GuideMe: Multi-Domain Task Guidance and Intervention in Streaming Video

Fang Liu[*1], Jinpeng Chen[*2], Ke Xu[3,1], Yuhao Liu[1], Huankang Guan[2], Xudong Lu[4], Yang Bo[2], Gerhard Hancke[1], Rui Liu[✉2], Rynson W.H. Lau[✉1,5]

[1] City University of Hong Kong [2] Huawei Research
[3] University of Science and Technology of China
[4] Chinese University of Hong Kong
[5] City University of Hong Kong (Dongguan)

Abstract. While multimodal Large Language Models (MLLMs) excel at offline video understanding, an interesting question of *how far they are from serving as a real-time procedural coach* remains unknown. Such a role typically requires an MLLM to continuously monitor the execution, detect mistakes, and provide corrective guidance in a closed-loop interaction. In this paper, we construct **GuideMe**, the first multi-domain benchmark for streaming video that supports training and evaluation of MLLMs for closed-loop interactive task guidance. It comprises 2,458 videos spanning 223.7 hours across diverse domains (*e.g.*, cooking, object manipulation, daily-life guidance, and fitness), with 47,775 interaction samples covering next-step instructions, completion feedback, error detection, and corrective guidance. To evaluate existing models on GuideMe, we design a three-component assessment framework to measure the capabilities of representative MLLMs, which consists of temporal-semantic bipartite matching for sequence-level alignment, behavioral classification for intervention timing, and LLM-as-a-Judge for content quality. Extensive experiments highlight a critical performance asymmetry: *despite excelling at providing instructions, existing MLLMs consistently fail to identify execution errors and respond with corrective feedback*. Code and data are released at homepage.

Keywords: MLLMs · Interactive Task Guidance · Benchmark

1 Introduction

Recent advances in Multimodal Large Language Models (MLLMs) have demonstrated strong performance on a wide range of visual understanding tasks. Representative commercial models include Gemini 3.1 Pro [16] and GPT-4o [20], while open-source models like Qwen3-VL [1] have also shown competitive capabilities. Unlocking their potential for real-time interaction naturally leads

[*] Joint first authors.
[✉] Corresponding authors.

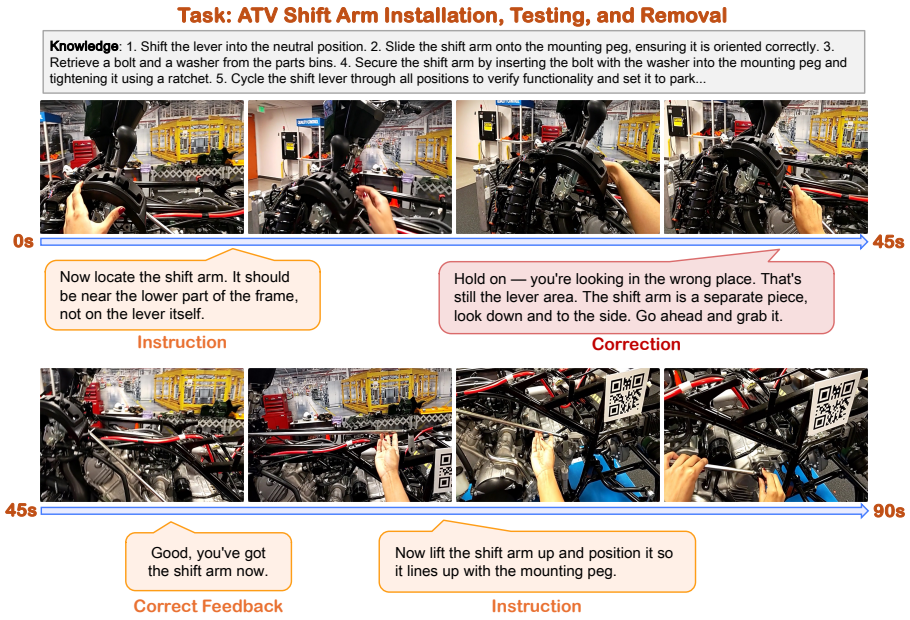


Fig. 1: Data examples in GuideMe. The benchmark captures the full closed-loop interaction cycle: the assistant issues a step-level instruction, monitors the user's execution, and upon detecting an error, provides corrective guidance to steer the user back on track. This cycle repeats throughout the procedure, enabling rigorous evaluation of interactive coaching ability.

to the *procedural task guidance*: enabling an AI assistant to act as a real-time coach for multi-step activities, such as assembling furniture, following a recipe, or performing a workout, by providing instructions, monitoring execution, and intervening when errors occur.

However, serving as a reliable procedural coach requires far more than understanding a video offline. In real-world interactions, users frequently deviate from the intended procedure: they skip steps, use wrong tools, or perform actions in the wrong order. An effective assistant must therefore operate in a *closed loop*: continuously observing the user's actions, detecting mistakes in real time, and providing corrective feedback that guides the user back to the correct trajectory. This cycle of instructing, observing, and correcting steps is the defining challenge that separates a passive narrator from an active coach. As shown in Tab. 1, existing procedural video datasets provide step-level annotations [11, 32, 41, 60] or post-hoc error labels [22, 34, 35, 37, 46], but none of them can evaluate this complete closed-loop cycle under realistic streaming constraints.

To address these gaps, we introduce **GuideMe**, the first multi-domain streaming benchmark designed to evaluate the closed-loop corrective capabilities of AI assistants in procedural tasks. GuideMe comprises 2,458 videos spanning

223.7 hours across four diverse domains (cooking, object manipulation, daily-life guidance, and fitness), yielding a total of 47,775 interaction samples. Each sample captures one of four assistant behaviors: issuing a next-step directive, confirming correct execution, detecting an error, or providing corrective guidance. This structured interaction format enables systematic evaluation of each stage in the closed loop, from basic instruction delivery to the more challenging tasks of error detection and correction.

GuideMe is constructed through a three-stage automated annotation pipeline. First, we extract ordered instructional activities from fine-grained action annotations, categorizing each as a correct action, a wrong action, or a correction action. Second, we generate task knowledge, including task descriptions and canonical step sequences, from the extracted activities via LLM prompting. Third, we generate natural conversational dialogues aligned with action boundaries and timestamps. To evaluate models under streaming conditions, we propose a three-component assessment framework: temporal-semantic bipartite matching for sequence-level alignment, behavioral classification for instance-level intervention timing, and LLM-as-a-Judge for content quality.

Experiments across proprietary models (*e.g.*, Gemini [16], Doubao [6, 7]) and open-source models (*e.g.*, Qwen3-VL [1], MMDuet2 [47]) reveal a striking asymmetry in closed-loop performance: models achieve reasonable scores on basic instruction delivery, but their performance degrades sharply on error detection and corrective guidance, which are fundamental for reliable interactive coaching. This finding exposes a fundamental limitation: *while current MLLMs can describe what should happen, they are far from coaching a user based on what is actually happening.* Our contributions are summarized as follows:

- We introduce GuideMe, the first multi-domain streaming benchmark for interactive procedural guidance. Unlike prior work restricted to offline annotations or single domains, GuideMe unifies four diverse domains (cooking, object manipulation, daily-life guidance, and fitness) into a closed-loop evaluation framework that covers instructions, execution feedback, error detection, and corrective guidance.
- We develop a scalable three-stage annotation pipeline that transforms heterogeneous procedural datasets into unified streaming interaction samples. We propose a three-component evaluation framework consisting of temporal-semantic bipartite matching, behavioral classification, and LLM-as-a-Judge for rigorous assessment under streaming conditions.
- We conduct comprehensive experiments across a wide range of models and uncover a critical capability gap. While current MLLMs can deliver step-by-step instructions, they fail at the harder stages of the closed loop, particularly detecting execution errors and providing corrective feedback, highlighting a clear direction for future research on interactive AI assistants.

