activities beyond common daily tasks to include expert-level and highly specialized domains. This would better reflect the diverse needs of real-world assistive scenarios and push models toward more generalized competency.
2. The current benchmark primarily focuses on video clips. Future efforts will extend toward multi-step task sequences spanning longer temporal contexts, enabling deeper evaluation of a model’s ability to offer coherent assistance throughout complex workflows.
3. We aim to include more sophisticated open-ended answers, such as multi-step plans, conditional instructions, and explanatory reasoning. This would challenge models to structure guidance in a clear, actionable, and context-aware manner.

Ultimately, we envision extending our framework toward interactive and real-time evaluation settings, where models must continuously perceive, reason, and respond as a task unfolds. Such a benchmark would more faithfully capture the demands of practical human–AI collaboration and offer a more powerful testbed for advancing next-generation MLLMs as capable real-world assistants.

References


- [1] Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Chunsheng Wu, Huajie Tan, Chunyuan Li, Jing Yang, Jie Yu, Xiyao Wang, Bin Qin, Yumeng Wang, Zizhen Yan, Ziyong Feng, Ziwei Liu, Bo Li, and Jiankang Deng. Llava-onevision-1.5: Fully open framework for democratized multimodal training. In *arxiv*, 2025. 4
- [2] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasapat, Naveen Sachdeva, Inderjit Dhillon, Marcel Blisstein, 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, 4
- [3] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 4
- [4] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 4
- [5] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023. 4
- [6] OpenAI. Gpt-5. <https://openai.com/index/introducing-gpt-5>, 2025. 4
- [7] Qwen Team. Qwen3 technical report, 2025. 4
- [8] 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. 4
- [9] Yi Wang, Xinhao Li, Ziang Yan, Yinan He, Jiashuo Yu, Xiangyu Zeng, Chenting Wang, Changlian Ma, Haiyan Huang, Jianfei Gao, Min Dou, Kai Chen, Wenhai Wang, Yu Qiao, Yali Wang, and Limin Wang. Internvideo2.5: Empowering video mllms with long and rich context modeling. *arXiv preprint arXiv:2501.12386*, 2025. 4
- [10] xAI. Grok 4. <https://x.ai/news/grok-4>, 2025. 4
- [11] Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl3: Towards long image-sequence understanding in multimodal large language models, 2024. 4
- [12] Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025. 4
- [13] 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. 4
- [14] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data, 2024. 4



Question: What are those red pieces on the cutting board?
Options: A. Sun-dried tomatoes.
 B. Roasted red bell peppers.
 C. Chopped Italian sausage.
 D. Fresh red chillies.
Answer: A

GPT-5: A	✓	LLaVA-NeXT-110B: C	✗
Gemini-2.5-Pro: A	✓	mPLUG-Owl3-7B: C	✗
Grok-4: C	✗	VideoLLaMA3-7B: C	✗
Qwen3-VL-30B-A3B-Thinking: C	✗	LLaVA-Video-7B: C	✗
InternVL3.5-38B: C	✗	LongVA-7B: C	✗
LLaVA-Onevision-72B: A	✓	InternVideo2.5-Chat-8B: C	✗

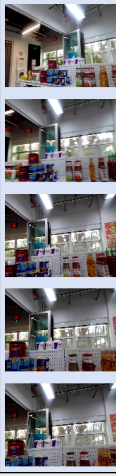
Capacity Dimension: *Static Environment Perception*
Reasoning: The chopped red ingredient on the cutting board is dark in color, irregular, and wrinkled—characteristics consistent with sun-dried tomatoes. It lacks the smooth surface of roasted peppers or the meaty texture of sausage.



Question: I just finished coating the meat. What's the next step?
Options: A. Heat the oil and fry it immediately.
 B. Place it on the nearby plate.
 C. Return it to the wet batter for another coat.
 D. Shake off excess flour and re-coat.
Answer: B

GPT-5: C	✗	LLaVA-NeXT-110B: D	✗
Gemini-2.5-Pro: C	✗	mPLUG-Owl3-7B: D	✗
Grok-4: D	✗	VideoLLaMA3-7B: D	✗
Qwen3-VL-30B-A3B-Thinking: B	✓	LLaVA-Video-7B: C	✗
InternVL3.5-38B: C	✗	LongVA-7B: D	✗
LLaVA-Onevision-72B: A	✗	InternVideo2.5-Chat-8B: D	✗

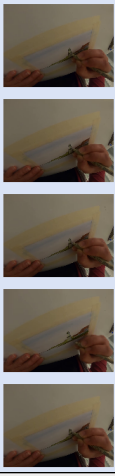
Capacity Dimension: *Dynamic Task Reasoning*
Reasoning: The cook just floured one piece. A plate on the right already holds several portions, indicating this is a batch-prep phase. The logical next step is to place the newly coated piece there and wait until all are ready before cooking.



