BENCHMARKING EGOCENTRIC COOKING VIDEO UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent advances in multimodal large language models (MLLMs) have pushed video understanding toward real-world assistance. Egocentric cooking is a demanding testbed for this goal, requiring fine-grained hand-object recognition, evolving state tracking, and sequential reasoning, yet existing video QA benchmarks largely emphasize general-purpose understanding or assume access to complete videos rather than past-only observations. To fill this gap, we present EgoChef, a benchmark comprising 61,328 open-ended question-answer pairs from 129 real-world videos and covering current-state perception, temporal reasoning, and operation-level understanding. Each question is anchored to a timestamp, with models restricted to video observed up to that point and future frames withheld. For reliable open-ended evaluation, we provide multiple reference answers through multi-model semantic fusion with human calibration. Across representative open- and closed-source MLLMs, the strongest model achieves only 55.68% overall accuracy, with temporal reasoning remaining particularly challenging. We believe EgoChef can serve as a focused benchmark for fine-grained and temporal understanding of egocentric cooking.

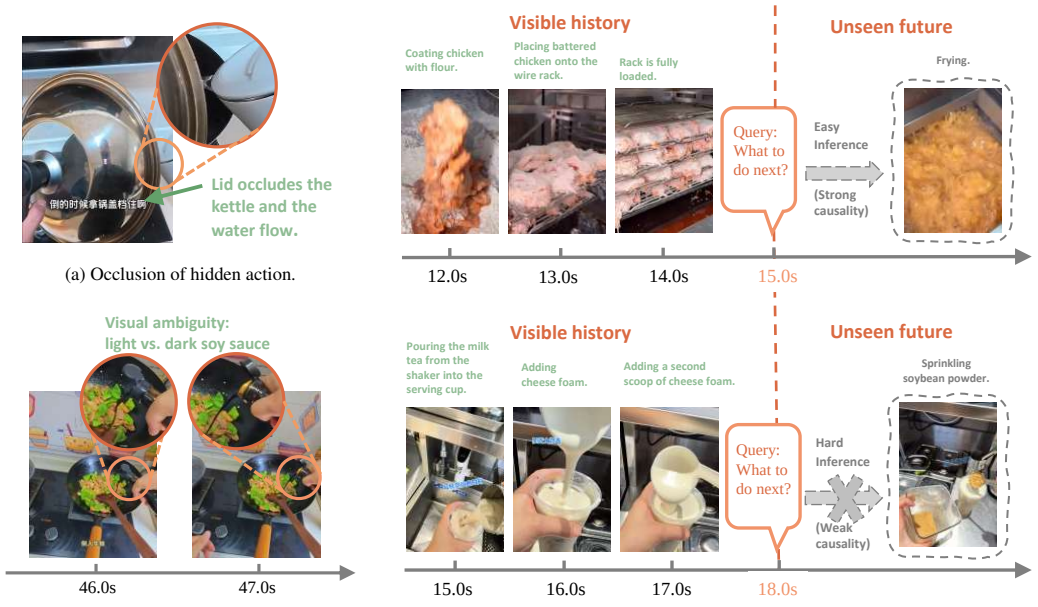
1 INTRODUCTION

Recent advances in multimodal large language models (MLLMs) (Alayrac et al., 2022; Li et al., 2023; Liu et al., 2023; Wang et al., 2024; OpenAI et al., 2024; Team et al., 2023) are pushing multimodal systems beyond offline content understanding toward interactive assistance in the physical world. Unlike an offline model that can inspect a complete event retrospectively, an online assistant must reason only from observations available so far, continuously identifying the current state, relating it to relevant history, and anticipating what may happen next. This requires more than recognizing visible content: the model must maintain and update a coherent procedural state from an evolving egocentric visual stream. We refer to this capability as *online egocentric procedural understanding*, which underlies household assistance (Li et al., 2026c), step-by-step instruction (Li et al., 2026a), and human-computer interaction (Li et al., 2026b).

Egocentric cooking provides a concentrated testbed for this capability (Perrett et al., 2025). Cooking unfolds through interdependent operations in which ingredients, tools, object states, and goals continually evolve, while first-person footage contains dense hand-object interaction, frequent occlusion, small manipulated objects, and subtle transient changes. Understanding such a process therefore goes beyond isolated action recognition: a model must perceive the current state, resolve fine-grained operational details, relate them to relevant prior steps, and anticipate plausible future operations. Egocentric cooking thus tightly couples *fine-grained perception, state tracking, temporal memory, and procedural anticipation* within a single evolving activity.

These capabilities become particularly difficult under partial observations and strict temporal constraints. As illustrated in Figure 1, critical operations may be hidden by occlusion, while fine-grained distinctions between ingredients or actions can be obscured by visual similarity, reflections, and motion blur. More fundamentally, procedural reasoning is temporally asymmetric: the present can be interpreted from observed history, whereas the future must be inferred without future visual evidence. Some continuations are strongly constrained by the observed procedure, such as frying after coating and rack placement, while others depend on weaker, recipe-specific cues, such as predicting soybean powder after cheese foam. The central difficulty is therefore not simply accumulating more

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) Occlusion of hidden action.

(b) Fine-grained action and object distinction.

(c) Strong and weak causality for predicting the unseen future.

Figure 1: Key challenges of online egocentric procedural understanding in EgoChef. (a) Critical actions may be hidden by egocentric occlusion. (b) Fine-grained ingredients and operations can be visually ambiguous. (c) Under past-only observation, future steps range from strongly constrained by the observed procedure to weakly constrained and knowledge-dependent.

history, but determining which past evidence remains relevant and how strongly it constrains the current and future procedural state.

Existing VideoQA benchmarks do not fully capture this setting. Conventional VideoQA often assumes access to complete videos (Xu et al., 2017), while many benchmarks focus on general-purpose understanding (Xiao et al., 2021), coarse-grained actions (Goyal et al., 2017), or third-person viewpoints. Recent egocentric benchmarks advance first-person understanding, but typically cover only subsets of the capabilities required for online procedural assistance: some emphasize perception without fine-grained operational reasoning, some evaluate procedural reasoning over complete videos, and others rely on multiple-choice questions (Mangalam et al., 2023). This distinction is substantive rather than cosmetic. Once post-query frames are available, next-step prediction can collapse into retrospective recognition. Strong performance on complete-video QA therefore does not necessarily indicate that a model can support an activity as it unfolds. What remains missing is an open-ended benchmark that jointly evaluates perception, fine-grained operations, temporal recall, and future anticipation under strict past-only access.

To fill this gap, we introduce EgoChef, an open-ended benchmark for online egocentric procedural understanding in real-world cooking videos. EgoChef contains 129 egocentric cooking videos and 61,328 question-answer pairs, with ten question types spanning three complementary dimensions: *current-state perception*, *temporal reasoning*, and *operation-level understanding*. For each question posed at timestamp t, the model can access only observations up to t, with all future frames withheld, preserving the information boundary faced by a real-time assistant. EgoChef further adopts open-ended answering and provides one to four reference answers through multi-model semantic fusion with human calibration, enabling reliable evaluation of semantically valid variations. Rather than probing isolated visual or temporal skills, EgoChef asks whether a model can continuously form, maintain, and update its understanding of a procedure as it evolves.