Table 1: Comparison of procedural video datasets. GuideMe is the only benchmark that supports all five evaluation dimensions, including the critical *closed-loop interaction* where the assistant's corrective guidance modifies the user's subsequent actions within the same session.

Dataset	Domain	#Videos	Hours	Step-level Instruction	Timed Feedback	Error Detection	Action-specific Correction	Closed-loop Interaction
Epic-Kitchen-100 [11]	Cooking	700	100	✓	×	×	×	×
EpicTent [21]	Tent Making	7	5.4	✓	×	✓	×	×
COIN [41]	Diverse	11,827	476	✓	×	×	×	×
YouCook2 [60]	Cooking	2,000	176	✓	×	×	×	×
IKEA [5]	Assembly	1,113	35.3	✓	×	×	×	×
HoloAssist [46]	Object Manip.	2,221	166	×	×	✓	✓	×
CaptainCook4D [35]	Cooking	384	94.5	×	×	✓	×	×
QEVD-Fit-Coach [34]	Fitness	223	13.5	✓	✓	✓	✓	×
Assembly101 [37]	Assembly	4,321	513	×	×	✓	✓	×
EgoPER [22]	Cooking	386	28	✓	×	✓	×	×
Ours	Diverse	2,458	223.7	✓	✓	✓	✓	✓

2 Related Works

2.1 Multimodal Large Language Models

Recent advances in Multimodal Large Language Models (MLLMs) have substantially broadened the scope of visual understanding [25–27]. Proprietary models such as GPT-4o [20] and Gemini 3 Pro [15] set new standards across diverse tasks spanning video understanding [13, 57], document recognition [14, 31], and mathematical reasoning [44, 58]. Open-source counterparts, including InternVL 3.5 [45], MiniCPM-V 4.5 [56], Qwen3 [1, 2, 51], and DeepSeek-VL2 [50], have rapidly narrowed the gap. However, these models are predominantly designed for offline analysis of pre-recorded content. Deploying them as interactive assistants that must observe, reason, and respond within a continuous video stream remains a largely open challenge.

2.2 Streaming MLLMs and Benchmarks

Early video MLLMs digest a clip offline and respond only after the full video, ill-suited to assistance that must react as events unfold. One line of streaming models thus targets *when* and *how* to respond live: VideoLLM-online [9] interleaves perception and generation, Dispider [36] disentangles perception, decision, and reaction, and StreamMind [12] reaches full frame rate via event-gated cognition. Another scales to long, unbounded streams (StreamingVLM [52], LiveStar [55]) or retrofits offline backbones into streaming responders (StreamBridge [43], MMDuet2 [47]). Orthogonally, memory-centric agents (M3-Agent [28], HippoMM [24]) target *post-hoc* QA over recorded video rather than proactive intervention. For evaluation, benchmarks such as StreamingBench [23], OVO-Bench [33], PhoStream [30], OmniMMI [49], SVBench [54], SpaceVista [39], and X-Stream [40] assess general streaming understanding and proactive reasoning,

while ProactiveVideoQA [48] further targets the user experience of proactive interaction. GuideMe is fundamentally harder: beyond understanding the stream, the model must *proactively* decide when to intervene, what to advise, and whether the user erred, demanding far stronger temporal reasoning and situational awareness.

2.3 Datasets for Procedural Activities

Procedural video datasets differ mainly in how much of the coaching loop they supervise, and none spans it from instruction to recovery (Tab. 1). The largest group pairs step-level instructions with expert demonstrations, ranging from cooking and assembly recordings [5, 11, 41, 60] to large-scale egocentric collections [4, 17–19, 32, 53]. Because every step is executed correctly, these data show what a good action looks like but never how to recover from a flawed one. Error-aware datasets [21, 22, 35, 37] restore the missing failure signal, yet stop at detecting mistakes and leave the user without guidance on how to fix them. Closest to genuine assistance, HoloAssist [46] records real instructor corrections and QEVD [34] adds timed, action-specific feedback; even so, neither closes the loop, since later guidance is scripted in advance rather than conditioned on whether the user actually recovered, and QEVD stays within a single fitness domain. All of these corpora, moreover, are curated offline. GuideMe is the first to combine four domains, model the full instruct-observe-correct cycle, and evaluate it under an online streaming protocol.

3 The GuideMe Benchmark

In this section, we present the GuideMe benchmark, covering task definition (Sec. 3.1), annotation pipeline (Sec. 3.2), dataset statistics (Sec. 3.3), inference pipeline (Sec. 3.4), and evaluation protocol (Sec. 3.5).

3.1 Task Definition

We formulate interactive procedural assistance as an online streaming task under a strict causal setting. Given a continuous video stream of user execution, the assistant observes frames sequentially and can only access visual evidence up to the current timestamp, without any future information. At each time step t, the assistant takes as input the recent video within a fixed sliding window and the full dialogue history accumulated so far. Based on this partial observation, it must decide whether to remain silent or to generate a guidance response.

If a response is required, the assistant produces one of three types of output: (i) a *next-step instruction* directing the upcoming action, (ii) *completion feedback* confirming that the current action has been successfully finished, or (iii) *corrective feedback* when an execution error is detected, specifying both the mistake and how to fix it. The user's subsequent action then updates the streaming context, forming a closed-loop interaction within a single session.

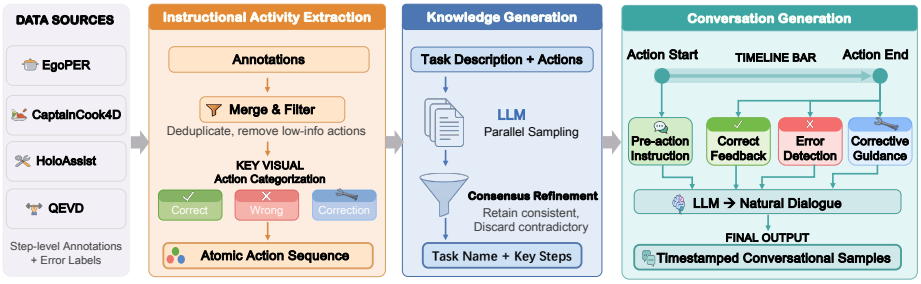


Fig. 2: Annotation pipeline of GuideMe. The process integrates three stages: instructional activity extraction with error categorization, knowledge generation from the extracted activities, and conversation generation.

3.2 Annotation Pipeline

We construct GuideMe through a three-stage automated annotation pipeline, illustrated in Fig. 2. Our data sources include EgoPER [22], CaptainCook4D [35], HoloAssist [46], and QEVD [34]. They span cooking, object manipulation, daily-life tasks, and fitness, and provide fine-grained step-level annotations of actions, mistakes, and corrective behaviors. Starting from these annotations, the pipeline proceeds from activity extraction to knowledge-guided dialogue generation.

Instructional Activity Extraction. We first extract a complete and ordered set of task instructions from the raw step-level annotations, where each instruction corresponds to an independent *atomic action* that can be executed, verified, or queried within a single dialogue turn. For datasets with fine-grained annotations, we merge duplicated or consecutive repeated actions and filter low-information ones (*e.g.*, *hold*, *rotate*) to ensure each entry conveys meaningful task progress. Each resulting action is categorized as a **correct action** (following the intended procedure), a **wrong action** (deviating from the workflow), or a **correction action** (rectifying a prior mistake), according to the dataset annotations. These labels are inherited from each source's task graph or human annotations, which already encode parallel valid orderings. This categorization enables modeling of both ideal procedural execution and realistic error-and-recovery behaviors, which form the hallmark of our closed-loop evaluation.