Question: I want to make a quick hot appetizer with this bottle of HERDEZ salsa, what's the best way?
Options: A. Pour it over cheese and microwave.
 B. Fry it with ground beef for taco filling.
 C. Top baked-chicken nachos with it.
 D. Simmer it with beans to make spicy bean soup.
Answer: A

GPT-5: A	✓	LLaVA-NeXT-110B: B	✗
Gemini-2.5-Pro: A	✓	mPLUG-Owl3-7B: C	✗
Grok-4: A	✓	VideoLLaMA3-7B: C	✗
Qwen3-VL-30B-A3B-Thinking: C	✗	LLaVA-Video-7B: B	✗
InternVL3.5-38B: C	✗	LongVA-7B: B	✗
LLaVA-Onevision-72B: A	✓	InternVideo2.5-Chat-8B: A	✓

Capacity Dimension: *Contextual Knowledge Retrieval*
Reasoning: The request is for a 'quick' hot appetizer. Spooning the salsa over a meltable cheese like Velveeta and microwaving for 30–60 seconds yields classic Chile con Queso almost instantly. Options B, C, and D require cooking meat, baking chicken, or simmering—much slower. Thus, A is the fastest method.




Question: I just finished drawing these lines with this pen. Is it safe to apply a layer of watercolor on top right now?
Options: A. No, the ink will smudge if it isn't dry yet.
 B. Yes, the ink from this pen is waterproof.
 C. No, watercolor will dissolve the ink lines.
 D. Yes, as long as you brush the paint on gently.
Answer: A

GPT-5: B	✗	LLaVA-NeXT-110B: B	✗
Gemini-2.5-Pro: A	✓	mPLUG-Owl3-7B: A	✓
Grok-4: B	✗	VideoLLaMA3-7B: B	✗
Qwen3-VL-30B-A3B-Thinking: A	✓	LLaVA-Video-7B: A	✓
InternVL3.5-38B: B	✗	LongVA-7B: A	✓
LLaVA-Onevision-72B: B	✗	InternVideo2.5-Chat-8B: C	✗

Capacity Dimension: *Safety and Feasibility Assessment*
Reasoning: The video shows the user using a Sakura Pigma Micron fineliner. Although the ink becomes waterproof after drying, the question specifies applying watercolor "right now," meaning the ink is still wet. Adding watercolor at this moment would cause the ink to smudge and blur.

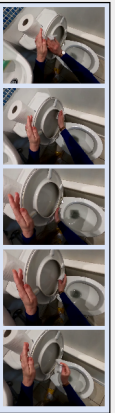
Figure 3. Examples of multiple-choice questions and model responses in LifeEval.



Question: I forgot the exact brand and model of this board. Can you check what's written on the box?
Answer: ASRock H81M-VG4 R3.0

GPT-5: H81M-VG4 R3.0	Score: 0.75
Gemini-2.5-Pro: H81M-VG4 R3.0	Score: 0.75
Grok-4: ASUS M5A99FX PRO R2.0	Score: 0
Qwen3-VL-30B-A3B-Thinking: H81M-VG4 R3.0	Score: 0.75
InternVL3.5-38B: ROG STRIX B450-F GAMING	Score: 0
LLaVA-Onevision-72B: ASUS ROG R3.0	Score: 0
LLaVA-NeXT-110B: H81M-UG4 R3.0	Score: 0.25
VideoLLaMA3-7B: ASRock Super Alloy	Score: 0

Capacity Dimension: *Static Environment Perception*
Reasoning: Between 0–3 seconds, a motherboard box is clearly visible on the desk. The side of the box displays large white text indicating the full brand and model: "ASRock H81M-VG4 R3.0."



