ALIVE: MAKING STREAMING VIDEO MODELS INTERACTIVE

Anonymous authors
Paper under double-blind review

ABSTRACT

Streaming video LLMs offer a natural foundation for multimodal interactive AI that continuously perceives the physical world and interacts with users. Yet existing models remain largely passive, responding to individual requests in isolation. To bridge this gap, we study how to transform streaming video LLMs from passive observers into interactive assistants. We formulate **interactive streaming video understanding** around three core capabilities: proactive monitoring, multi-task coordination, and feedback seeking. Based on this formulation, we introduce **ALIVE-234K**, a large-scale dataset for post-training, together with **FSI-Bench**, a comprehensive benchmark for evaluating interactivity in streaming video LLMs. We further investigate post-training strategies and introduce **ALIVE**, a family of models that equips offline video LLMs with rich interactive behaviors without requiring specialized model architectures. Across multiple streaming video benchmarks, ALIVE achieves highly competitive performance, preserving strong video understanding while enabling rich interactions over evolving video streams. Our results demonstrate that interactive behaviors can emerge through appropriate data and post-training, providing a path toward multimodal assistants that continuously perceive and interact with users as events unfold.

1 INTRODUCTION

Advances in large language models (LLMs) are pushing conversational AI beyond multi-turn dialogues toward interactive assistants that can respond to diverse user requests, manage multiple concurrent tasks, and, when appropriate, proactively engage users as those tasks evolve (Ye et al., 2025; Lu et al., 2024). Bringing these capabilities into the physical world points toward a long-standing vision of multimodal interactive AI: systems such as *JARVIS in Iron Man* that continuously perceive streams of video, audio, and other sensory inputs while seamlessly interacting with human users, augmenting our ability to understand and act in complex, dynamic environments.

Streaming video LLMs lie at the very heart of this vision. Recent models (Chen et al., 2024; Xiong et al., 2025; Xu et al., 2026) can continuously process incoming video, reason about past observations, describe ongoing activities, and monitor events as they occur. Yet today's streaming video LLMs remain largely passive, often responding to individual requests in isolation. They lack a critical capability of interactive assistants: the ability to maintain and revise user tasks over time and proactively engage users as the video unfolds. This gap in interactivity substantially limits their potential as a foundation for multimodal interactive AI.

In this work, we aim to transform streaming video LLMs from passive observers into interactive assistants. Given that modern video LLMs largely inherit the architecture of LLMs, we ask a key question: *Can interactive capabilities emerge through appropriate data and learning, without requiring specialized model designs*? To address this question, we first identify a set of core capabilities required for interactive streaming video understanding and curate a large-scale video interaction dataset for training and evaluation. We then develop a learning framework that explicitly support these capabilities and establish a comprehensive benchmark for interactive streaming video LLMs.

Our experiments show that an offline video LLM can indeed be transformed through post-training into a capable interactive streaming video LLM. The resulting model preserves its ability to understand a streaming video while acquiring substantially richer interactive behaviors: it can simultaneously manage multiple user-specified tasks, modify or cancel existing tasks, and proactively

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.

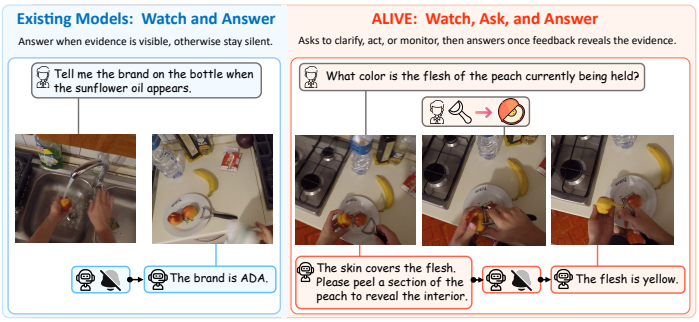


Figure 1: **From passive to interactive streaming video LLMs.** ALIVE extends *watch-and-answer* (left) to *watch-ask-answer* (right), enabling task coordination and feedback seeking.

seek user feedback based on prior requests and events unfolding in the video. These results suggest a path toward streaming video LLMs that do not merely watch and respond, but sustain ongoing interactions with users as the video streams.

Our **contributions** are summarized as follows:

- We formulate **interactive streaming video understanding** as a set of core capabilities, including proactive monitoring, multi-task coordination, and feedback seeking. Building on this formulation, we introduce **ALIVE-234K**, a large-scale dataset for post-training, together with **FSI-Bench** a comprehensive benchmark for evaluation.
- We investigate post-training strategies and introduce a new training objective for enabling interactivity in streaming video LLMs. Together, our learning framework enables *ALIVE*, a new family of interactive streaming video LLMs that requires no specialized model architectures.
- We show that our models achieves highly competitive performance across multiple streaming video benchmarks, preserving strong video understanding capabilities while substantially improving performance on interactive tasks.

2 Related Work

We provide additional discussion of VLMs and streaming VLMs in the Appendix A. Here, we mainly contrast the dataset and benchmark coverage of our ALIVE with prior work. Table 1 summarizes the task coverage of representative streaming benchmarks and training datasets.

Streaming video benchmarks increasingly assess models under temporally partial and causally evolving observations. At the perception and context level, StreamingBench and OVO-Bench evaluate real-time understanding, retrospective retrieval, and future-oriented response. SVBench, StreamBench, and OmniMMI cover temporal multi-turn dialogue, memory, multimodal interaction, and proactive reasoning (Lin et al., 2024; Li et al., 2025; Yang et al., 2025b; Xiong et al., 2025; Wang et al., 2025g). At the interaction level, ProactiveVideoQA evaluates the timing and quality of proactive responses through PAUC (Wang et al., 2025f). Streamo-Bench and RIVER Bench span instruction following, temporal grounding, narration, retrospective memory, and live perception (Xia et al., 2026; Shi et al., 2026); IPIBench adds proactive monitoring, task updates, and interleaved reactive–proactive requests (Li et al., 2026a). Yet no benchmark jointly tests whether a streaming model can detect that a request remains unresolved, select among Ask-for-Information, Ask-for-Execution, and Ask-for-Monitor, and complete the original task after incorporating user feedback.