Knowledge Generation. Given the filtered instructional activity sequence from the extraction step and the original task description, we instruct the LLM to (1) infer a descriptive task name and (2) generate high-level procedural knowledge in the form of numbered key steps. Because the input actions may still include erroneous or corrective steps, the prompt emphasizes that the LLM must rely on commonsense reasoning to identify essential and logically consistent ones.

To improve robustness, we perform parallel sampling (N=10) and aggregate all candidates in a second-round refinement prompt, where steps consistently appearing across candidates are retained and contradictory ones discarded.

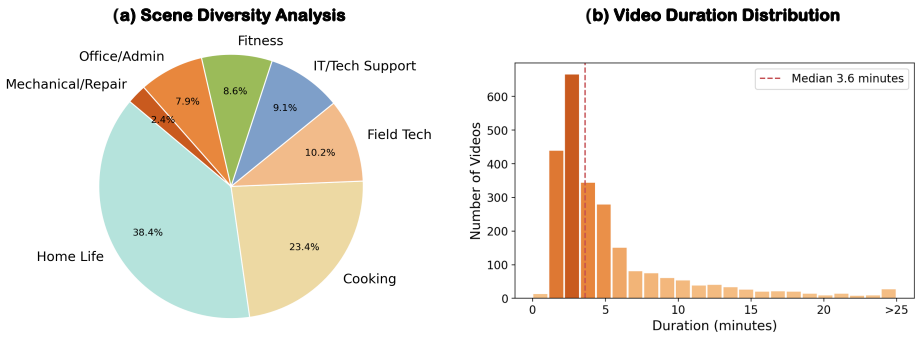


Fig. 3: Dataset statistics of GuideMe. (a) Scene Diversity: Distribution across seven categories, dominated by Home Life (38.4%) and Cooking (23.4%), with technical domains (Field Tech and IT Support) comprising 38.2%. (b) Duration Distribution: Temporal spread of 2,458 videos with a median of 3.6 minutes. The distribution includes long-form sequences up to 41.2 minutes, providing challenging samples for evaluating long-range temporal reasoning.

Conversation Generation. In the final stage, we generate instructional dialogues aligned with the *atomic action* boundaries, guided by the task description and generated procedural knowledge. For each action, the assistant produces utterances at two temporal points: (1) at the *start*, a pre-action instruction directs the upcoming step; (2) at the *end*, the assistant issues either correct execution feedback confirming successful completion, or error detection with corrective guidance that identifies the mistake at the end timestamp with explicit corrective instructions for the proper procedure.

The source of correction instructions varies by dataset. HoloAssist provides explicit correction annotations, while CaptainCook4D and EgoPER require deriving corrections from their task graphs: when an action deviates from the expected node, we identify the discrepancy and use the correct subsequent step as the corrective instruction.

We serialize each video's atomic instructions in temporal order, with timestamps attached to each structured entry (*e.g.*, "`[05:32]` *Tighten the screw on the left panel*"). Because long videos may contain hundreds of entries, we split the sequence into *clips* of at most 200 entries, each conditioned on global context (task summary and high-level steps) and local context (fine-grained sub-steps in the current temporal range). The LLM then transforms these structured clips into natural conversational dialogues.

3.3 Dataset Statistics and Analysis

The final dataset contains 2,458 video instances with a combined duration of 223.7 hours, drawn from CaptainCook4D [35] (295 videos), EgoPER [22] (280), QEVD [34] (212), and HoloAssist [46] (1,671). Video lengths range from 0.5 to 41.2 minutes (average 5.5 min, median 3.6 min), yielding 47,775 streaming

interaction samples across cooking, daily tasks, fitness, and embodied assistance, of which 9,876 form the held-out test set used for all evaluations. GuideMe is split into **GuideMe-Train** (1,985 videos, 177.0 hours) and **GuideMe-Test** (473 videos, 46.7 hours); all reported evaluations use GuideMe-Test, while GuideMe-Train supports task-specific adaptation and is used only for the fine-tuned Qwen3-VL-8B result in Tab. 2.

As shown in Fig. 3, the domain distribution bridges simple daily routines and complex professional tasks: Home Life and Cooking form the foundational core, while Field Tech and IT Support provide the variety required for generalized evaluation. Temporally, the high density around the 3.6-minute median captures concise goal-oriented actions, while the extension to 41.2 minutes provides the depth needed for evaluating long-range reasoning and multi-step planning.

3.4 Inference Pipeline

We design an inference pipeline that mirrors the constraints of real-time interactive assistance. A single task-guidance query is issued once at the start of the video; thereafter the model consumes the stream strictly in chronological order and must *proactively* decide when to intervene, with no external trigger. This departs fundamentally from question-answering benchmarks [23, 33, 48], which issue queries at fixed timestamps and only evaluate response timing; our pipeline instead tests whether a model can determine *when* to speak at all.

Inference is run at discrete timestamps, for which we consider two sampling strategies. **Dense** sampling queries the model at a fixed 1 s interval across the entire stream, so silent timestamps vastly outnumber the ground-truth interventions. **Anchor-Based** sampling (our default) instead queries at every ground-truth intervention plus an equal number of silent timestamps, yielding a balanced mix of speak and stay-silent cases. We compare the two in Sec. 4.4.

At each timestamp, the model observes the most recent 60 seconds of video through a sliding window [29, 30], together with the full textual dialogue history (*i.e.*, all previous assistant utterances and their timestamps). The sliding window bounds visual memory and enforces the online setting. Given this context, the model must decide whether to remain silent or intervene. It outputs the single word `"Silent"` when no guidance is needed, and generates the appropriate instruction or feedback once sufficient evidence is available (*e.g.*, when the user completes an action or commits an error).

3.5 Evaluation Protocol

Evaluating interactive streaming guidance requires assessing not only *what* the model says but also *when* it chooses to speak or remain silent. Traditional streaming metrics such as TimeDiff [9] presume a rigid one-to-one alignment to the reference, and LM-PPL [9] needs token probabilities that closed API models do not expose. Crucially, none of them evaluates the silence-versus-speak decision central to proactive guidance. We address this with a three-component evaluation framework: (1) temporal-semantic bipartite matching for sequence-level alignment,

(2) LLM-as-a-Judge scoring for content quality, and (3) behavioral classification for instance-level intervention timing.

Temporal-Semantic Bipartite Matching. We define the generated sequence as $\mathcal{G} = \{(g_i, t_i)\}_{i=1}^{N}$ and the reference sequence as $\mathcal{R} = \{(r_j, \hat{t}_j)\}_{j=1}^{M}$, where g_i and r_j denote the textual content of the i-th generated and j-th reference response, and t_i, $\hat{t}_j$ are their corresponding timestamps. To ensure strict temporal localization, we partition the video into K segments $\{\mathcal{S}_k\}_{k=1}^{K}$ based on the midpoints of consecutive reference timestamps: $B_j = \frac{\hat{t}_j + \hat{t}_{j+1}}{2}$. This partitioning ensures that each segment $\mathcal{S}_k$ contains exactly one active (non-silent) reference event, providing a constrained search space for matching.