Our evaluation reveals limitations beyond the overall performance gap. The strongest evaluated MLLM reaches only 55.68% overall accuracy, with temporal reasoning remaining the most challenging dimension. More importantly, this difficulty cannot be explained simply by insufficient visual history. Multi-frame input generally outperforms a single current frame, confirming the value of temporal context, yet failure cases show that models often carry obsolete ingredients, tools, or actions from earlier steps into current predictions. Visual history is therefore both necessary and potentially distracting: an online model must retain what remains procedurally relevant while suppressing what has become obsolete. Domain-specific prompting and fine-tuning improve many cate-

gories, but do not consistently resolve fine-grained failures such as tool recognition. Taken together, these results suggest that the bottleneck is not seeing more frames, remembering longer contexts, or acquiring more domain knowledge, but *maintaining the right procedural state from an evolving visual history*. Online egocentric procedural understanding requires not only memory, but selective retention and timely updating; not only perception, but anticipation grounded in observed procedure.

Our contributions are threefold. (1) We introduce **EgoChef**, an open-ended benchmark for *online egocentric procedural understanding*, with 129 real-world cooking videos and 61,328 QA pairs spanning perception, temporal reasoning, and fine-grained operation understanding. (2) We establish a strict timestamp-constrained past-only protocol for real-time assistance, with multi-reference open-ended evaluation constructed through multi-model semantic fusion and human calibration. (3) We benchmark representative open- and closed-source MLLMs and show that current models struggle not only with temporal reasoning and anticipation, but also with maintaining and updating procedural state over evolving visual history.

2 Related Work

Cooking and Egocentric Video QA. First-person cooking videos capture fine-grained hand-object interactions, tool use, and multi-step procedures, making them a natural setting for egocentric procedural reasoning and real-time assistance (Li et al., 2026b; Perrett et al., 2025; Damen et al., 2018; Li et al., 2026a; Abdelsalam et al., 2023; Peddi et al., 2024). Early VideoQA benchmarks such as MovieQA (Tapaswi et al., 2016), TGIF-QA (Jang et al., 2017), and TVQA (Lei et al., 2018) focused on multiple-choice evaluation, while MSRVTT-QA and MSVD-QA Xu et al. (2017) scaled automatic annotation and ActivityNet-QA (Yu et al., 2019) extended the task to open-ended questions on longer videos. Egocentric benchmarks, however, still largely emphasize recognition, detection, or localization (Plizzari et al., 2024): Ego4D (Grauman et al., 2022) and EPIC-KITCHENS (Damen et al., 2018) provide dense action and object annotations, whereas EgoTaskQA (Jia et al., 2022) and EgoSchema (Mangalam et al., 2023) use multiple-choice evaluation. EgoBlind (Xiao et al., 2025) supports timestamp-aware open-ended QA with future frames withheld, but mainly targets navigation and safety. In contrast, EgoChef focuses on timestamped open-ended QA for fine-grained procedural understanding in real-time first-person cooking assistance.

MLLMs for Video Understanding. By integrating visual inputs with pre-trained LLMs (Alayrac et al., 2022; Li et al., 2023; Liu et al., 2023), MLLMs extend open-ended reasoning to video (Li et al., 2025c; Maaz et al., 2024; Lin et al., 2024; Tang et al., 2025). Although most target general-purpose or third-person settings, recent work increasingly studies egocentric reasoning (He et al., 2026; Zamani & Ziaeetabar, 2026). EgoThink (Cheng et al., 2024b) and VidEgoThink (Cheng et al., 2024a) evaluate first-person reasoning, MM-Ego (Ye et al., 2025) targets embodied assistance, and EgoCross (Li et al., 2026c) reveals weak cross-domain generalization. EgoBlind further reports around 60% accuracy for leading MLLMs on egocentric assistive QA, compared with 87.4% human performance. Despite this progress, procedural reasoning in first-person cooking remains underexplored (Li et al., 2026a; 2025b; Perrett et al., 2025; Tateno et al., 2026; Aggarwal et al., 2026). EgoChef therefore evaluates representative MLLMs under timestamp-restricted observations, targeting persistent weaknesses in fine-grained video understanding (Fu et al., 2025; Li et al., 2024a), particularly temporal reasoning and procedural state tracking.

Multi-Model Fusion for Annotation Construction. Reliable open-ended VideoQA requires high-quality annotations that accommodate semantic variation (Luo et al., 2021; Han et al., 2025; Ruiz et al., 2025). Single-model predictions can be unstable under visual ambiguity and model bias, while majority-vote self-consistency cannot fully reconcile semantically equivalent answers with different phrasings (Wang et al., 2023; Oh & Lee, 2026; Knappe et al., 2024; Liang et al., 2024). We therefore use a multi-model semantic fusion pipeline combining synonym normalization, outlier filtering, cross-model disagreement resolution, and human calibration to reduce lexical noise and prediction conflicts while producing reliable timestamp-aligned reference answers.

3 The EgoChef Dataset

Building EgoChef requires more than attaching questions to collected videos. We emphasize two requirements: *causal validity*, ensuring each question is answerable from information available at its timestamp, and *semantic reliability*, ensuring open-ended references capture valid linguistic vari-

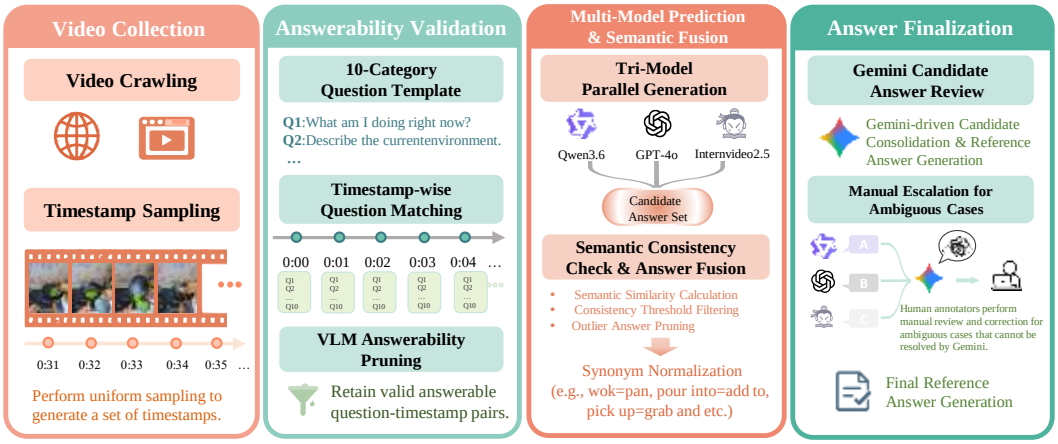


Figure 2: Overview of the EgoChef construction pipeline. Candidate timestamps are paired with question templates, filtered for past-only answerability, and annotated via multi-model prediction, semantic fusion, disagreement resolution, and human calibration.

ation without ambiguity. As shown in Figure 2, EgoChef uses four stages: video collection and timestamp sampling, question formulation for real-time assistance, past-only answerability filtering, and multi-reference annotation via multi-model semantic fusion and human calibration.

3.1 Video Collection

We collect 129 long-form egocentric cooking videos from public video-sharing platforms, including YouTube, Bilibili, and TikTok, with an average duration of 74 seconds. Captured using GoPro cameras, chest-mounted phones, or handheld smartphones, these videos document complete cooking procedures spanning home-style stir-fries, baking, ingredient preparation, and other kitchen activities. Unlike short action clips, they preserve natural procedural transitions, evolving object states, and dense hand-object interactions in real kitchens, providing the temporal structure needed for online procedural reasoning. We uniformly sample timestamps from each video as candidate query moments, which serve as temporal anchors for subsequent question construction.