3 Streaming Interactivity: Problem Formulation and Datasets

3.1 Problem Formulation

We consider a streaming video model that continuously receives video observations and user utterances. At decision time t, its output depends only on the observed video prefix $V_{1:t}$ and the interaction history H_t, comprising received user utterances and prior model outputs. All mappings below describe causal rollouts: each output uses only information available at its own decision time.

Table 1: Comparison of representative streaming video benchmarks and training datasets by interaction properties, task-family coverage, and scale.

Type	Dataset & Benchmark	Multi-turn	Visual Source	Scale	PM	STC	FSI
Eval	StreamingBench Lin et al. (2024)	✓	Both	4.5K	✓	✗	✗
	OVO-Bench Li et al. (2025)	✗	Both	2.8K	✓	✗	✗
	SVBench Yang et al. (2025b)	✓	Both	50.0K	✗	✗	✗
	ProactiveVideoQA Wang et al. (2025f)	✗	Both	1.4K	✓	✗	✗
	OmniMMI Wang et al. (2025g)	✓	Both	2.3K	✓	✗	✗
	RIVER Shi et al. (2026)	✗	Both	4.3K	✓	✗	✗
	EgoSAT Lei et al. (2026)	✗	Ego	4.8K	✗	✗	✗
	IPIBench Li et al. (2026a)	✓	Both	3.7K	✓	✓	✗
	FSI-Bench (Ours)	✓	Both	N/R	✓	✗	✓
Train	LIVE Chen et al. (2024)	✗	Ego	134K	✓	✗	✗
	MMDuet2-Data Wang et al. (2025e)	✗	Both	52.8K	✓	✗	✗
	Streamo-Instruct Xia et al. (2026)	✗	Both	465K	✓	✗	✗
	AURA Lu et al. (2026)	✗	Both	174K	✓	✗	✗
	JoyAI-VL-Interaction Yao et al. (2026)	✓	Both	>4M	✓	✓	✗
	VideoChat3-OL617K Li et al. (2026b)	✗	Both	617K	✓	✗	✗
	ProAssist Zhang et al. (2025)	✓	Ego	30.1K	✓	✓	✗
	ALIVE-234K (Ours)	✓	Both	234K	✓	✓	✓

PM: Proactive Monitoring; **STC**: Streaming Task Coordination; **FSI**: Feedback-Seeking Interaction. ✓ denotes explicit support, and ✗ denotes no explicit support. Scale uses each work's native counting unit (K = thousand, M = million).

A **passive streaming** model executes a fixed query $Q_{t_0} \in H_t$ issued at time t_0:

$$f_\theta(Q_{t_0}, V_{1:t}) \to (r_{t_0}, \ldots, r_t), \tag{1}$$

where $(r_{t_0}, \ldots, r_t)$ are responses from t_0 to t including the choice of remaining silence (`SILENCE`). This formulation covers immediate question answering and delayed or repeated event monitoring. Here, *passive* refers to executing a fixed request without accommodating new requests or seeking additional user input, not to the absence of proactive responses.

An **interactive streaming** model retains the same streaming capabilities as the passive setting, while supporting a broader range of outputs and evolving interactions. Before receiving further user input, its rollout is similar yet with richer output

$$f_\theta(Q_{t_0}, V_{1:t}) \to (o_{t_0}, \ldots, o_t), \;\; \text{with} \;\; o_\tau = \begin{cases} r_\tau, & \text{respond or remain silent,} \\ c_\tau, & \text{request clarification or feedback,} \\ a_\tau, & \text{request user execution.} \end{cases} \tag{2}$$

Unlike passive streaming, interactive streaming maintains an evolving interaction history H_t. At a later decision time $\tau > t$, a new interaction ΔH_τ updates this state:

$$H_\tau = \mathrm{Update}(H_t, \Delta H_\tau), \tag{3}$$

where ΔH_τ may correspond to a new user query that adds, cancels, or modifies a request, user feedback that resolves a previous clarification request c_t, or newly observed evidence resulting from an execution request a_t. With the updated state and retained interaction history, the model continues with a similar streaming processing

$$f_\theta(H_\tau, V_{1:\tau'}) \to O_{\tau:\tau'}, \qquad \tau' \geq \tau. \tag{4}$$

This process unfolds along a shared arrow of time, where both the user and model can initiate interaction, update one another, and continue acting on the evolving video stream. Interactive streaming thus enables bidirectional communication within a single continuous trajectory.

3.2 Interactive Streaming Task Taxonomy

As aforementioned, passive streaming models primarily focus on proactive monitoring, while interactive streaming further extends to task coordination and feedback seeking. Accordingly, we briefly introduce the core interactive streaming capabilities across three complementary families.

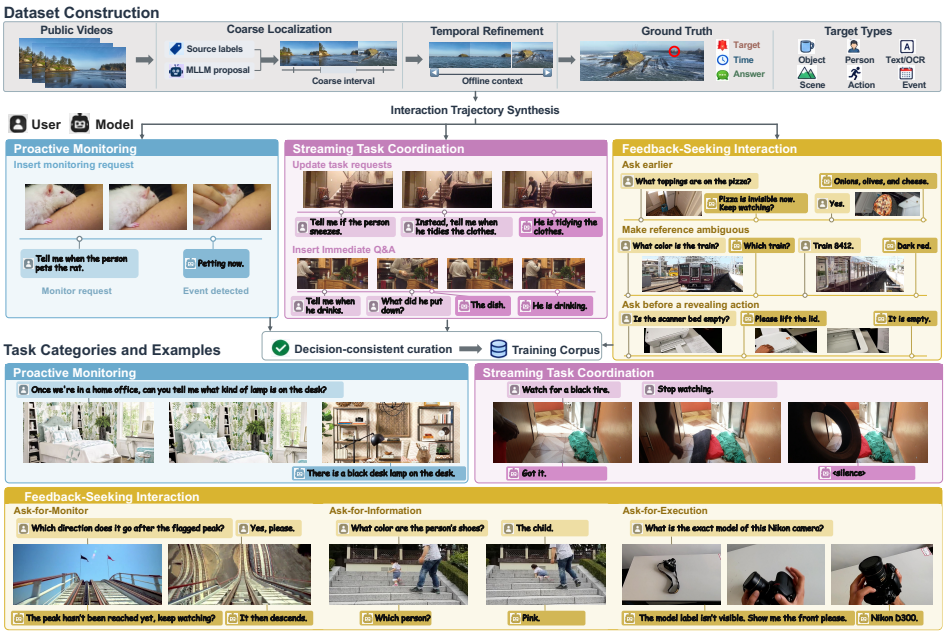


Figure 2: **Overview of our data construction pipeline**. Starting from public video datasets, we ground targets and their temporal evidence, construct interaction trajectories for each task family, and apply decision-consistent curation to produce the final training corpus.