Question: I'm holding the toilet lid and seat with one hand. How can I free it up to wipe more easily?
Answer: Lift both the lid and the seat fully back against the tank.

GPT-5: Lean the lid seat ring against the water tank.	Score: 1.0
Gemini-2.5-Pro: Turn the lid and the seat entirely.	Score: 1.0
Grok-4: Stand the toilet lid and the seat completely.	Score: 1.0
Qwen3-VL-30B-A3B-Thinking: Hold the tissue between your fingers.	Score: 0
InternVL3.5-38B: Wipe with a tissue.	Score: 0
LLaVA-Onevision-72B: Hold it up with your knees.	Score: 0
LLaVA-NeXT-110B: Step on the toilet seat with your foot.	Score: 0
VideoLLaMA3-7B: Place the tissue on the toilet seat.	Score: 0

Capacity Dimension: *Goal-oriented Planning*
Reasoning: In the video, the user is struggling to support both the toilet seat and lid with one hand, which limits their ability to clean effectively. The best solution is to lift the seat and lid entirely and let them rest stably against the tank, freeing both hands for easier and more thorough cleaning.

Figure 4. Examples of open-ended questions and model responses in LifeEval. Scores are evaluated by GPT-5.

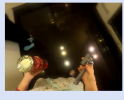


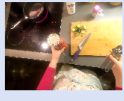
	Question1: Looking at the ingredients on my stovetop, what dish am I preparing? Answer1: Tomato chicken stew.			
	GPT-5: Tomato and Vegetable Soup. Score: 0.25	InternVL3.5-38B: Stewed meat. Score: 0.25		
	Gemini-2.5-Pro: Chicken Stewed with Tomatoes. Score: 1.0	LLaVA-Onevision-72B: Stir-frying dishes Score: 0		
	Grok-4: Provençal Vegetable stew. Score: 0	LLaVA-NeXT-110B: Scrambled eggs with tomatoes. Score: 0		
	Qwen3-VL-30B-A3B-Thinking: Make tomato pasta. Score: 0	VideoLLaMA3-7B: I'm preparing a dish. Score: 0		
	Question2: Okay. To keep the zucchini from becoming too mushy and retain some texture, when should I add it to the pot? Answer2: In the last 10–15 minutes of cooking.			
	GPT-5: Add it a few minutes before finished. Score: 0.25	InternVL3.5-38B: Before the stewed meat is well cooked. Score: 0.25		
	Gemini-2.5-Pro: Put it in when it's almost done. Score: 0.75	LLaVA-Onevision-72B: Put it in halfway through stewing. Score: 0.25		
	Grok-4: Put it in the pot in the last few minutes. Score: 0.5	LLaVA-NeXT-110B: Add it halfway through stewing. Score: 0.25		
	Qwen3-VL-30B-A3B-Thinking: Add it when it's almost done. Score: 0.75	VideoLLaMA3-7B: Add it after it boils. Score: 0		
	Question3: Good idea. Before adding the zucchini to the stew, what else can I do to help it stay firmer during cooking? Answer3: Fry the surface with a little oil first.			
	GPT-5: Sprinkle salt and squeeze out the water. Score: 0	InternVL3.5-38B: Blanching. Score: 0		
	Gemini-2.5-Pro: Salt it and squeeze out the water. Score: 0	LLaVA-Onevision-72B: Salt it and fry it slightly in oil. Score: 1.0		
	Grok-4: Salting and releasing the water. Score: 0	LLaVA-NeXT-110B: Slice first and then cook. Score: 0		
	Qwen3-VL-30B-A3B-Thinking: Salt it first. Score: 0	VideoLLaMA3-7B: Cut into small pieces. Score: 0		
	Capacity Dimension: Multi-turn Interactive Collaboration			
	Reasoning: The video shows chicken thighs being stewed together with zucchini and a can of tomato products—this combination strongly suggests a tomato-based chicken stew. To preserve the zucchini's crispness in a long-simmering dish, easily overcooked vegetables like zucchini should be added only near the end of cooking, typically in the final 10–15 minutes. Before adding it to the pot, lightly searing the cut surfaces in a small amount of oil helps create a protective layer that improves the zucchini's ability to keep its shape and texture during stewing.			

Figure 5. Examples of open-ended multi-turn interactive collaboration in LifeEval. Scores are evaluated by GPT-5.