Within each segment $\mathcal{S}_k$, we seek the optimal assignment between $\mathcal{G}$ and $\mathcal{R}$ by solving a minimum weight bipartite matching problem, inspired by [8, 10]. We define a cost matrix $\mathbf{C}$, where the cost $C_{i,j}$ for pairing g_i with r_j integrates textual alignment and temporal distance: $C_{i,j} = \mathcal{L}_{\text{text}}(g_i, r_j) + \mathcal{L}_{\text{dist}}(t_i, \hat{t}_j)$. The Textual Cost $\mathcal{L}_{\text{text}}$ is:

$$\mathcal{L}_{\text{text}}(g_i, r_j) = \begin{cases} 0, & \text{if } g_i, r_j \in \text{Silence} \\ 1 - |\cos(\mathbf{e}_i, \hat{\mathbf{e}}_j)|, & \text{otherwise} \end{cases} \tag{1}$$

where $\mathbf{e}_i, \hat{\mathbf{e}}_j$ are embeddings from a pre-trained Sentence-Transformer. The "otherwise" case captures all mismatches, including the model generating text during a silent ground-truth window or vice versa. The temporal Distance Cost $\mathcal{L}_{\text{dist}}$ penalizes time-shifted responses via a Gaussian-decay function:

$$\mathcal{L}_{\text{dist}}(t_i, \hat{t}_j) = 1 - \exp\left(-\sigma |t_i - \hat{t}_j|^2\right), \tag{2}$$

Let $\mathcal{A}$ denote the optimal assignment obtained from the bipartite matching. To reward models that capture the essence of an action even if the wording differs from the reference, we define soft versions of precision and recall that use continuous semantic scores $S_{i,j}$ as weights instead of binary counts:

$$\text{sPrecision} = \frac{\sum_{(i,j)\in\mathcal{A}} S_{i,j}}{\tilde{N}_{\text{gen}}}, \quad \text{sRecall} = \frac{\sum_{(i,j)\in\mathcal{A}} S_{i,j}}{\tilde{M}_{\text{ref}}}, \tag{3}$$

where $S_{i,j} = |\cos(\mathbf{e}_i, \hat{\mathbf{e}}_j)|$ denotes the cosine similarity between embeddings $\mathbf{e}_i$ and $\hat{\mathbf{e}}_j$. The terms $\tilde{N}_{\text{gen}}$ and $\tilde{M}_{\text{ref}}$ represent the total count of active talk events in the generated sequence $\mathcal{G}$ and reference sequence $\mathcal{R}$, respectively. Their harmonic mean gives the Soft-F_1 score: $\text{sF}_1 = \frac{2 \cdot \text{sPrecision} \cdot \text{sRecall}}{\text{sPrecision} + \text{sRecall}}$, which provides a unified, threshold-free measure that balances semantic accuracy with temporal coverage.

LLM-as-a-Judge. After bipartite matching identifies the best-aligned prediction-reference pairs, we further evaluate the response quality of each matched pair using an LLM-as-a-Judge [59]. The judge is given the query context, the model prediction, and the reference answer, and assigns an integer score from 0 to 5

based on relevance, factual plausibility, and causal reasoning. We linearly rescale the score to 0–100 and average it over all matched pairs, yielding $\textbf{Score}_m$. Unlike sPrecision, sRecall, and sF_1, which measure temporal-semantic coverage, Score_m isolates the quality of model responses at successfully matched intervention points.

Behavioral Classification. While bipartite matching evaluates sequence-level alignment, we also need to assess whether the model intervenes at the right individual moments. For each video, we construct two complementary evaluation instances: (1) **assistant-anchored instances**, where each ground-truth intervention timestamp serves as a temporal anchor and the model must produce the appropriate guidance; and (2) **silent instances**, sampled from intervals between consecutive interventions (at least 5 s from any anchor), where the model should remain silent. The two types are balanced in number.

Each prediction is categorized into one of four mutually exclusive outcomes: Correct Silent (**CS**, ↑), where the model correctly remains silent for a silent instance; False Alarm (**FA**, ↓), where the model speaks during a silent instance; No Response (**NR**, ↓), where the model fails to respond at an intervention-anchored timestamp; and Partly Correct (**PC**, ↑), where the model produces a response at an intervention-anchored timestamp. These categories capture two complementary timing failures: *over-intervention*, reflected by FA, and *under-intervention*, reflected by NR.

We further summarize these behavior outcomes with an aggregate response-behavior metric, the **Score**. CS is assigned a score of 100, as the model correctly remains silent; FA and NR are assigned a score of 0, corresponding to false intervention and erroneous silence, respectively; and each PC instance receives an LLM-as-a-Judge [59] quality score, linearly scaled to the range of 0–100. The **Score** reported in our tables is the average of these values over all evaluation instances, rewarding both correct silence and high-quality interventions while penalizing false alarms and missed responses.

4 Experiments

In this section, we present experiments on GuideMe. We describe the experimental setup (Sec. 4.1), report baseline and visual results (Sec. 4.2 and Sec. 4.3), and conduct ablation studies on inference settings (Sec. 4.4). Additional per-domain breakdowns and prompt details are provided in the supplementary material.

4.1 Experiment Setup

We evaluate three categories of zero-shot baselines on GuideMe using the Inference Pipeline (Sec. 3.4): (1) proprietary MLLMs (Doubao-Seed-1.8 [7], GPT-5.2 [38], Gemini 3 Pro [15], Gemini 3.1 Pro [16]), (2) open-source MLLMs (Qwen2.5-7B [3], Qwen3-VL-8B [2], Qwen3-VL-30B-A3B [2], Qwen3.5-397B-A17B [42]), and (3) open-source streaming MLLMs (VideoLLM-online [9], Dispider [36],

Table 2: Main evaluation results on GuideMe. We assess model responses from two complementary perspectives: *Temporal Alignment* evaluates whether the model provides guidance at the appropriate time with relevant content, including Score_m for the LLM-as-a-Judge quality of matched prediction-reference pairs. *Response Behavior* categorizes each instance into Correct Silent (**CS**, ↑), No Response (**NR**, ↓), False Alarm (**FA**, ↓), and Partly Correct (**PC**, ↑), and summarizes them with **Score**. The best result in each category is **bolded**. † denotes fine-tuning on the GuideMe training split.

Model	Param.	Temporal Alignment				Response Behavior				
		sPrecision ↑	sRecall ↑	sF1 ↑	Score_m ↑	CS ↑	NR ↓	FA ↓	PC ↑	Score ↑
Proprietary Multimodal Models										
Doubao-Seed-1.8 [7]	-	30.6	**47.5**	36.8	63.1	8.8	6.1	34.8	44.2	38.8
GPT-5.2 [38]	-	30.2	36.8	32.5	64.4	13.4	16.8	30.6	39.2	38.9
Gemini 3 Pro [15]	-	39.3	36.9	36.5	**66.6**	17.5	22.0	19.0	41.5	44.8
Gemini 3.1 Pro [16]	-	29.3	46.9	35.7	61.5	7.4	6.3	36.4	49.9	37.8
Open-source Multimodal Models										
Qwen2.5-7B [3]	7B	30.1	30.9	29.3	57.5	19.7	20.4	25.4	34.5	37.5
Qwen3-VL-8B [2]	8B	29.7	46.8	35.7	62.0	5.2	6.7	37.7	50.5	31.9
Qwen3-VL-30B-A3B [2]	30B	29.5	46.9	35.6	61.6	6.7	6.0	36.5	50.9	33.3
Qwen3.5-397B-A17B [42]	397B	38.4	20.2	21.7	64.6	35.2	37.9	7.9	19.0	**45.7**
Open-source Multimodal Streaming Models										
VideoLLM-online [9]	8B	22.3	41.1	28.7	40.8	0.0	**0.0**	43.6	**56.4**	22.6
Dispider [36]	7B	1.0	0.1	0.1	46.7	**43.6**	56.3	**0.0**	0.1	43.6
LiveStar [55]	8B	20.7	19.0	19.1	14.4	21.2	24.4	22.2	32.3	25.0
MMDuet2 [47]	3B	3.3	0.2	0.4	43.3	41.7	57.9	0.2	0.2	41.8
Fine-tuned on GuideMe Train Set										
Qwen3-VL-8B† [2]	8B	**40.3**	43.7	**39.7**	58.1	19.5	20.6	24.5	35.3	42.1