Proactive Monitoring. The user specifies a condition to monitor, and the model observes the stream until sufficient evidence supports a response:

- *Proactive Timing*: The model issues a notification when the target condition is first verifiably satisfied, completing the task.
- *Proactive Understanding*: The model answers an associated visual question once both the trigger and sufficient answer evidence are available.
- *Repeated Proactiveness*: The model maintains the request across multiple occurrences and responds once for each distinct occurrence.

Streaming Task Coordination. The model maintains and updates active commitments as user requests evolve or coexist:

- *Task State Update*: The user cancels or modifies an existing monitoring task, and the model updates subsequent behavior accordingly.
- *Concurrent Requests*: The model handles multiple monitoring objectives or interleaved immediate questions while preserving all remaining active commitments.

Feedback-Seeking Interaction. When the current context is insufficient to resolve a request, the model identifies the missing prerequisite and actively seeks the required feedback:

- *Ask-for-Monitor*: When the answer depends on future observations, the model asks whether the user wants to establish a monitoring task and responds once the relevant evidence appears.
- *Ask-for-Information*: When the request is ambiguous, the model asks for clarification and uses the additional information to resolve the original query.
- *Ask-for-Execution*: When necessary visual evidence is unavailable, the model requests a concrete user action that reveals the missing information before answering.

Across these tasks, the model must preserve valid commitments and incorporate subsequent user feedback or visual evidence into its ongoing interaction state.

3.3 Data Construction and Datasets

Our construction starts from publicly available video datasets and organizes annotation around when relevant evidence becomes available and what remains unresolved at each decision point. We detail the source datasets in Appendix C. As showin in Fig. 2, we first temporally ground target events and

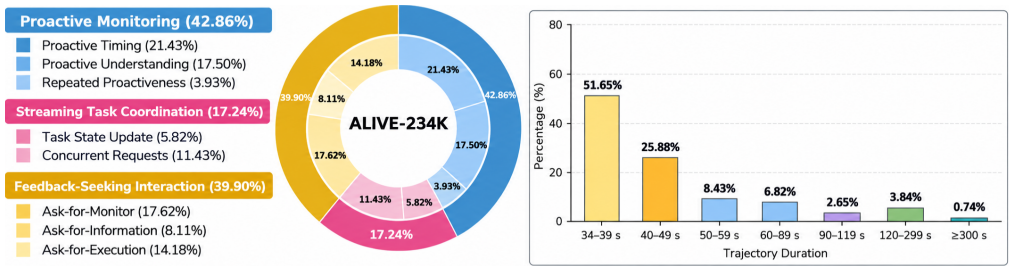


Figure 3: **Dataset statistics**. ALIVE-234K, designed for post-training, covers 3 interaction categories and 8 task types, with trajectories spanning from tens of seconds to several minutes.

evidence, then synthesize interaction trajectories for proactive monitoring, streaming task coordination, and feedback seeking.

Temporal coarse localization and refinement. For videos with existing descriptions and approximate timestamps, we use them to identify candidate intervals; otherwise, an offline MLLM first detects relevant targets and coarsely localizes their occurrence. We then refine each candidate using an offline MLLM with access to video context both before and after the candidate event. This bidirectional temporal context helps identify the earliest sampled time at which the required visual condition is met, such as entity recognizability, text readability, scene visibility, or action completion. We additionally verify that each assigned response is supported by the video prefix and dialogue available at that time.

Interaction trajectory synthesis. Based on the grounded evidence and timestamps, we synthesize three families of interaction trajectories:

• ***Proactive monitoring***. We convert temporally grounded conditions into monitoring requests and align notifications with verified trigger times. Single-occurrence requests terminate after one response, while repeated requests persist across occurrences. For proactive understanding, we add a visual question beyond the trigger and verify that both are supported at response time.

• ***Streaming task coordination***. We extend monitoring trajectories with changing or concurrent commitments. Cancellation suppresses subsequent responses for the affected task, modification replaces its original condition, and concurrent requests retain independent triggers and completion rules. We also insert immediate visual questions before a pending trigger, requiring the model to answer the new request while preserving the original monitoring commitment.

• ***Feedback-seeking interaction***. We identify decision points where the current request is not yet resolvable and construct feedback according to the missing prerequisite. *Ask-for-Monitor.* The requested information is not yet observable, but can be resolved through continued observation. We place the request before the evidence appears, generate a monitoring offer and user acceptance, and assign the final response to the later evidence-grounded time. *Ask-for-Information.* Multiple visible instances make the target referentially ambiguous. We construct questions for which different candidate instances yield different answers, then generate a clarification that uniquely identifies the intended target before producing the final response. *Ask-for-Execution.* The answer requires an additional grounded operation. We place the request before the recorded action, require a specific action that reveals the missing evidence, and assign the final answer only after the corresponding visual outcome becomes available.

We finally verify that the intended behavior is justified by the evidence available at each decision point: sufficient evidence supports direct answering, referential ambiguity supports clarification, a necessary operation supports execution, and unresolved future evidence supports monitoring. Ambiguous or invalid cases are filtered or rechecked.

ALIVE-234K for post-training. ALIVE-234K comprises 234,371 interaction trajectories spanning all three task families and eight sub-tasks. As shown in Fig. 3, we retain a notable proportion of proactive monitoring examples, as this capability interconnect with more advanced interactive behaviors. Feedback-seeking interactions are also well represented, while streaming task coordination occupies a smaller portion, since overly structured cancellation and modification patterns may interfere with the model's broader interaction capabilities. The trajectories further span diverse temporal

horizons: 51.65% last 34–39 seconds, 25.88% last 40–49 seconds, and over 20% extend beyond 50 seconds, including sequences lasting several minutes.