3.2 Question Formulation

Rather than organizing EgoChef around generic VideoQA categories, we define its question space around what a real-time cooking assistant needs to understand at any moment: the current state, its relation to the surrounding procedure, and the fine-grained details of the ongoing operation. This yields ten question types spanning three complementary dimensions. (1) Current-state perception. Current action, cooking step, and environment description characterize *what is happening now*. (2) Temporal reasoning. Previous-action and next-action questions probe the procedural context immediately before and after the current moment. (3) Operation-level understanding. Ingredient, tool use, quantity estimation, ingredient preparation, and safety warning require fine-grained grounding of objects, states, interactions, and potential risks. Applying these templates to the sampled timestamps produces candidate question-timestamp pairs for subsequent validation. Representative examples are shown in Figure 3.

3.3 Past-Only Answerability Filtering

A central and persistent challenge in timestamped VideoQA is that *not every question is answerable at every moment*. Ingredient quantity, for example, cannot be reliably inferred when the relevant ingredient is not observable, while asking for the next operation near the end of a procedure may be ill-defined. Retaining such pairs would confound model capability with missing evidence and potentially introduce noisy supervision.

We therefore validate answerability *before* constructing reference answers. For each candidate question-timestamp pair, InternVideo2.5 determines whether the question can be answered using

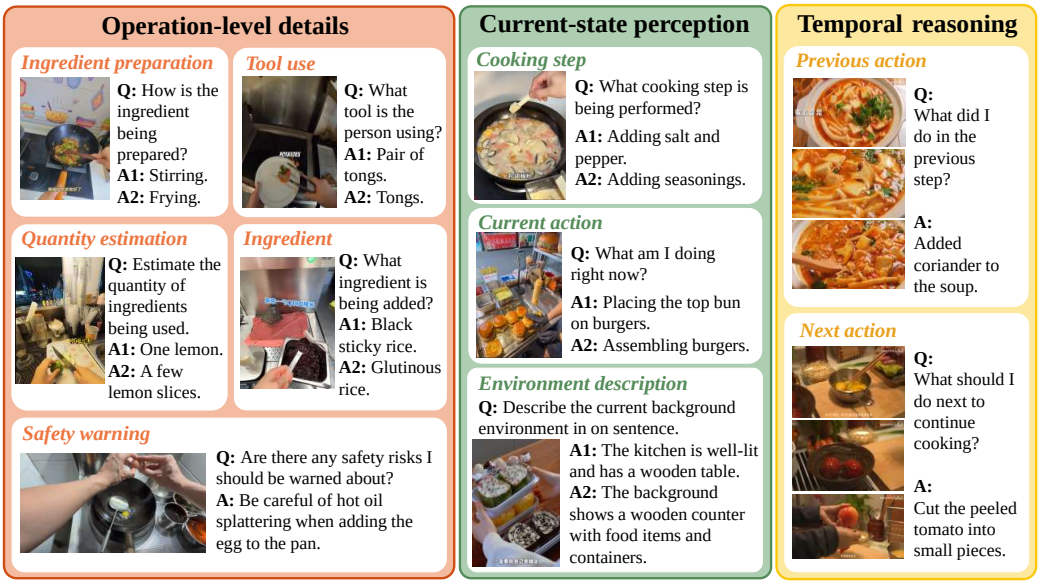
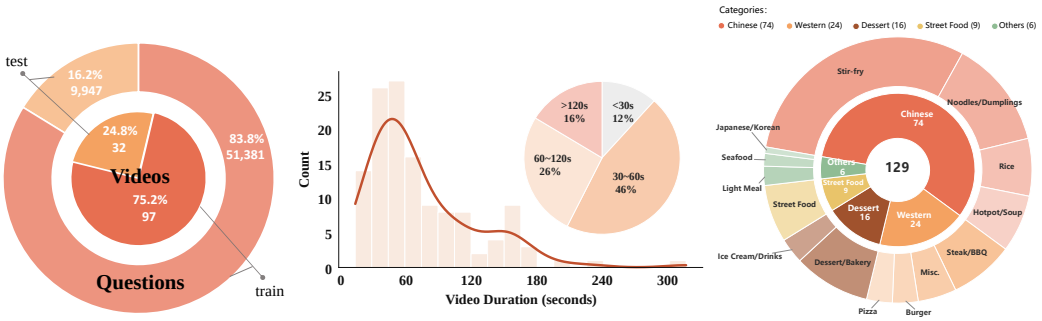


Figure 3: Representative question types in EgoChef. The benchmark spans current-state perception, temporal reasoning, and operation-level understanding, covering ongoing actions, procedural context, ingredients, tools, quantities, preparation methods, and safety risks.



(a) Data split. (b) Video duration. (c) Food categories.

Figure 4: Dataset statistics of EgoChef. (a) Video-level train-test split and corresponding QA distribution. (b) Distribution of video durations. (c) Distribution of food categories.

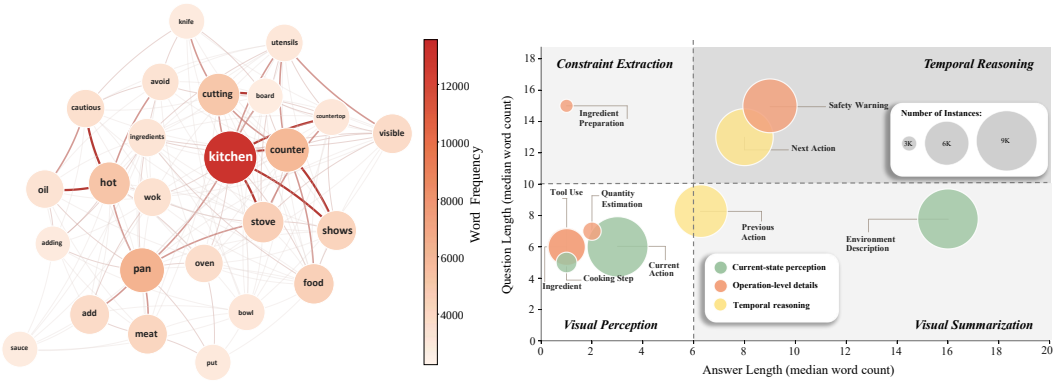
only observations up to and including the query timestamp, with no access to future frames. Pairs lacking sufficient evidence are discarded, removing 32.3% of the initial candidates. This step retains only questions supported by information genuinely available to an online assistant and, importantly, enforces the same past-only information boundary during benchmark construction and evaluation.

3.4 Answer Annotation via Multi-Model Semantic Fusion

Reference answers are standardized as concise phrases or single actionable instructions, then produced by a three-stage pipeline: (1) independent prediction by GPT-4o, Qwen3.6, and InternVideo2.5. (2) semantic fusion via token-overlap similarity and synonym normalization. (3) disagreement resolution with human calibration. This yields up to four reference answers per question, all following the same past-only constraint. Full details are provided in Appendix A.1.