LiveStar [55], MMDuet2 [47]). We additionally report Qwen3-VL-8B fine-tuned on the GuideMe training split to assess whether task-specific adaptation improves streaming guidance. By default we provide the ground-truth dialogue history at each step so that scores isolate each model's single-step timing and content decisions from error accumulation across turns. We report the two metric families defined in Sec. 3.5: *Temporal Alignment* (sPrecision, sRecall, sF_1, Score_m) and *Response Behavior* (CS, FA, NR, PC, Score).

4.2 Quantitative Results

Tab. 2 reports aggregate results on the full GuideMe test set, while Tab. 3 further stratifies representative models by ground-truth response category to analyze where failures occur. Overall, no zero-shot baseline provides reliable streaming coaching: models either intervene too often, remain silent too often, or fail to detect when an intervention should change from routine guidance to error correction.

Response frequency trades off with quality. The results show a clear split between aggressive models, which respond frequently but raise many false alarms,

Table 3: Per-category breakdown in the main results. Metrics are stratified by ground-truth response category on the full test set.

Model	Instruction				Error				Correction			
	sF_1 ↑	NR↓	PC↑	Score↑	sF_1 ↑	NR↓	PC↑	Score↑	sF_1 ↑	NR↓	PC↑	Score↑
Gemini 3 Pro [15]	42.0	33.3	66.7	33.0	33.8	39.6	60.4	20.2	34.3	41.7	58.3	23.1
Doubao-Seed-1.8 [7]	49.6	8.9	91.1	41.9	41.9	14.8	85.2	26.2	42.7	15.0	85.0	30.0

and conservative models, which avoid false alarms but miss required interventions. For example, Qwen3-VL-30B-A3B reaches 50.9% PC but only 33.3 Score, while Qwen3.5 obtains the highest Score (45.7) largely by speaking less. Score_m further shows that matched-response quality is relatively close among non-streaming MLLMs (57.5–66.6), suggesting that deciding *when* to intervene is the more limiting factor.

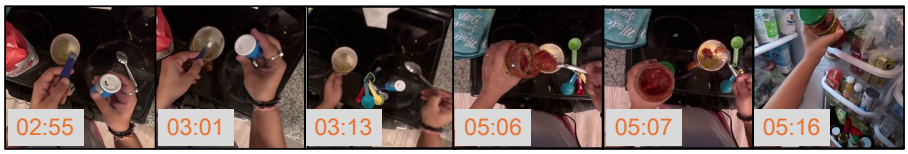
Scaling and streaming pretraining are insufficient. They further show that general scaling and streaming-oriented pretraining do not solve the task. Qwen3-VL-8B and Qwen3-VL-30B-A3B obtain nearly identical sF1 scores (35.7 vs. 35.6), while the much larger Qwen3.5-397B-A17B performs worse on temporal alignment (sF1 21.7). Streaming-trained models show the same issue more severely: VideoLLM-online over-responds (FA 43.6, nearly never silent), whereas Dispider and MMDuet2 collapse to near-silence (sF1 < 1, NR $> 55\%$). This indicates that knowing *when* to speak, not only how to describe video content, is the core missing capability.

Fine-tuning improves alignment but shifts behavior. Fine-tuning on GuideMe improves Qwen3-VL-8B at the sequence level, raising sPrecision from 29.7 to 40.3 and sF1 from 35.7 to 39.7. At the instance level, however, the model becomes more conservative: CS increases from 5.2 to 19.5 and FA decreases from 37.7 to 24.5, indicating better silence calibration, but PC also drops from 50.5 to 35.3 and NR rises from 6.7 to 20.6. At the content level, Score improves from 31.9 to 42.1, yet remains below the strongest zero-shot models. Thus, the training split supports task-specific adaptation, but robust closed-loop coaching still requires jointly improving temporal alignment, intervention decisions, and response quality.

Error correction remains the hardest behavior. Tab. 3 shows that both Gemini 3 Pro and Doubao-Seed-1.8 perform best on Instruction and degrade on Error and Correction. For Gemini 3 Pro, sF1 drops from 42.0 on Instruction to 33.8 on Error and 34.3 on Correction, while Score drops from 33.0 to 20.2 and 23.1. Doubao follows the same trend, with sF1 dropping from 49.6 to 41.9 and 42.7. Current MLLMs are therefore stronger at routine step guidance than at timely error detection and corrective intervention.

4.3 Qualitative Results

Fig. 4 shows a representative cooking example (a microwave mug pizza), comparing the ground-truth guidance with predictions from six models at two error-correction intervention points. The example highlights that models often continue



User

I'm going to make a microwave mug pizza. Can you walk me through the steps and let me know if I make any mistakes along the way?

• • •

Assistant (03:01)

GT: Hold on a second. It looks like you're about to add the salt, but please make sure you're using a clean, dry measuring spoon so the ingredients don't stick.

Doubao-Seed-1.8: Great, herbs are added. Now go ahead and add a tablespoon of mozzarella cheese to the mug.

GPT-5.2: *[Silent]*

Gemini 3.1 Pro: Perfect. Now add 1/16 teaspoon of baking soda.

Qwen2.5-7B: Now, add 2 tablespoons of pizza sauce.

MMDuet2: *[Silent]*

Qwen3-VL-8B (SFT): That's the right amount of herbs. Now, add 1/16 teaspoon of black pepper to the mix.

• • •

Assistant (05:07)

GT: Just a heads-up: it looks like you're reaching for the sauce, but make sure you use marinara, not salsa.

Doubao-Seed-1.8: Great, you've grabbed the pizza sauce. Now add 2 tablespoons of it to the mug.

GPT-5.2: Now add 1 tablespoon of pizza sauce on top, do not stir it in. Spread it lightly.

Gemini 3.1 Pro: Now, spoon 1 tablespoon of marinara sauce onto the surface of the dough and spread it around.

Qwen2.5-7B: Microwave for 1 minute and 20 seconds.

MMDuet2: *[Silent]*

Qwen3-VL-8B (SFT): It looks like you're trying to add the tomato sauce, but you're spilling some of it while you're at it.

• • •

Fig. 4: Qualitative example. A microwave mug pizza task with two error-correction interventions. The ground-truth (GT) response is shown in orange alongside predictions from six representative models.

routine instruction even when corrective feedback is needed. At 03:01, most responding models advance to later recipe steps instead of warning about the measuring spoon, while GPT-5.2 and MMDuet2 remain silent. At 05:07, Gemini 3.1 Pro correctly identifies marinara sauce, but several models either give generic next-step guidance or miss the correction.