FSI-Bench for evaluation. No existing benchmark evaluates whether a streaming model can identify an unresolved request, seek the missing prerequisite, and then complete the original request (Table 1). We introduce FSI-Bench to evaluate this closed loop. Each episode pairs a video prefix with a query at decision time t_0, an answer window $[t_s, t_e]$, and a reference answer, with resolving feedback additionally provided for queries that are not immediately answerable. FSI-Bench contains 1,163 verified episodes: 450 Ask-for-Information, 412 Ask-for-Execution, 101 Ask-for-Monitor, and 200 Answer episodes. Additional details on benchmark construction, quality control, and evaluation are provided in Appendix F.

4 Learning Interactive Streaming Video LLMs

A streaming interactive video model must learn two complementary competencies: selecting the appropriate response strategy when speaking, and recognizing when the current interaction state calls for silence rather than a response. We first describe controlled contrastive data that provides training signal for each competency, then present a unified training procedure that turns interaction trajectories into conditionally weighted training instances.

Output format. At each decision time t, the interactive streaming model (Eq. 2) observes the video prefix $V_{1:t}$ and interaction history H_t, and produces an output $o_t \sim p_\theta(\cdot \mid V_{1:t}, H_t)$. For training, we decompose this into two complementary decisions, that are implemented through a single autoregressive distribution p_θ without separate prediction heads.

The *response gate* determines whether the model responses (R) or remains silent (S), yielding $g_t \in \{R, S\}$. This decision depends on the current H_t: whether a valid task exists, whether the observed evidence matches the active query, and whether any required user feedback has been received. While, the *action strategy* determines the output form when $g_t = R$. Following the output space in Eq. 2, the model selects among a direct answer r_t, a clarification request c_t, or an execution request a_t, according to what evidence is available in $V_{1:t}$ and what prerequisite is missing.

Response-gate contrasts: When to respond? Whether a visual event warrants a response is determined not by the event alone, but by the current interaction history. The same target appearance may require a notification under an active monitoring task, silence after that task has been canceled, and continued observation while the model awaits user feedback. A model trained only on correlations between visual events and responses will respond to salient changes regardless of whether the underlying task remains valid.

We construct response-gate contrasts by fixing the visual prefix and a shared comparison time while varying one interaction condition, so that the correct label changes between R and S. Three families of conditions produce these contrasts. *Query relevance*: substituting a non-matching query for the original, so that the visual event no longer corresponds to the active monitoring target. *Task lifecycle*: placing the same event sequence under an active repeated task, a fulfilled single-occurrence task, or a canceled task. *Feedback state*: presenting the same later evidence under accepted versus rejected monitoring proposals, or with versus without a clarification that resolves an ambiguous reference. These contrasts create counterfactual groups: sets of branches that share visual context but assign opposite R and S labels at identified decision points.

Action-strategy contrasts: What actions to take? When the model responds, the appropriate strategy depends on whether the current evidence is sufficient to answer, and if not, what kind of gap prevents it. We construct controlled training pairs that vary the evidence condition while preserving the underlying question, so that each gap type is paired with a direct-answer counterpart.

Three types of gaps determine the appropriate strategy. If the answer depends on visual evidence that is not yet available, the model should ask the user to perform a concrete action that would reveal it when such an action exists, or propose to keep observing when it does not. If the user's reference is genuinely ambiguous among multiple visible instances that would yield different answers, the model should ask the user to clarify the intended target. If none of these gaps exists, the model should answer directly. We construct these contrasts by repositioning the question across video prefixes with different evidence states and by introducing or removing referential ambiguity, verifying the

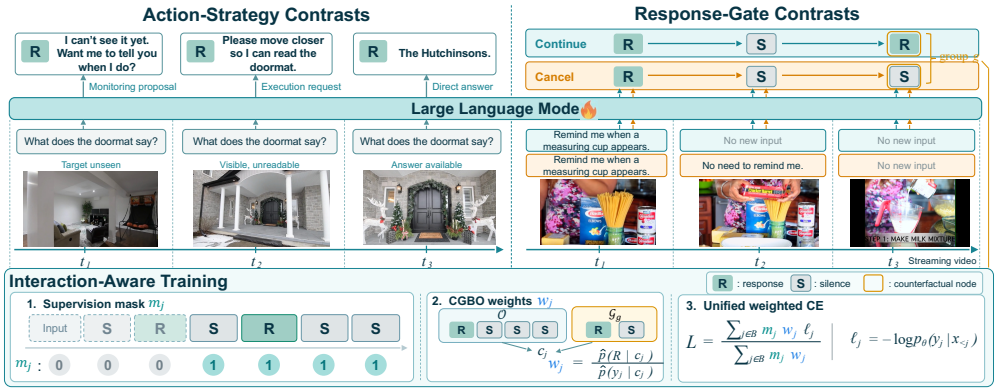


Figure 4: **Overview of our training. Left**: Action-strategy contrasts vary the available evidence while fixing the question, so the model learns to select among monitoring proposals, execution requests, and direct answers. **Right**: Response-gate contrasts fix the visual event while varying the interaction state, producing counterfactual groups where the same observation requires a response under one condition and silence under another. **Bottom**: The unified weighted cross-entropy objective combines standard response-text supervision with CGBO-balanced control tokens.

available evidence at each configuration independently. These contrasts present R on both sides with different response texts. The model learns to select among strategies through standard next-token supervision on the response content.

Interaction-Aware Training. Given contrasts augmented interaction trajectories, we describe how each trajectory is converted into a training instance, what tokens are supervised, and how critical decision points are weighted. We anchor each instance at the last response in the effective visual window and sample its endpoint from a random offset in $[a, a+M]$, allowing instances to extend into post-response silence rather than ending immediately after the response. The preceding N frames provide the visual and interaction context. For response-gate contrast pairs, the shared comparison time takes priority, ensuring that the negative branch reaches its critical decision point. Within this window, we supervise the anchor response, including its control token and text, together with all subsequent silence tokens. Earlier responses remain as context but are not supervised. This captures both routine inter-event waiting and the critical silence introduced by response-gate contrasts.