3.5 Dataset Statistics and Analysis

As shown in Figure 4, EgoChef contains 61,328 QA pairs across 129 videos. We assign 32 videos (9,947 QA pairs) to the test set and the remaining videos to training, split at the video level. Videos average 74 seconds, with over 45% lasting 30-60 seconds, and span six cooking categories. Figure 5 shows that reference answers form compositional semantic clusters (a), and that question length is



(a) Answer co-occurrence network. (b) Question-answer distribution.

Figure 5: Semantic and question-answer analysis of EgoChef. (a) Co-occurrence network of reference answers, where node size and color encode term frequency and edges denote co-occurrence. (b) Median question and answer lengths across question types; bubble size indicates instance frequency and colors denote the three reasoning dimensions.

not a reliable proxy for difficulty: short operation-level questions remain challenging due to weak linguistic priors and the need for precise visual grounding (b). Full statistics are in Appendix A.2.

3.6 Comparison with Existing Benchmark Datasets

Table 1 compares EgoChef with existing general and specialized VideoQA benchmarks along the dimensions most relevant to online procedural understanding.

(1) General Benchmarks. General-purpose datasets range from third-person VideoQA benchmarks such as MSRVTT-QA, Video-MME, and MVBench to first-person life-logging datasets such as EgoTaskQA, EgoSchema, and EgoLife. Although many support current-state perception or temporal reasoning, they largely lack fine-grained operation-level supervision. Moreover, few combine strict past-only evaluation with multiple reference answers, thereby severely limiting their ability to evaluate open-ended real-time assistance.

(2) Specialized Benchmarks. Specialized benchmarks exhibit a complementary mismatch. EgoProceVQA and HD-EPIC provide richer procedural or operation-level reasoning but remain offline and rely on limited answer supervision, while timestamp-aware benchmarks such as EgoBlind and EgoTextVQA target comparatively less structured assistive tasks. EgoChef bridges these settings by jointly supporting current-state perception, temporal reasoning, and operation-level understanding under strict past-only access with multi-reference open-ended evaluation.

4 Experiments

4.1 Experiment Settings

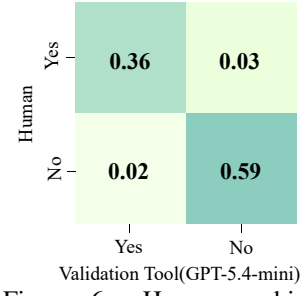


Figure 6: Human-machine agreement.

Evaluation Setup and Metric. To simulate real-time assistance, all models are evaluated under the same strict past-only constraint used in EgoChef construction. For each question posed at timestamp t, the model receives only video observations up to and including t, with all future frames withheld. Since EgoChef adopts open-ended evaluation, models are instructed to answer concisely, typically with a short phrase or a single sentence, reducing variation due purely to response verbosity. We use GPT-5.4-mini as an LLM-based judge to assess semantic equivalence between model predictions and reference answers. For each sample, the judge receives the question, the prediction, and all one to four references, and accepts the prediction if it is semantically equivalent to any valid reference. We refine the judging prompt through manual inspection and human spot-checking on a subset of the test set. As shown in Figure 6, the resulting evaluator achieves a Cohen's κ of 0.89 with human annotators, indicating strong agreement despite linguistic variation in open-ended responses.

Table 1: Comparison of EgoChef with existing video question answering benchmarks. **Per.**: Current-state perception; **Temp.**: Temporal reasoning; **Oper.**: Operation-level details; **RT**: Answer based on video content up to the question timestamp (Past-only). **MA**: Multiple reference answers for each question. **OE/MC**: Open-Ended/Multi-Choice.

Category	Datasets	# V	# Q	Persp.	Reasoning Capability			RT	MA	Task
					Per.	Temp.	Oper.			
Gen.	EgoTaskQA (Jia et al., 2022)	2K	40K	1st	✓	✗	✗	✗	✗	OE
	EgoSchema (Mangalam et al., 2023)	5K	5K	1st	✓	✓	✗	✗	✗	MC
	EgoMemoria (Ye et al., 2025)	629	7K	1st	✓	✓	✗	✗	✗	MC
	EgoThink (Cheng et al., 2024b)	595	700	1st	✓	✓	✗	✗	✗	OE
	EgoLife (Yang et al., 2025)	6	6K	1st	✓	✓	✓	✗	✗	MC
	MSRVTT-QA (Xu et al., 2017)	10K	243K	3rd	✓	✗	✗	✗	✗	OE
	Video-MME (Fu et al., 2025)	900	2.7K	3rd	✓	✓	✗	✗	✗	MC
	MMBench-Video (Fang et al., 2024)	600	2K	3rd	✓	✓	✗	✗	✗	OE
	MVBench (Li et al., 2024a)	4K	20K	3rd	✓	✓	✗	✗	✗	MC
Spec.	EgoBlind (Xiao et al., 2025)	1.3K	5.3K	1st	✓	✗	✗	✓	✓	OE
	EgoProceVQA (Li et al., 2026a)	632	3.6K	1st	✓	✓	✓	✗	✗	OE
	EgoTextVQA (Zhou et al., 2025)	1.5K	7K	1st	✓	✗	✗	✓	✗	OE
	HD-EPIC (Perrett et al., 2025)	156	26K	1st	✓	✓	✓	✗	✗	MC
	EgoChef (Ours)	**129**	**61K**	1st	✓	✓	✓	✓	✓	OE

Table 2: Per-category and overall accuracy of models on the EgoChef test set. Best and second-best results within each group (closed- vs. open-source) are highlighted in red and blue respectively.

Methods	LLM	Size	#F	Per.	Temp.	Oper.	Overall
Closed-source Models							
GPT-5.6 Terra	-	ori	8	65.36%	55.92%	47.94%	55.68%
Gemini 2.5 Flash	-	ori	1fps	57.31%	46.34%	47.88%	50.89%
Gemini 3 Flash	-	ori	1fps	53.06%	39.72%	52.08%	49.94%
Qwen3.7-Flash	-	ori	1fps	49.71%	21.51%	36.40%	38.11%
Open-source Models							
Qwen2.5-VL (Wang et al., 2024)	Qwen2.5-7B	ori	1fps	41.15%	20.36%	32.96%	33.32%
LLaVA-OV (Li et al., 2025a)	Qwen2-7B	384^2	16	40.69%	18.30%	28.85%	30.90%
Video-LLaMA3 (Zhang et al., 2025)	Qwen2.5-7B	ori	1fps	33.47%	16.75%	27.86%	27.61%
MiniCPM-V 2.6 (Yao et al., 2024)	Qwen2.5-7B	384^2	1fps	36.69%	10.83%	23.55%	25.63%
ShareGPT4Video (Chen et al., 2024a)	LLaMA3-8B	ori	16	26.53%	6.77%	14.22%	17.06%
Video-LLaVA (Lin et al., 2024)	Vicuna-7B	224^2	8	21.47%	4.51%	15.59%	15.44%
LLaMA-VID (Li et al., 2024b)	Vicuna-7B	224^2	1fps	17.76%	3.66%	18.63%	15.32%
InternVL2.5-8B (Chen et al., 2024b)	InternLM2_5-7B	448^2	8	17.25%	2.31%	13.30%	12.49%

Baseline Models. We evaluate representative closed- and open-source MLLMs with diverse visual resolutions, frame-sampling strategies, and video-processing architectures. Closed-source models include GPT-5.6 Terra, Gemini 2.5 Flash, Gemini 3 Flash, and Qwen3.7-Flash; open-source models include Qwen2.5-VL, LLaVA-OV, Video-LLaMA3, MiniCPM-V 2.6, ShareGPT4Video, Video-LLaVA, LLaMA-VID, and InternVL2.5-8B. To prevent annotation leakage, we exclude models involved in reference construction, including GPT-4o, Qwen3.6, and InternVideo2.5. Evaluating them against references they helped generate would confound generalization with self-consistency.