Table 4: Effect of sampling and dialogue-history settings. The default setting (*) uses Anchor-Based sampling with ground-truth (GT) dialogue history. We compare Dense and Anchor-Based sampling, both defined in Sec. 3.4, under GT dialogue history, and further evaluate **w/o GT** by replacing GT history with the model's own predictions. Results are evaluated on a randomly sampled 100-video subset (25 videos from each source dataset). The best result per model in each column is **bolded**.

Model	Setting	Temporal Alignment			Response Behavior				
		sPrecision ↑	sRecall ↑	sF1 ↑	CS ↑	NR ↓	FA ↓	PC ↑	Score ↑
Doubao-Seed-1.8 [7]	Dense	23.1	26.3	21.9	**89.3**	**4.4**	**4.7**	1.6	**90.3**
	Anchor*	30.6	**47.9**	**36.9**	7.8	6.1	34.3	**51.9**	38.6
	w/o GT	**43.5**	14.2	20.0	33.9	50.9	8.4	6.8	38.3
GPT-5.2 [38]	Dense	10.7	**40.6**	14.8	**47.4**	**3.1**	46.4	3.1	**49.0**
	Anchor*	30.4	37.0	**32.8**	12.7	17.1	29.8	**40.4**	38.8
	w/o GT	**31.7**	17.9	21.6	26.1	43.3	**16.1**	14.4	34.1
Gemini 3 Pro [15]	Dense	10.0	**53.2**	14.5	**45.4**	**2.1**	48.4	4.1	**47.9**
	Anchor*	**39.7**	35.6	**35.7**	18.1	23.1	19.0	**39.9**	43.7
	w/o GT	35.7	25.1	28.0	27.6	36.4	**14.6**	21.3	39.1
Qwen2.5-7B [3]	Dense	29.1	15.4	18.6	**90.5**	5.5	**3.3**	0.7	**91.0**
	Anchor*	**32.7**	**33.0**	**31.4**	19.2	**19.2**	25.5	**36.2**	37.3
	w/o GT	16.3	25.5	17.1	13.8	27.6	23.4	35.1	26.3

4.4 Ablation Studies

We conduct four ablations in two groups. Tab. 4 studies three inference-protocol settings on a randomly sampled 100-video subset (25 videos from each source dataset). The *Dense* and *Anchor* settings follow Sec. 3.4, and both use ground-truth (GT) dialogue history; Dense queries every 1 s, while Anchor queries intervention timestamps and balanced silent timestamps. To further evaluate fully autonomous context, the *w/o GT* setting keeps Anchor-Based sampling but replaces GT dialogue history with the model's own predictions. Tab. 5 then analyzes Gemini 3 Pro with two sensitivity tests: under Anchor-Based sampling, we vary the sliding window size to control the visible visual context; under Dense sampling, we vary the sampling interval to control query frequency.

Anchor-Based vs. Dense Sampling. Dense sampling exposes a strong silence imbalance: Doubao and Qwen2.5 collapse to silence (CS+NR >93%), inflating Score through the silent bonus (*e.g.*, Doubao 38.6→90.3) while their coaching ability (PC) collapses. GPT-5.2 and Gemini 3 Pro instead over-respond at nearly half of all timestamps (FA >46%). Anchor-Based sampling avoids this distributional bias, so we adopt it as the default.

Ground-Truth vs. Predicted Context. Removing GT context consistently degrades every model: sF1 and Score drop while missed interventions (NR) rise sharply (*e.g.*, Doubao sF1 36.9 → 20.0, NR 6.1 → 50.9; Qwen2.5 Score 37.3 → 26.3), as early mistakes compound into an increasingly unreliable dialogue history. This confirms GT history as the fairer evaluation protocol and quantifies the headroom that remains for fully autonomous deployment.

Table 5: Sensitivity analysis with Gemini 3 Pro. *Sliding Window Size* under the Anchor-Based sampling strategy (default: 60 s window) and *Sampling Interval* under the Dense sampling strategy (default: 1 s interval). The best result within each block is **bolded**.

Setting	Temporal Alignment			Response Behavior				
	sPrecision ↑	sRecall ↑	sF1 ↑	CS ↑	NR ↓	FA ↓	PC ↑	Score ↑
Sliding Window Size under Anchor-Based Sampling								
30 s	31.6	40.1	34.8	13.7	22.8	23.1	40.2	40.3
60 s	33.8	**40.3**	36.5	17.5	22.0	19.0	**41.5**	44.8
120 s	**35.1**	39.8	**37.1**	**19.9**	**21.9**	**16.7**	**41.5**	**47.6**
Sampling Interval under Dense Sampling								
0.5 s	3.6	**56.5**	5.2	44.9	**1.0**	51.9	2.0	46.2
1 s	10.0	53.2	14.5	**45.4**	2.1	**48.4**	4.1	**47.9**
2 s	**11.6**	54.0	**16.8**	35.2	2.6	52.8	**9.3**	40.8

Sliding Window Size. Tab. 5 shows that the 60 s visual window provides a reasonable balance between fidelity and cost. Shrinking it to 30 s lowers Score by 4.5 as procedural context is lost, while enlarging it to 120 s yields only a marginal +2.8 gain at 2× token cost.

Sampling Interval. A 2 s interval drops Score by 7.1 as it misses fast actions, whereas 0.5 s gives comparable Score at 2× cost, confirming 1 s as the default. Overall, GuideMe's conclusions remain robust to these choices rather than being artifacts of a particular sampling configuration.

5 Conclusion

This paper presents GuideMe, the first multi-domain streaming benchmark for evaluating closed-loop interactive task guidance. GuideMe comprises 2,458 videos spanning 223.7 hours across four domains with 47,775 interaction samples, built through a scalable three-stage annotation pipeline and assessed with a three-component evaluation framework: Temporal-Semantic Bipartite Matching for sequence-level alignment, Behavioral Classification for instance-level intervention timing, and LLM-as-a-Judge for content quality. Across proprietary, open-source, and streaming models, we find that current MLLMs polarize into two failure modes: aggressive models intervene frequently but raise many false alarms, whereas conservative ones stay silent and miss most interventions. As a result, no model coaches reliably: the best-balanced Gemini 3 Pro still misses 22% of needed interventions, while Qwen3.5 obtains the highest Score (45.7) despite correctly responding at only 19% of intervention-anchored instances. This exposes a fundamental gap in the instruct-observe-correct cycle: models can *describe* procedures, but they cannot yet *coach* them. We hope GuideMe serves as a practical testbed for developing AI assistants that move beyond passive narration toward genuinely interactive procedural coaching.