Among the supervised silence tokens, most represent ordinary waiting, while the critical tokens at counterfactual decision points are rare. A global response-to-silence ratio would assign them identical weight, making a silence after task cancellation indistinguishable from one that merely waits for a target. We resolve this through *Counterfactual Group Balancing Objective* (CGBO). Let $\mathcal{G}_1, \ldots, \mathcal{G}_K$ denote the counterfactual groups produced by response-gate contrasts, and let $\mathcal{O} = \mathcal{B} \setminus \bigcup_{g=1}^{K} \mathcal{G}_g$ denote the ordinary pool containing all remaining supervised nodes. Each node j belongs to exactly one condition $c_j \in \{\mathcal{O}, \mathcal{G}_1, \ldots, \mathcal{G}_K\}$. Within each condition, we rebalance by inverse label frequency:

$$w_j = \hat{p}(R \mid c_j)/\hat{p}(y_j \mid c_j), \tag{5}$$

where $\hat{p}$ denotes the empirical label frequency within condition c_j. For response nodes this reduces to $w_j = 1$. For ordinary silence nodes ($j \in \mathcal{O}$), the weight becomes $w_j^{(\mathcal{O})} = n^R(\mathcal{O})/n^S(\mathcal{O})$. For silence nodes in counterfactual group g ($j \in \mathcal{G}_g$), the weight becomes $w_j^{(g)} = n^R(\mathcal{G}_g)/n^S(\mathcal{G}_g)$, precomputed offline from the complete group. Response-text tokens receive unit weight throughout. Group membership is training metadata and introduces no additional model tokens.

Let $m_j \in \{0, 1\}$ denote the token-level supervision mask defined above, and $\ell_j = -\log p_\theta(y_j \mid V_{1:t}, y_{<j})$ the cross-entropy for token j. The complete model is trained with a weighted objective:

$$\mathcal{L} = \sum\nolimits_{j \in \mathcal{B}} m_j \, w_j \, \ell_j / \sum\nolimits_{j \in \mathcal{B}} m_j \, w_j. \tag{6}$$

Action-strategy contrasts contribute through standard response-text supervision, while response-gate contrasts contribute through conditionally balanced control tokens. The unified objective subsumes both strategy selection and response timing within a single autoregressive framework, requiring only weighted cross-entropy over the model's existing vocabulary.

Table 2: Results on OVOBench. We report accuracy (%) across three scenario categories: Real-Time Visual Perception, Backward Tracing, and Forward Active Responding. Best completed values in each column are in **bold**.

Model	Real-Time Visual Perception							Backward Tracing				Forward Active Responding				Overall
	OCR	ACR	ATR	STU	FPD	OJR	Avg.	EPM	ASI	HLD	Avg.	REC	SSR	CRR	Avg.	
VideoLLM-online-8B	8.1	23.9	12.1	14.0	45.5	21.2	20.8	22.2	18.8	12.2	17.7	-	-	-	–	–
Flash-VStream-7B	24.2	29.4	28.5	33.7	25.7	28.8	28.4	39.1	37.2	5.9	27.4	8.0	67.3	60.0	45.1	33.6
Dispider-7B	57.7	49.5	62.1	44.9	61.4	51.6	54.6	48.5	55.4	4.3	36.1	18.1	37.4	48.8	34.7	41.8
VideoChat3-4B	89.3	70.6	76.7	63.5	70.3	65.8	72.7	55.2	67.6	32.8	51.9	30.1	68.2	57.1	51.8	58.8
AURA-8B	89.9	79.8	80.2	**70.2**	**77.2**	**81.5**	**79.8**	54.9	67.6	58.6	60.4	30.4	**75.8**	61.3	**55.8**	65.3
JoyAI-VL-Interaction-8B	89.9	67.9	76.7	62.9	75.3	70.7	73.9	57.9	**68.9**	38.2	55.0	31.1	70.3	57.1	52.8	60.6
Qwen3-VL-4B-Inst. (standard64)	77.9	54.1	75.9	51.7	75.2	62.0	66.1	54.9	60.8	16.1	43.9	41.5	60.9	43.3	48.6	52.9
Qwen3-VL-8B-Inst. (standard64)	75.2	58.7	75.0	55.1	67.3	62.0	65.5	54.5	70.9	9.7	45.1	**44.3**	65.2	46.7	52.0	54.2
ALIVE-4B (Ours)	**92.6**	75.2	75.9	57.3	68.3	79.9	74.9	59.3	60.1	**87.1**	**68.8**	31.2	67.4	60.8	53.2	65.6
ALIVE-8B (Ours)	89.2	**80.0**	**82.8**	65.2	64.0	**81.5**	77.1	**61.1**	**68.9**	74.2	68.1	31.4	66.0	**66.7**	54.7	**66.6**

5 Experiments and Results

In this section, we first introduce implementation details, benchmarks, evaluation protocol, and metrics. We then highlight the main results and key findings. Additional details on implementation and further experimental results are provided in the Appendix B and D.

Implementation Details. We build two model variants, ALIVE-4B and ALIVE-8B, on top of Qwen3-VL-4B-Instruct and Qwen3-VL-8B-Instruct, respectively. The training data combines ALIVE-234K with 153K offline samples curated from LLaVA-Video (Zhang et al., 2024), Ego4D (Grauman et al., 2022), and STAR (Wu et al., 2024). During training, the vision encoder and the connector are frozen, and only the language model is updated. We train for a single epoch with a batch size of 256 and a learning rate of 1×10^{-5}. Input videos are sampled at 1 FPS. Following the instance construction in §4, we set the context window length to $N{=}30$ frames and the maximum post-anchor offset to $M{=}4$ positions.