4.2 Quantitative Analysis

Table 2 reports performance across the three EgoChef dimensions. Three observations emerge.

(1) Overall performance remains limited. The strongest evaluated model, GPT-5.6 Terra, achieves 55.68% overall accuracy, whereas InternVL2.5-8B reaches only 12.49%. These results remain far from what would be expected if EgoChef reduced to conventional video question answering. The key distinction is that EgoChef requires understanding a procedure as it unfolds, without access to future frames that could retrospectively resolve ambiguity in the current state.

(2) Temporal reasoning is the most consistent bottleneck. Across models, performance is generally highest on current-state perception, lower on operation-level understanding, and lowest on temporal reasoning. GPT-5.6 Terra reaches 55.92% on temporal reasoning, while most open-source models remain below 25%, with InternVL2.5-8B falling to 2.31%. This gap reflects a fundamental dif-

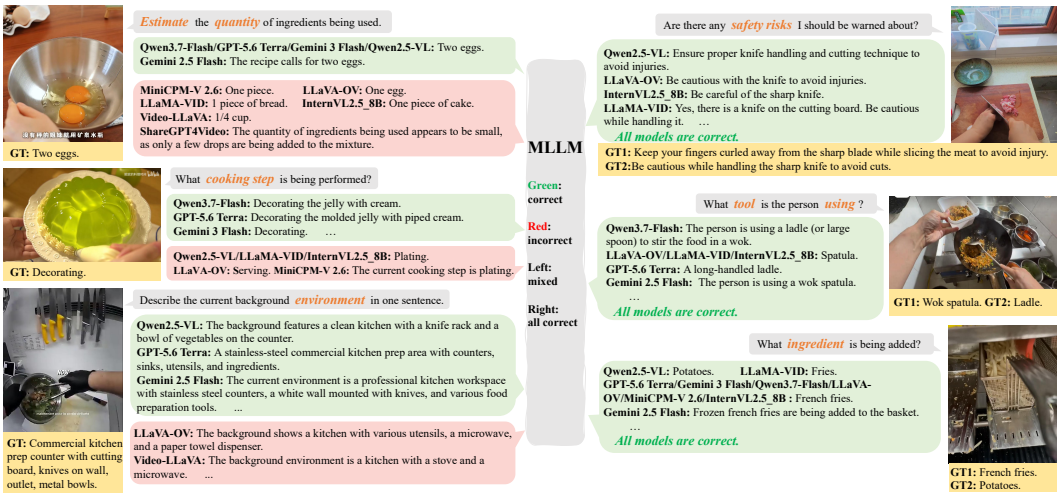
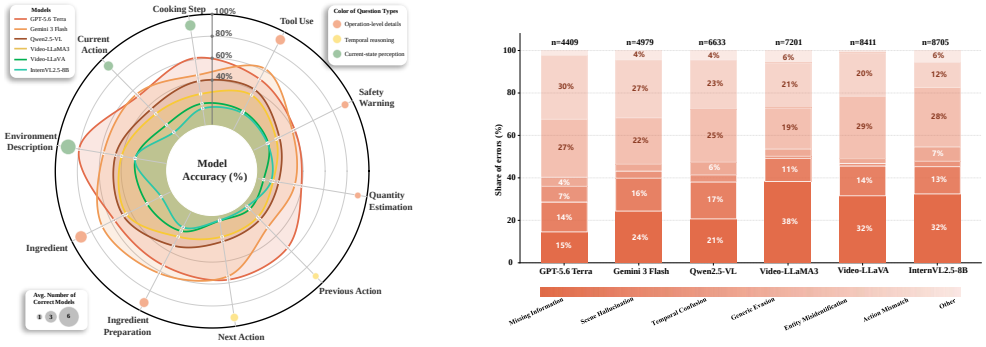


Figure 7: Qualitative comparison of MLLM predictions on EgoChef. Left: mixed cases with correct and incorrect predictions. Right: all-correct cases where all models succeed.



(a) Per-category acc. (bubble: avg. correct models/type). (b) Error type distribution of MLLMs on EgoChef.

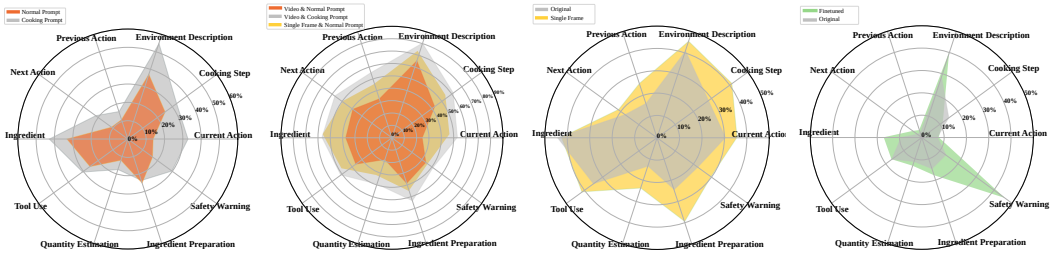
Figure 8: Model performance and error analysis on EgoChef. (a) Per-category accuracy with bubble size showing avg. correct models per type. (b) Error type distribution of MLLMs on EgoChef.

ference in the information required: perceptual questions can often be answered from the current observation, whereas temporal questions require relating the present to prior operations or, for next-action prediction, inferring a plausible continuation without future evidence. The degradation therefore cannot be explained by weak visual grounding alone. Rather, it exposes difficulty in organizing observed history into a temporally grounded procedural state that can support both retrospective reasoning and prospective inference.

(3) A substantial closed-source and open-source gap remains. The three strongest closed-source models each exceed 49% overall accuracy, whereas the best open-source model, Qwen2.5-VL, reaches 33.32%. This gap persists across open-source models with different resolutions and video-processing strategies. Because these systems also differ in architecture, scale, training data, and alignment, the gap cannot be attributed to any single factor. Nevertheless, the results suggest that stronger visual input alone is insufficient: effective online procedural understanding also requires temporal selection, state tracking, and continual updating of procedural context.

4.3 Qualitative Analysis

Figure 7 compares model predictions on identical inputs to reveal where models diverge, rather than only how often they fail. The left column shows mixed cases in which some models succeed while others fail on the same observation. In the cooking-step example, Qwen3.7-Flash, GPT-5.6 Terra, and Gemini 3 Flash correctly identify the action as *decorating*, whereas most other models produce the coarser label *plating*. A similar pattern appears across the mixed cases: correct predictions



(a) LLaVA-OV. (b) GPT-5.6 Terra. (c) Video-LLaMA3. (d) InternVL2.5-8B.

Figure 9: Ablation on prompt design, single-frame versus video input, and fine-tuning on EgoChef.

identify the specific ongoing action, while incorrect ones often substitute a nearby but less specific procedural step. These errors therefore reflect insufficient procedural granularity rather than gross visual misperception.