References

1. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., Ge, W., Guo, Z., Huang, Q., Huang, J., Huang, F., Hui, B., Jiang, S., Li, Z., Li, M., Li, M., Li, K., Lin, Z., Lin, J., Liu, X., Liu, J., Liu, C., Liu, Y., Liu, D., Liu, S., Lu, D., Luo, R., Lv, C., Men, R., Meng, L., Ren, X., Ren, X., Song, S., Sun, Y., Tang, J., Tu, J., Wan, J., Wang, P., Wang, P., Wang, Q., Wang, Y., Xie, T., Xu, Y., Xu, H., Xu, J., Yang, Z., Yang, M., Yang, J., Yang, A., Yu, B., Zhang, F., Zhang, H., Zhang, X., Zheng, B., Zhong, H., Zhou, J., Zhou, F., Zhou, J., Zhu, Y., Zhu, K.: Qwen3-vl technical report (2025), https://arxiv.org/abs/2511.21631 1, 3, 4
2. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., Ge, W., Guo, Z., Huang, Q., Huang, J., Huang, F., Hui, B., Jiang, S., Li, Z., Li, M., Li, M., Li, K., Lin, Z., Lin, J., Liu, X., Liu, J., Liu, C., Liu, Y., Liu, D., Liu, S., Lu, D., Luo, R., Lv, C., Men, R., Meng, L., Ren, X., Ren, X., Song, S., Sun, Y., Tang, J., Tu, J., Wan, J., Wang, P., Wang, P., Wang, Q., Wang, Y., Xie, T., Xu, Y., Xu, H., Xu, J., Yang, Z., Yang, M., Yang, J., Yang, A., Yu, B., Zhang, F., Zhang, H., Zhang, X., Zheng, B., Zhong, H., Zhou, J., Zhou, F., Zhou, J., Zhu, Y., Zhu, K.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025) 4, 10, 11
3. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923 (2025) 10, 11, 14
4. Bao, Y., Yu, K., Zhang, Y., Storks, S., Bar-Yossef, I., de la Iglesia, A., Su, M., Zheng, X., Chai, J.: Can foundation models watch, talk and guide you step by step to make a cake? In: EMNLP. pp. 12325–12341 (2023) 5
5. Ben-Shabat, Y., Yu, X., Saleh, F., Campbell, D., Rodriguez-Opazo, C., Li, H., Gould, S.: The ikea asm dataset: Understanding people assembling furniture through actions, objects and pose. In: WACV. pp. 847–859 (2021) 4, 5
6. ByteDance: Doubao-seed-1.6. Available at Volcengine ARK Platform (2025) 3
7. ByteDance: Doubao-seed-1.8. Available at Volcengine ARK Platform (2026) 3, 10, 11, 12, 14
8. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: ECCV. pp. 213–229. Springer (2020) 9
9. Chen, J., Lv, Z., Wu, S., Lin, K.Q., Song, C., Gao, D., Liu, J.W., Gao, Z., Mao, D., Shou, M.Z.: Videollm-online: Online video large language model for streaming video. In: CVPR. pp. 18407–18418 (2024) 4, 8, 10, 11
10. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: CVPR. pp. 1290–1299 (2022) 9
11. Damen, D., Doughty, H., Farinella, G.M., Fidler, S., Furnari, A., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., et al.: The epic-kitchens dataset: Collection, challenges and baselines. IEEE TPAMI **43**(11), 4125–4141 (2020) 2, 4, 5
12. Ding, X., Wu, H., Yang, Y., Jiang, S., Zhang, Q., Bai, D., Chen, Z., Cao, T.: Streammind: Unlocking full frame rate streaming video dialogue through event-gated cognition. In: ICCV. pp. 13448–13459 (2025) 4
13. Fu, C., Dai, Y., Luo, Y., Li, L., Ren, S., Zhang, R., Wang, Z., Zhou, C., Shen, Y., Zhang, M., et al.: Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In: CVPR. pp. 24108–24118 (2025) 4

14. Fu, L., Yang, B., Kuang, Z., Song, J., Li, Y., Zhu, L., Luo, Q., Wang, X., Lu, H., Huang, M., Li, Z., Tang, G., Shan, B., Lin, C., Liu, Q., Wu, B., Feng, H., Liu, H., Huang, C., Tang, J., Chen, W., Jin, L., Liu, Y., Bai, X.: Ocrbench v2: An improved benchmark for evaluating large multimodal models on visual text localization and reasoning (2024) 4
15. Google DeepMind: Gemini 3 pro model card. Available at Google DeepMind Model Cards (2025) 4, 10, 11, 12, 14
16. Google DeepMind: Gemini 3.1 pro model card. Available at Google DeepMind Model Cards (2026) 1, 3, 10, 11
17. Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., et al.: Ego4d: Around the world in 3,000 hours of egocentric video. In: CVPR. pp. 18995–19012 (2022) 5
18. Grauman, K., Westbury, A., Torresani, L., Kitani, K., Malik, J., Afouras, T., Ashutosh, K., Baiyya, V., Bansal, S., Boote, B., et al.: Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In: CVPR. pp. 19383–19400 (2024) 5
19. Huang, Y., Chen, G., Xu, J., Zhang, M., Yang, L., Pei, B., Zhang, H., Dong, L., Wang, Y., Wang, L., et al.: Egoexolearn: A dataset for bridging asynchronous ego-and exo-centric view of procedural activities in real world. In: CVPR. pp. 22072–22086 (2024) 5
20. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024) 1, 4
21. Jang, Y., Sullivan, B., Ludwig, C., Gilchrist, I., Damen, D., Mayol-Cuevas, W.: Epic-tent: An egocentric video dataset for camping tent assembly. In: ICCVW. pp. 0–0 (2019) 4, 5
22. Lee, S.P., Lu, Z., Zhang, Z., Hoai, M., Elhamifar, E.: Error detection in egocentric procedural task videos. In: CVPR. pp. 18655–18666 (2024) 2, 4, 5, 6, 7
23. Lin, J., Fang, Z., Chen, C., Wan, Z., Luo, F., Li, P., Liu, Y., Sun, M.: Streamingbench: Assessing the gap for mllms to achieve streaming video understanding. arXiv preprint arXiv:2411.03628 (2024) 4, 8
24. Lin, Y., Zhang, J., Wang, Q., Ye, H., Fu, Y., Liu, Y., Li, H.H., Chen, Y.: Hippomm: Hippocampal-inspired multimodal memory for long audiovisual event understanding. In: CVPR. pp. 5968–5977 (2026) 4
25. Liu, F., Liu, Y., Kong, Y., Xu, K., Zhang, L., Yin, B., Hancke, G., Lau, R.: Referring image segmentation using text supervision. In: ICCV. pp. 22124–22134 (2023) 4
26. Liu, F., Liu, Y., Xu, K., Hancke, G.P., Lau, R.W.: Gensplat: Bridging the generalization gap in 3dgs language comprehension. In: CVPR. pp. 5221–5231 (2026) 4
27. Liu, F., Liu, Y., Xu, K., Ye, S., Hancke, G.P., Lau, R.W.: Language-guided salient object ranking. In: CVPR. pp. 29803–29813 (2025) 4
28. Long, L., He, Y., Ye, W., Pan, Y., Lin, Y., Li, H., Zhao, J., Li, W.: Seeing, listening, remembering, and reasoning: A multimodal agent with long-term memory. In: ICLR (2026) 4
29. Lu, X., Bo, Y., Chen, J., Li, S., Guo, X., Guan, H., Liu, F., Xu, D., Sun, P., Sun, H., et al.: Aura: Always-on understanding and real-time assistance via video streams. arXiv preprint arXiv:2604.04184 (2026) 8
30. Lu, X., Guan, H., Bo, Y., Chen, J., Guo, X., Li, S., Liu, F., Sun, P., Li, X., Zhang, W., et al.: Phostream: Benchmarking real-world streaming for omnimodal assistants in mobile scenarios. In: ICML (2026) 4, 8