Benchmarks. We adopt two established benchmarks (OVOBench and IPIBench), as well as our own FSI-Bench. OVOBench (Li et al., 2025) evaluates online video understanding, testing whether a model can answer questions using only the video content observed up to the query time, without access to future frames. IPIBench (Li et al., 2026a) assesses proactive interaction capabilities of streaming video models, including the ability to monitor events and respond at appropriate times. FSI-Bench further evaluates feedback-seeking interactions.

Evaluation protocol. For OVOBench and IPIBench, we report official results when they are publicly available; otherwise, we evaluate models by strictly following the code released in the official benchmark repositories (Li et al., 2025; 2026a). For our own models, we adopt a streaming evaluation setup: frames are sampled at 1 FPS and the model receives a sliding window of the most recent 30 frames. For FSI-Bench, evaluation follows the causal rollout of §3: models observe only the frames and utterances available so far, without future frames or reference answers. At t_0, each model must choose among answering, waiting, or requesting information/action, matching Eq. 2; malformed responses are classified by the judge. The action label is then removed, and the rollout continues in the model's native streaming format until three steps after t_e, with user feedback injected at its annotated time. The initial decision evaluates the action strategy, while the continuation evaluates the response gate (§4).

Metrics. For OVOBench and IPIBench, we directly adopt the metrics defined in the original papers (Li et al., 2025; 2026a). For FSI-Bench, we report accuracy, defined as the fraction of interaction trajectories that are fully resolved: a trajectory is considered correct only when the initial decision at t_0 is correct *and* the subsequent response is correct and falls within the valid time window. Both the decision and the response are assessed by an LLM judge; details are provided in Appendix F.3.

5.1 Main Results

OVOBench and IPIBench. As shown Tab. 2 and 3, Across both OVOBench and IPIBench, ALIVE consistently improves the capabilities most closely tied to interactive streaming rather than only static visual recognition. On OVOBench, ALIVE-8B achieves the best overall score (66.6), with particularly strong gains in backward tracing and forward active responding, indicating improved

Table 3: Results on IPIBench. We report accuracy (%) across three task groups: Proactive Monitoring, Proactive Task Management, and Interleaved Reactive-Proactive requests. Best completed values in each column are in **bold**.

Model	Proactive Monitoring				Proactive Task Management				Interleaved Reactive-Proactive				Overall
	Timing	Under.	Repeat.	Avg.	Cancel	Modify	Multi	Avg.	R2P	RuP	RaP	Avg.	
VideoLLM-online-8B	14.6	0.1	1.8	5.5	7.9	1.7	6.6	5.4	0.0	1.1	1.6	0.9	3.9
Flash-VStream-7B	5.9	0.0	0.0	2.0	2.3	0.0	0.0	0.8	0.0	0.0	0.8	0.3	1.0
Dispider-7B	18.7	0.7	1.8	7.1	-	-	-	-	-	-	-	-	-
VideoChat3-4B	25.1	7.8	2.4	11.8	22.2	14.8	3.6	13.5	19.5	5.7	6.7	10.6	12.0
AURA-8B	36.4	25.5	0.9	21.0	32.4	25.8	12.1	23.4	31.3	11.3	18.4	20.3	21.6
JoyAI-VL-Interaction-8B	44.0	16.4	3.3	21.2	44.4	11.7	1.0	19.0	30.0	7.0	24.1	20.4	20.2
Qwen3-VL-4B-Inst.	49.2	29.5	9.4	29.4	7.9	20.5	14.9	14.4	24.8	**25.3**	28.5	26.2	23.3
Qwen3-VL-8B-Inst.	43.1	24.5	4.3	24.0	1.4	17.5	9.2	9.4	24.4	19.8	23.1	22.4	18.6
ALIVE-4B (Ours)	56.7	**30.6**	12.7	33.3	**56.5**	**29.2**	18.2	34.6	36.2	20.2	37.1	31.1	33.0
ALIVE-8B (Ours)	**64.6**	25.6	**35.9**	**42.0**	55.6	28.9	**28.8**	**37.8**	**38.1**	24.9	**44.0**	**35.7**	**38.5**

Table 4: Results on FSI-Bench. We report accuracy (%): a trajectory is correct only when the initial decision and the subsequent response are both correct. Answer is the direct-answer control. Avg. is the macro average over the four task types. Best results are in **bold**.

Model	Ask-for-Monitor	Ask-for-Information	Ask-for-Execution	Answer	Avg.
VideoChat3-4B	6.7	16.2	10.6	3.0	9.1
AURA-8B	5.5	21.7	9.0	22.2	14.6
JoyAI-VL-Interaction-8B	0.6	0.2	0.0	67.2	17.0
Qwen3-VL-4B-Inst.	3.0	1.9	6.5	61.6	18.3
Qwen3-VL-8B-Inst.	4.9	9.3	6.1	48.0	17.1
ALIVE-4B (Ours)	30.3	**65.3**	30.9	67.7	48.6
ALIVE-8B (Ours)	**37.6**	63.7	**35.1**	**68.7**	**51.3**

reasoning over the arrow of time. The advantage becomes more pronounced on IPIBench: ALIVE-8B reaches 38.5 overall, substantially outperforming prior streaming models across all tasks. Notably, the largest gains appear on interaction-intensive tasks such as repeated proactiveness, multi-task coordination, and reactive-to-proactive transitions. We attribute these gains to our carefully designed interaction-aware training objective and high-quality interactive streaming data.

FSI-Bench. Table 4 shows that existing streaming models struggle to complete feedback-seeking interactions, often answering prematurely, remaining passive, or issuing ineffective requests. In contrast, ALIVE consistently handles the full feedback-seeking loop: it recognizes when the current evidence is insufficient, requests the appropriate information, action, or monitoring commitment, and incorporates the resulting feedback to complete the original request. Its strong performance on the Answer control further suggests that this behavior is selective rather than indiscriminate, highlighting ALIVE's improved ability to decide when feedback is necessary and when the available evidence already supports a direct response.