The right column shows cases in which all listed models answer correctly, including questions about safety risks, tool identity, and the ingredient being added. These examples indicate that the relevant visual evidence is not inherently ambiguous: when the required object or action is directly and clearly observable, predictions converge across models despite substantial differences in model scale and design. The contrast between the two columns therefore localizes the difficulty to fine-grained procedural-state resolution rather than task unanswerability.

Figure 8 quantifies these patterns over the full evaluation set. The polar plot in (a) reports per-category accuracy across the ten question types, with bubble size indicating the average number of models answering correctly in each category. The two temporal-reasoning categories, *previous-action* and *next-action*, exhibit smaller bubbles than most perception-oriented categories, such as *current-action* and *environment description*, consistent with the quantitative results in Section 4.2. The distinction is intuitive: previous- and next-action questions require reasoning beyond the current frame, whereas categories such as environment description or ingredient identification can often be resolved from the present observation alone.

The stacked bars in (b) decompose errors by type across six models, ordered by increasing total error count, with n ranging from 1,687 to 8,703. *Entity misidentification* and *missing information* account for a larger share of errors among models with higher total error counts, whereas stronger models retain a comparatively larger proportion of *temporal confusion* and *generic evasion*. This shift suggests that model improvement does not eliminate all failure modes uniformly. As basic entity grounding improves, temporal and generic-response errors constitute a larger fraction of the residual failures rather than disappearing at the same rate.

Taken together, the quantitative and qualitative analyses identify a consistent bottleneck: models are least reliable when answering requires reasoning about procedural states that are not directly observable in the current frame, whether those states lie in the past or must be inferred into the future. A detailed analysis of representative failure cases is provided in Appendix B.1.

4.4 Model Analysis

We examine the effects of temporal context, prompt design, and task-specific adaptation in Figure 9. We find that multi-frame video input generally helps but is not always better than a single frame, that cooking-specific prompting improves alignment across most categories, and that LoRA (Hu et al., 2022) fine-tuning yields overall gains but does not resolve fine-grained failures such as tool recognition. Detailed analysis is provided in Appendix B.2.

5 Conclusion

We present EgoChef, an open-ended benchmark for online egocentric procedural understanding under strict past-only visual access. Our evaluation shows that current MLLMs remain substantially limited, particularly in temporal reasoning, where they struggle to relate the current state to relevant history and anticipate plausible next steps. Together, these results suggest that the key challenge is not simply accessing more visual context, but maintaining and updating the right procedural state as an activity unfolds. We hope EgoChef provides a practical testbed for advancing real-time egocentric assistants with stronger procedural reasoning and anticipation.

AI USE STATEMENT

In this work, we used generative AI tools for language polishing and for generating Python code for figures and LaTeX tables. We have not used generative AI tools for generating research ideas, designing experiments, or analyzing data, and the rest of the required disclosure tasks are not applicable to this work. We have reviewed all AI-assisted work. For example, all LLM-generated code was tested and verified for correctness by running it and checking the outputs, and all AI-assisted text was manually reviewed and edited by the authors. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Mohamed A. Abdelsalam, Samrudhdhi B. Rangrej, Isma Hadji, Nikita Dvornik, Konstantinos G. Derpanis, and Afsaneh Fazly. Gepsan: Generative procedure step anticipation in cooking videos. In *ICCV*, 2023.

Lavisha Aggarwal, Vikas Bahirwani, and Andrea Colaco. From videos to conversations: Egocentric instructions for task assistance, 2026.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning. In *NeurIPS*, 2022.

Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Bin Lin, Zhenyu Tang, Li Yuan, Yu Qiao, Dahua Lin, Feng Zhao, and Jiaqi Wang. Sharegpt4video: Improving video understanding and generation with better captions. In *NeurIPS*, 2024a.

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. *Sci. China Inf. Sci.*, 67(12):220101, 2024b.

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

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*, 2024b.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The epic-kitchens dataset. In *ECCV*, 2018.

Xinyu Fang, Kangrui Mao, Haodong Duan, Xiangyu Zhao, Yining Li, Dahua Lin, and Kai Chen. Mmbench-video: A long-form multi-shot benchmark for holistic video understanding. In *NeurIPS*, 2024.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, Ran He, and Xing Sun. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *CVPR*, 2025.

Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, Florian Hoppe, Christian Thurau, Ingo Bax, and Roland Memisevic. The "something something" video database for learning and evaluating visual common sense. In *ICCV*, 2017.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *CVPR*, 2022.

Songhao Han, Wei Huang, Hairong Shi, Le Zhuo, Xiu Su, Shifeng Zhang, Xu Zhou, Xiaojuan Qi, Yue Liao, and Si Liu. VideoEspresso: A large-scale chain-of-thought dataset for fine-grained video reasoning via core frame selection. In *CVPR*, 2025.

Yuping He, Yifei Huang, Guo Chen, Lidong Lu, Baoqi Pei, Jilan Xu, Tong Lu, and Yoichi Sato. Bridging perspectives: A survey on cross-view collaborative intelligence with egocentric-exocentric vision. *Int. J. Comput. Vis.*, 134(2):62, 2026.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022.

Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. TGIF-QA: Toward spatio-temporal reasoning in visual question answering. In *CVPR*, 2017.

Baoxiong Jia, Ting Lei, Song-Chun Zhu, and Siyuan Huang. Egotaskqa: Understanding human tasks in egocentric videos. In *NeurIPS*, 2022.

Tim Knappe, Ryan Li, Ayushi Chauhan, Kaylee Chhua, Kevin Zhu, and Sean O'Brien. Semantic self-consistency: Enhancing language model reasoning via semantic weighting, 2024.

Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L. Berg. Tvqa: Localized, compositional video question answering. In *EMNLP*, 2018.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-OneVision: Easy visual task transfer. *Trans. Mach. Learn. Res.*, 2025a.

Junlong Li, Huaiyuan Xu, Sijie Cheng, Kejun Wu, Kim-Hui Yap, Lap-Pui Chau, and Yi Wang. Building egocentric procedural AI assistant: Methods, benchmarks, and challenges, 2025b.

Junlong Li, Junxi Li, Yuxiang Yang, Wenbin Zou, Lap-Pui Chau, and Yi Wang. Egoprocevqa: A novel egocentric procedural understanding task with self-skill-exploration agent. arXiv preprint arXiv:2607.13792, 2026a.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. Mvbench: A comprehensive multi-modal video understanding benchmark. In *CVPR*, 2024a.

Kunchang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: chat-centric video understanding. *Sci. China Inf. Sci.*, 68(10):200102, 2025c.

Xiang Li, Heqian Qiu, Lanxiao Wang, Hanwen Zhang, Chenghao Qi, Linfeng Han, Huiyu Xiong, and Hongliang Li. Challenges and trends in egocentric vision: A survey. *Mach. Intell. Res.*, 23 (1):1–33, 2026b.

Yanjun Li, Yuqian Fu, Tianwen Qian, Qi'Ao Xu, Silong Dai, Danda Pani Paudel, Luc Van Gool, and Xiaoling Wang. Egocross: Benchmarking multimodal large language models for cross-domain egocentric video question answering. In *AAAI*, 2026c.

Yanwei Li, Chengyao Wang, and Jiaya Jia. LLaMA-VID: An image is worth 2 tokens in large language models. In *ECCV*, 2024b.