31. Mathew, M., Karatzas, D., Jawahar, C.: Docvqa: A dataset for vqa on document images. In: WACV. pp. 2200–2209 (2021) 4
32. Miech, A., Zhukov, D., Alayrac, J.B., Tapaswi, M., Laptev, I., Sivic, J.: Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In: ICCV. pp. 2630–2640 (2019) 2, 5
33. Niu, J., Li, Y., Miao, Z., Ge, C., Zhou, Y., He, Q., Dong, X., Duan, H., Ding, S., Qian, R., et al.: Ovo-bench: How far is your video-llms from real-world online video understanding? In: CVPR. pp. 18902–18913 (2025) 4, 8
34. Panchal, S., Bhattacharyya, A., Berger, G., Mercier, A., Böhm, C., Dietrichkeit, F., Pourreza, R., Li, X., Madan, P., Lee, M., et al.: What to say and when to say it: Live fitness coaching as a testbed for situated interaction. NeurIPS **37**, 75853–75882 (2024) 2, 4, 5, 6, 7
35. Peddi, R., Arya, S., Challa, B., Pallapothula, L., Vyas, A., Gouripeddi, B., Zhang, Q., Wang, J., Komaragiri, V., Ragan, E., et al.: Captaincook4d: A dataset for understanding errors in procedural activities. NeurIPS **37**, 135626–135679 (2024) 2, 4, 5, 6, 7
36. Qian, R., Ding, S., Dong, X., Zhang, P., Zang, Y., Cao, Y., Lin, D., Wang, J.: 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. pp. 24045–24055 (2025) 4, 10, 11
37. Sener, F., Chatterjee, D., Shelepov, D., He, K., Singhania, D., Wang, R., Yao, A.: Assembly101: A large-scale multi-view video dataset for understanding procedural activities. In: CVPR. pp. 21096–21106 (2022) 2, 4, 5
38. Singh, A., Fry, A., Perelman, A., Tart, A., Ganesh, A., El-Kishky, A., McLaughlin, A., Low, A., Ostrow, A., Ananthram, A., et al.: Openai gpt-5 system card. arXiv preprint arXiv:2601.03267 (2025) 10, 11, 14
39. Sun, P., Lang, S., Wu, D., Ding, Y., Feng, K., Liu, H., Ye, Z., Liu, R., Liu, Y.H., Wang, J., et al.: Spacevista: All-scale visual spatial reasoning from mm to km. In: ICML (2026) 4
40. Sun, P., Lu, X., Liu, H., Bo, Y., Wu, D., Guan, H., Cai, M., Chen, J., Guo, X., Li, S., et al.: X-stream: Exploring mllms as multiplexers for multi-stream understanding. In: ECCV (2026) 4
41. Tang, Y., Ding, D., Rao, Y., Zheng, Y., Zhang, D., Zhao, L., Lu, J., Zhou, J.: Coin: A large-scale dataset for comprehensive instructional video analysis. In: CVPR. pp. 1207–1216 (2019) 2, 4, 5
42. Team, Q.: Qwen3.5: Accelerating productivity with native multimodal agents (February 2026), https://qwen.ai/blog?id=qwen3.5 10, 11
43. Wang, H., Feng, B., Lai, Z., Xu, M., Li, S., Ge, W., Dehghan, A., Cao, M., Huang, P.: Streambridge: Turning your offline video large language model into a proactive streaming assistant. arXiv preprint arXiv:2505.05467 (2025) 4
44. Wang, K., Pan, J., Shi, W., Lu, Z., Ren, H., Zhou, A., Zhan, M., Li, H.: Measuring multimodal mathematical reasoning with math-vision dataset. NeurIPS **37**, 95095–95169 (2024) 4
45. Wang, W., Gao, Z., Gu, L., Pu, H., Cui, L., Wei, X., Liu, Z., Jing, L., Ye, S., Shao, J., et al.: Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. arXiv preprint arXiv:2508.18265 (2025) 4
46. Wang, X., Kwon, T., Rad, M., Pan, B., Chakraborty, I., Andrist, S., Bohus, D., Feniello, A., Tekin, B., Frujeri, F.V., et al.: Holoassist: an egocentric human interaction dataset for interactive ai assistants in the real world. In: ICCV. pp. 20270–20281 (2023) 2, 4, 5, 6, 7

47. Wang, Y., Liu, S., Wang, D., Xu, N., Wan, G., Zhang, H., Zhao, D.: Mmduet2: Enhancing proactive interaction of video mllms with multi-turn reinforcement learning. arXiv preprint arXiv:2512.06810 (2025) 3, 4, 11
48. Wang, Y., Meng, X., Wang, Y., Zhang, H., Zhao, D.: Proactivevideoqa: A comprehensive benchmark evaluating proactive interactions in video large language models. arXiv preprint arXiv:2507.09313 (2025) 5, 8
49. Wang, Y., Wang, Y., Chen, B., Wu, T., Zhao, D., Zheng, Z.: Omnimmi: A comprehensive multi-modal interaction benchmark in streaming video contexts. In: CVPR. pp. 18925–18935 (2025) 4
50. Wu, Z., Chen, X., Pan, Z., Liu, X., Liu, W., Dai, D., Gao, H., Ma, Y., Wu, C., Wang, B., et al.: Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. arXiv preprint arXiv:2412.10302 (2024) 4
51. Xu, J., Guo, Z., Hu, H., Chu, Y., Wang, X., He, J., Wang, Y., Shi, X., He, T., Zhu, X., Lv, Y., Wang, Y., Guo, D., Wang, H., Ma, L., Zhang, P., Zhang, X., Hao, H., Guo, Z., Yang, B., Zhang, B., Ma, Z., Wei, X., Bai, S., Chen, K., Liu, X., Wang, P., Yang, M., Liu, D., Ren, X., Zheng, B., Men, R., Zhou, F., Yu, B., Yang, J., Yu, L., Zhou, J., Lin, J.: Qwen3-omni technical report (2025), https://arxiv.org/abs/2509.17765 4
52. Xu, R., Xiao, G., Chen, Y., He, L., Peng, K., Lu, Y., Han, S.: Streamingvlm: Real-time understanding for infinite video streams. arXiv preprint arXiv:2510.09608 (2025) 4
53. Yang, J., Liu, S., Guo, H., Dong, Y., Zhang, X., Zhang, S., Wang, P., Zhou, Z., Xie, B., Wang, Z., et al.: Egolife: Towards egocentric life assistant. In: CVPR. pp. 28885–28900 (2025) 5
54. Yang, Z., Hu, Y., Du, Z., Xue, D., Qian, S., Wu, J., Yang, F., Dong, W., Xu, C.: Svbench: A benchmark with temporal multi-turn dialogues for streaming video understanding. arXiv preprint arXiv:2502.10810 (2025) 4
55. Yang, Z., Zhang, K., Hu, Y., Wang, B., Qian, S., Wen, B., Yang, F., Gao, T., Dong, W., Xu, C.: Livestar: Live streaming assistant for real-world online video understanding. arXiv preprint arXiv:2511.05299 (2025) 4, 11
56. Yu, T., Wang, Z., Wang, C., Huang, F., Ma, W., He, Z., Cai, T., Chen, W., Huang, Y., Zhao, Y., et al.: Minicpm-v 4.5: Cooking efficient mllms via architecture, data, and training recipe. arXiv preprint arXiv:2509.18154 (2025) 4
57. Yu, Z., Xu, D., Yu, J., Yu, T., Zhao, Z., Zhuang, Y., Tao, D.: Activitynet-qa: A dataset for understanding complex web videos via question answering. In: AAAI. pp. 9127–9134 (2019) 4
58. Zhang, R., Jiang, D., Zhang, Y., Lin, H., Guo, Z., Qiu, P., Zhou, A., Lu, P., Chang, K.W., Qiao, Y., et al.: Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In: ECCV. pp. 169–186. Springer (2024) 4
59. Zheng, L., Chiang, W.L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., et al.: Judging llm-as-a-judge with mt-bench and chatbot arena. NeurIPS **36**, 46595–46623 (2023) 9, 10
60. Zhou, L., Xu, C., Corso, J.: Towards automatic learning of procedures from web instructional videos. In: AAAI. vol. 32 (2018) 2, 4, 5