6 Conclusion

In this work, we studied how to transform streaming video LLMs from passive observers into interactive assistants. We formulated *interactive streaming video understanding* around three core capabilities: proactive monitoring, multi-task coordination, and feedback seeking. Based on this formulation, we introduced *ALIVE-234K* for post-training and *FSI-Bench* for comprehensive evaluation. We further investigated post-training strategies for interactive streaming video LLMs and developed *CGBO*, a conditional balancing objective that up-weights critical silence decisions at counterfactual decision points. Together, these components enable *ALIVE*, a family of interactive streaming video LLMs that requires no specialized model architectures. Our experiments show that ALIVE preserves strong streaming video understanding while acquiring substantially richer interactive behaviors. More broadly, our results suggest that interactivity in streaming video LLMs can emerge through appropriate data and post-training, without fundamentally redesigning the underlying model architecture. We hope this work takes a step toward multimodal assistants that do not merely watch and respond, but continuously perceive an evolving world and sustain ongoing interactions with users over time.

AI USE STATEMENT

We used generative AI tools during dataset construction and evaluation. Specifically, an offline multimodal LLM was used for coarse temporal localization of target events in source videos, and language models assisted in synthesizing interaction trajectories, generating monitoring requests, clarification exchanges, and response formulations during the construction of ALIVE-234K. For FSI-Bench evaluation, an LLM judge assessed decision validity, request quality, and answer correctness as described in Appendix F.3. We also used generative AI tools for drafting and editing portions of the manuscript. All AI-generated dataset content was verified through decision-consistent curation and visual evidence checks as described in §3.3. Evaluation judgments were validated against the defined scoring criteria. Code and numerical results were checked through consistency verification. We take responsibility for the final content of this work, including all AI-assisted text, data, and figures.

ETHICS STATEMENT

ALIVE-234K is constructed from 41 publicly available video datasets (Appendix C); all source videos are used in accordance with their original distribution terms. The interaction trajectories are synthetically constructed around recorded video content and do not involve human subjects performing scripted interactions. FSI-Bench evaluation relies on LLM-based judging with clearly defined criteria; no human participants were recruited for evaluation beyond the standard research team. Our work studies interactive streaming video understanding and does not involve generating or manipulating video content. We document the capabilities and limitations of the evaluated models to support responsible assessment of interactive streaming video systems.

REPRODUCIBILITY STATEMENT

§3 describes the problem formulation, task taxonomy, and data construction pipeline, with Fig. 2 illustrating the construction process and Fig. 3 summarizing dataset statistics. §4 specifies the training formulation, contrastive data construction, and the CGBO objective. §5 and Appendix D report model architectures, base models, training hyperparameters (learning rate, batch size, epoch count), and frame sampling settings. §5 and Appendix F detail the FSI-Bench composition, rollout protocol, judging procedure, and scoring criteria. Appendix C lists all 41 source video datasets with trajectory counts. For existing benchmarks, we specify the evaluation setup and identify which results use official numbers versus our reproduction.

REFERENCES

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

Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval, 2022. URL https://arxiv.org/abs/2104.00650.

Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

Vivek Chavan, Yasmina Imgrund, Tung Dao, Sanwantri Bai, Bosong Wang, Ze Lu, Oliver Heimann, and Jörg Krüger. Indego: A dataset of industrial scenarios and collaborative work for egocentric assistants, 2025. URL https://arxiv.org/abs/2511.19684.

Joya Chen, Zhaoyang Lv, Shiwei Wu, Kevin Qinghong Lin, Chenan Song, Difei Gao, Jia-Wei Liu, Ziteng Gao, Dongxing Mao, and Mike Zheng Shou. Videollm-online: Online video large language model for streaming video, 2024. URL https://arxiv.org/abs/2406.11816.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. The epic-kitchens dataset: Collection, challenges and baselines, 2020. URL https://arxiv.org/abs/2005.00343.

Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, and Chen Change Loy. Mevis: A large-scale benchmark for video segmentation with motion expressions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2694–2703, October 2023.

Dave Epstein, Boyuan Chen, and Carl Vondrick. Oops! predicting unintentional action in video, 2019. URL https://arxiv.org/abs/1911.11206.

Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query, 2017. URL https://arxiv.org/abs/1705.02101.

Mingji Ge, Qirui Chen, Zeqian Li, and Weidi Xie. Densestep2m: A scalable, training-free pipeline for dense instructional video annotation, 2026. URL https://arxiv.org/abs/2604.26565.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18995–19012, 2022.

Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, Eugene Byrne, Zach Chavis, Joya Chen, Feng Cheng, Fu-Jen Chu, Sean Crane, Avijit Dasgupta, Jing Dong, Maria Escobar, Cristhian Forigua, Abrham Gebreselasie, Sanjay Haresh, Jing Huang, Md Mohaiminul Islam, Suyog Jain, Rawal Khirodkar, Devansh Kukreja, Kevin J Liang, Jia-Wei Liu, Sagnik Majumder, Yongsen Mao, Miguel Martin, Effrosyni Mavroudi, Tushar Nagarajan, Francesco Ragusa, Santhosh Kumar Ramakrishnan, Luigi Seminara, Arjun Somayazulu, Yale Song, Shan Su, Zihui Xue, Edward Zhang, Jinxu Zhang, Angela Castillo, Changan Chen, Xinzhu Fu, Ryosuke Furuta, Cristina Gonzalez, Prince Gupta, Jiabo Hu, Yifei Huang, Yiming Huang, Weslie Khoo, Anush Kumar, Robert Kuo, Sach Lakhavani, Miao Liu, Mi Luo, Zhengyi Luo, Brighid Meredith, Austin Miller, Oluwatumininu Oguntola, Xiaqing Pan, Penny Peng, Shraman Pramanick, Merey Ramazanova, Fiona Ryan, Wei Shan, Kiran Somasundaram, Chenan Song, Audrey Southerland, Masatoshi Tateno, Huiyu Wang, Yuchen Wang, Takuma Yagi, Mingfei Yan, Xitong Yang, Zecheng Yu, Shengxin Cindy Zha, Chen Zhao, Ziwei Zhao, Zhifan Zhu, Jeff Zhuo, Pablo Arbelaez, Gedas Bertasius, David Crandall, Dima Damen, Jakob Engel, Giovanni Maria Farinella, Antonino Furnari, Bernard Ghanem, Judy Hoffman, C. V. Jawahar, Richard Newcombe, Hyun Soo Park, James M. Rehg, Yoichi Sato, Manolis Savva, Jianbo Shi, Mike Zheng Shou, and Michael Wray. Ego-exo4d: Understanding skilled human activity from first- and third-person perspectives, 2024. URL https://arxiv.org/abs/2311.18259.