Xun Liang, Shichao Song, Zifan Zheng, Hanyu Wang, Qingchen Yu, Xunkai Li, Rong-Hua Li, Yi Wang, Zhonghao Wang, Feiyu Xiong, and Zhiyu Li. Internal consistency and self-feedback in large language models: A survey. *arXiv preprint arXiv:2407.14507*, 2024.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-LLaVA: Learning united visual representation by alignment before projection. In *EMNLP*, 2024.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023.

Man Luo, Shailaja Keyur Sampat, Riley Tallman, Yankai Zeng, Manuha Vancha, Akarshan Sajja, and Chitta Baral. 'just because you are right, doesn't mean i am wrong': Overcoming a bottleneck in the development and evaluation of open-ended vqa tasks. In *EACL*, 2021.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *ACL*, 2024.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. In *NeurIPS*, 2023.

Jungsuk Oh and Jay-Yoon Lee. Latent self-consistency for reliable majority-set selection in short- and long-answer reasoning. In *AAAI*, 2026.

OpenAI, 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.

Rohith Peddi, Shivvrat Arya, Bharath Challa, Likhitha Pallapothula, Akshay Vyas, Bhavya Gouripeddi, Qifan Zhang, Jikai Wang, Vasundhara Komaragiri, Eric Ragan, Nicholas Ruozzi, Yu Xiang, and Vibhav Gogate. CaptainCook4D: A dataset for understanding errors in procedural activities. In *NeurIPS*, 2024.

Toby Perrett, Ahmad Darkhalil, Saptarshi Sinha, et al. Hd-epic: A highly-detailed egocentric video dataset. In *CVPR*, 2025.

Chiara Plizzari, Gabriele Goletto, Antonino Furnari, Siddhant Bansal, Francesco Ragusa, Giovanni Maria Farinella, Dima Damen, and Tatiana Tommasi. An outlook into the future of egocentric vision. *Int. J. Comput. Vis.*, 2024.

Tomas Ruiz, Tanalp Agustoslu, and Carsten Schwemmer. Rethinking ground truth: A case study on human label variation in mllm benchmarking. In *BigData*, 2025.

Yunlong Tang, Jing Bi, Siting Xu, Luchuan Song, Susan Liang, Teng Wang, Daoan Zhang, Jie An, Jingyang Lin, Rongyi Zhu, Ali Vosoughi, Chao Huang, Zeliang Zhang, Pinxin Liu, Mingqian Feng, Feng Zheng, Jianguo Zhang, Ping Luo, Jiebo Luo, and Chenliang Xu. Video understanding with large language models: A survey. *IEEE Trans. Circuits Syst. Video Technol.*, 36(2):1355–1376, 2025.

Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Movieqa: Understanding stories in movies through question-answering. In *CVPR*, 2016.

Masatoshi Tateno, Gido Kato, Hirokatsu Kataoka, Yoichi Sato, and Takuma Yagi. Handyvqa: A video QA benchmark for fine-grained hand-object interaction dynamics. In *CVPR*, 2026.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jia 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.

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. arXiv preprint arXiv:2409.12191, 2024.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *ICLR*, 2023.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *CVPR*, 2021.

Junbin Xiao, Nanxin Huang, Hao Qiu, Zhulin Tao, Xun Yang, Richang Hong, Meng Wang, and Angela Yao. Egoblind: Towards egocentric visual assistance for the blind. In *NeurIPS*, 2025.

Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *ACM MM*, 2017.

Jingkang Yang, Shuai Liu, Hongming Guo, Yuhao Dong, Xiamengwei Zhang, Sicheng Zhang, Pengyun Wang, Zitang Zhou, Binzhu Xie, Ziyue Wang, Bei Ouyang, Zhengyu Lin, Marco Cominelli, Zhongang Cai, Bo Li, Yuanhan Zhang, Peiyuan Zhang, Fangzhou Hong, Joerg Widmer, Francesco Gringoli, Lei Yang, and Ziwei Liu. Egolife: Towards egocentric life assistant. In *CVPR*, 2025.

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.

Hanrong Ye, Haotian Zhang, Erik Daxberger, Lin Chen, Zongyu Lin, Yanghao Li, Bowen Zhang, Haoxuan You, Dan Xu, Zhe Gan, Jiasen Lu, and Yinfei Yang. Mm-ego: Towards building egocentric multimodal llms for video qa. In *ICLR*, 2025.

Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. ActivityNet-QA: A dataset for understanding complex web videos via question answering. In *AAAI*, 2019.

Mohammad Zamani and Fatemeh Ziaeetabar. Vision-language models for egocentric video: From hand-object interaction to embodied ai, 2026.

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, Peng Jin, Wenqi Zhang, Fan Wang, Lidong Bing, and Deli Zhao. VideoLLaMA 3: Frontier multimodal foundation models for image and video understanding, 2025.

Sheng Zhou, Junbin Xiao, Qingyun Li, Yicong Li, Xun Yang, Dan Guo, Meng Wang, Tat-Seng Chua, and Angela Yao. Egotextvqa: Towards egocentric scene-text aware video question answering. In *CVPR*, 2025.

A Details

A.1 Answer Annotation via Multi-Model Semantic Fusion

Once *when* a question is answerable has been established, the remaining challenge is determining *what* constitutes a valid answer. This is particularly important for open-ended QA, where semantically equivalent responses may differ substantially in wording. We first standardize the expected output as either a concise phrase (e.g., "cutting" or "pouring") or a single actionable instruction, and then construct reference answers through three stages.

(1) Independent prediction. GPT-4o, Qwen3.6, and InternVideo2.5 independently generate candidate answers for every valid question-timestamp pair. Using multiple independent annotators reduces reliance on any single model while exposing cases of disagreement. (2) Semantic fusion. We assess consistency among candidate answers using token-overlap similarity together with synonym normalization. When at least two models agree semantically, we retain the most informative consistent responses, remove duplicates, and preserve up to four reference answers. Minor color discrepancies caused by illumination, video filtering, or compression are ignored during matching to avoid rejecting otherwise equivalent answers. (3) Disagreement resolution and human calibration. When candidate answers disagree substantially or remain uncertain, Gemini 3.5 Flash reviews the visual history up to the query timestamp and either consolidates a reliable answer or escalates the pair for human inspection. Human annotators then resolve ambiguous cases and additionally spot-check 30% of the resulting QA pairs. Both stages follow the same past-only constraint, ensuring that no post-query visual evidence is introduced during annotation.

Together, answerability filtering and multi-model fusion address two distinct sources of benchmark noise: the former ensures that each question is supported by the visual evidence available at its timestamp, while the latter reconciles linguistic variation and model disagreement into reliable, timestamp-aligned, multi-reference supervision.

A.2 Dataset Statistics and Analysis

(1) Data Split. EgoChef contains 61,328 QA pairs across 129 videos. We assign 32 videos and their 9,947 QA pairs to the test set, with the remaining videos forming the training set for model adaptation and fine-tuning. The split is performed at the video level to keep training and evaluation procedures disjoint. Dataset statistics are summarized in Figure 4.

(2) Video Duration and Food Diversity. As shown in Figure 4(b), EgoChef videos average 74 seconds, with more than 45% lasting 30-60 seconds. These videos are long enough to capture meaningful procedural evolution across multiple cooking stages, while a subset exceeding 120 seconds introduces additional long-range temporal variation. The dataset further spans diverse cooking categories, including Chinese, Western, Dessert, Japanese/Korean, Street Food, and Light Meal. Chinese cuisine forms the largest group with 74 videos, including 39 stir-fry videos and 17 wheat-based dishes such as noodles and dumplings. This diversity exposes the same question space to different ingredients, tools, visual environments, and procedural conventions.

(3) Answer Semantics. Figure 5(a) visualizes the frequency and co-occurrence structure of reference-answer terms after removing common stop words. The resulting graph contains semantic clusters spanning cookware (e.g., stove, pan, wok), food entities (e.g., meat, ingredients), physical states (e.g., hot), and actions (e.g., add, put). These concepts are not independent: state descriptors co-occur with objects undergoing change, while action verbs are closely associated with their corresponding tools. This structure reflects the compositional nature of EgoChef, where answering a question often requires jointly reasoning about *what object is involved, how its state changes, and what operation is being performed*.

(4) Question-Answer Characteristics. Figure 5(b) compares median question and answer lengths across question types. Question length is not a reliable proxy for difficulty. Several operation-level questions occupy the short-question, short-answer regime yet remain challenging because they provide weak linguistic priors and require precise visual grounding of small or transient objects. By contrast, questions such as safety warning and next action provide richer contextual cues and therefore stronger semantic priors. EgoChef thus spans both visually grounded fine-grained perception

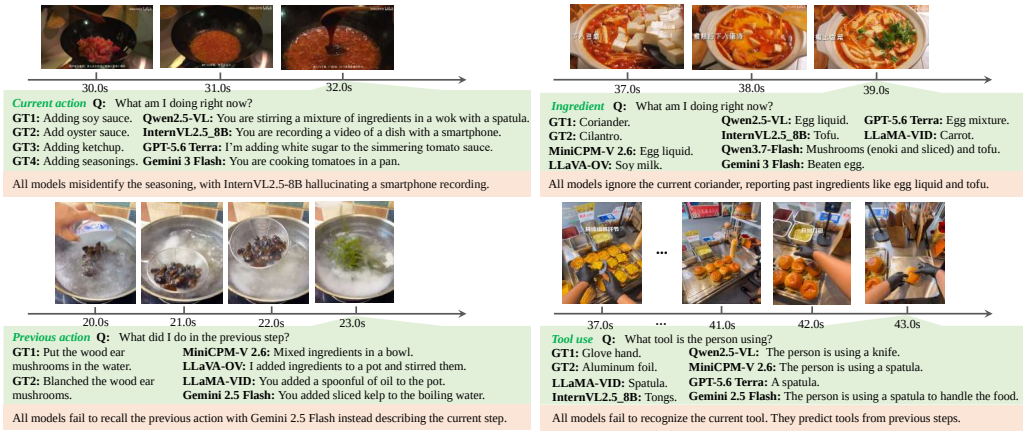


Figure 10: Representative failure cases on EgoChef. Current MLLMs struggle with fine-grained visual grounding, separating current evidence from historical context, retrieving preceding operations, and updating tool-related procedural states.

and context-dependent procedural reasoning, rather than conflating linguistic complexity with reasoning difficulty.

B Experiments

B.1 Failure Case Analysis

Figure 10 presents four representative failure modes that expose complementary weaknesses in current MLLMs.

In the *current-action* case (top-left), models misidentify the seasoning being added, while InternVL2.5-8B produces an unrelated description involving a smartphone. The relevant evidence is fine-grained and transient, further complicated by occlusion and visual similarity among ingredients. This example exposes a basic grounding limitation: reliable procedural reasoning first requires predictions to remain anchored to subtle evidence in the current observation.

The *ingredient* case (top-right) reveals a different problem. Although coriander is visible at the current moment, several models instead predict ingredients observed earlier, such as egg liquid or tofu. This behavior persists even when the prompt explicitly emphasizes the latest visual evidence. The failure is therefore not simply a lack of temporal context. Rather, historical observations can become distracting when models fail to distinguish evidence that remains relevant from evidence that has become obsolete. Visual history is useful, but useful memory must be selective.

The *previous-action* case (bottom-left) exposes weak temporal indexing. Several models describe the current operation instead of retrieving the immediately preceding one, with Gemini 2.5 Flash exhibiting this behavior explicitly. Both events are present in the observed history, yet the models assign them to the wrong temporal roles. Online reasoning therefore requires more than retaining past frames: history must be organized according to *what has happened*, *what is happening now*, and *what may happen next*.

A similar pattern appears in the *tool-use* case (bottom-right). Models frequently predict tools used in earlier steps rather than the tool currently involved. Fine-grained tool recognition is itself difficult because tools may be small, partially occluded, or entangled with complex hand-object interactions. However, the repeated carryover of previously observed tools also indicates difficulty in revising the active procedural state as the operation changes.

Taken together, these failures reveal a common limitation. Current MLLMs do not simply lack visual evidence or memory. They struggle to decide which historical evidence should remain active, which should be suppressed, and how new observations should revise the current state. The cen-

tral challenge is therefore not merely accumulating more visual history, but *maintaining the right procedural state from that history*.

B.2 Ablation Studies

(1) Temporal context helps, but more history is not always better. Replacing multi-frame video input with a single frame generally degrades performance, confirming the value of visual history. However, single-frame input remains competitive on several current-state categories, while the failure cases show that historical evidence can also become stale or distracting. The challenge is therefore not simply to provide more frames, but to retain the history that remains relevant and suppress what no longer is. (2) Domain-specific prompting improves task alignment. Replacing the cooking-specific prompt with a generic VQA prompt consistently reduces performance, particularly for Current Action and Environment Description. For both LLaVA-OV and GPT-5.6 Terra, cooking-specific prompting improves all evaluated categories. Explicit procedural framing therefore helps models interpret visual evidence in a more task-relevant manner, but does not eliminate the underlying temporal and fine-grained errors. (3) Fine-tuning improves adaptation but not all bottlenecks. LoRA fine-tuning on EgoChef improves InternVL2.5-8B across most categories, yet Cooking Step slightly decreases and Tool Use remains nearly unchanged. Thus, additional task-specific supervision improves overall adaptation but remains insufficient for categories dominated by subtle visual cues and rapidly changing hand-object interactions.

Overall, these ablations suggest that the key challenge is not more context or supervision, but *selectively organizing and updating* visual history into an accurate procedural state.

C Limitations

While EgoChef provides a focused testbed for online egocentric procedural understanding, it has several limitations.

First, annotation and evaluation bias. Our reference answers are constructed via multi-model semantic fusion (GPT-4o, Qwen3.6, InternVideo2.5) with human calibration on 30% of the QA pairs. The remaining 70% are unverified and may inherit systematic biases or hallucinations from the generating models; when the annotating models collectively err on the same question type, that error can be consolidated into the reference set. Similarly, our LLM-based judge (GPT-5.4-mini), despite a Cohen's κ of 0.89 against human annotators on a subset, may favor certain phrasing or admit fine-grained visual differences. Neither bias can be fully eliminated by partial human verification, and their direction (over- or under-estimating model capability) remains unclear.

Second, robustness to editing and temporal ambiguity. Because EgoChef videos are collected from public platforms with jump cuts and retrospective narration, the strict past-only boundary may be blurred at edited transitions; how timestamp alignment and answerability filtering should handle such discontinuity remains an open challenge for real-time assistant evaluation.

Despite these limitations, we believe EgoChef serves as a necessary step toward strict past-only online procedural evaluation and provides a solid foundation for future improvements.