Chunhui Gu, Chen Sun, David A. Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, Cordelia Schmid, and Jitendra Malik. Ava: A video dataset of spatio-temporally localized atomic visual actions, 2018. URL https://arxiv.org/abs/1705.08421.

Mingfei Han, Liang Ma, Kamila Zhumakhanova, Ekaterina Radionova, Jingyi Zhang, Xiaojun Chang, Xiaodan Liang, and Ivan Laptev. Roomtour3d: Geometry-aware video-instruction tuning for embodied navigation, 2025a. URL https://arxiv.org/abs/2412.08591.

Mingfei Han, Linjie Yang, Xiaojun Chang, Lina Yao, and Heng Wang. Shot2story: A new benchmark for comprehensive understanding of multi-shot videos, 2025b. URL https://arxiv.org/abs/2312.10300.

Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language, 2017. URL https://arxiv.org/abs/1708.01641.

Jie Lei, Tamara L. Berg, and Mohit Bansal. Qvhighlights: Detecting moments and highlights in videos via natural language queries, 2021. URL https://arxiv.org/abs/2107.09609.

Yijia Lei, Jinzhao Li, Yichi Zhang, Jiacheng Hua, Yin Li, and Miao Liu. Egosat: A comprehensive benchmark of ego centric s treaming inter act ion understanding. In *European Conference on Computer Vision*, pp. 576–593. Springer, 2026.

Jinzhao Li, Yinuo Chen, Wenxuan Song, Yijia Lei, Yichi Zhang, Honglei Yan, Panwang Pan, and Miao Liu. Ipibench: Evaluating interactive proactive intelligence of mllms under continuous streams. *arXiv preprint arXiv:2605.27074*, 2026a.

Xinhao Li, Yuhan Zhu, Xiangyu Zeng, Yuhao Dong, Haoning Wu, Zhiqiu Zhang, Yuandong Yang, Changlian Ma, Qingyu Zhang, Yansong Shi, et al. Videochat3: Fully open video mllm for efficient and generalist video understanding. *arXiv preprint arXiv:2607.14935*, 2026b.

Yifei Li, Junbo Niu, Ziyang Miao, Chunjiang Ge, Yuanhang Zhou, Qihao He, Xiaoyi Dong, Haodong Duan, Shuangrui Ding, Rui Qian, Pan Zhang, Yuhang Zang, Yuhang Cao, Conghui He, and Jiaqi Wang. Ovo-bench: How far is your video-llms from real-world online video understanding?, 2025. URL https://arxiv.org/abs/2501.05510.

Shuo Liang, Yiwu Zhong, Zi-Yuan Hu, Yeyao Tao, and Liwei Wang. Fine-grained spatiotemporal grounding on egocentric videos, 2025. URL https://arxiv.org/abs/2508.00518.

Junming Lin, Zheng Fang, Chi Chen, Zihao Wan, Fuwen Luo, Peng Li, Yang Liu, and Maosong Sun. Streamingbench: Assessing the gap for mllms to achieve streaming video understanding, 2024. URL https://arxiv.org/abs/2411.03628.

Xudong Lu, Yang Bo, Jinpeng Chen, Shuhan Li, Xintong Guo, Huankang Guan, Fang Liu, Dunyuan Xu, Peiwen Sun, Heyang Sun, Rui Liu, and Hongsheng Li. Aura: Always-on understanding and real-time assistance via video streams, 2026. URL https://arxiv.org/abs/2604.04184.

Yaxi Lu, Shenzhi Yang, Cheng Qian, Guirong Chen, Qinyu Luo, Yesai Wu, Huadong Wang, Xin Cong, Zhong Zhang, Yankai Lin, Weiwen Liu, Yasheng Wang, Zhiyuan Liu, Fangming Liu, and Maosong Sun. Proactive agent: Shifting llm agents from reactive responses to active assistance, 2024. URL https://arxiv.org/abs/2410.12361.

Lingni Ma, Yuting Ye, Fangzhou Hong, Vladimir Guzov, Yifeng Jiang, Rowan Postyeni, Luis Pesqueira, Alexander Gamino, Vijay Baiyya, Hyo Jin Kim, Kevin Bailey, David Soriano Fosas, C. Karen Liu, Ziwei Liu, Jakob Engel, Renzo De Nardi, and Richard Newcombe. Nymeria: A massive collection of multimodal egocentric daily motion in the wild, 2024. URL https://arxiv.org/abs/2406.09905.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. Videogpt+: Integrating image and video encoders for enhanced video understanding, 2024. URL https://arxiv.org/abs/2406.09418.

Andreea-Maria Oncescu, João F. Henriques, Yang Liu, Andrew Zisserman, and Samuel Albanie. Queryd: A video dataset with high-quality text and audio narrations, 2021. URL https://arxiv.org/abs/2011.11071.

Viorica Pătrăucean, Lucas Smaira, Ankush Gupta, Adrià Recasens Continente, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Joseph Heyward, Mateusz Malinowski, Yi Yang, Carl Doersch, Tatiana Matejovicova, Yury Sulsky, Antoine Miech, Alex Frechette, Hanna Klimczak, Raphael Koster, Junlin Zhang, Stephanie Winkler, Yusuf Aytar, Simon Osindero, Dima Damen, Andrew Zisserman, and João Carreira. Perception test: A diagnostic benchmark for multimodal video models, 2023. URL https://arxiv.org/abs/2305.13786.

Rui Qian, Shuangrui Ding, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Dahua Lin, and Jiaqi Wang. Dispider: Enabling video llms with active real-time interaction via disentangled perception, decision, and reaction, 2025. URL https://arxiv.org/abs/2501.03218.

Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, and Manfred Pinkal. Grounding action descriptions in videos. *Transactions of the Association for Computational Linguistics*, 1:25–36, 2013. doi: 10.1162/tacl_a_00207. URL https://aclanthology.org/Q13-1003/.

Seonguk Seo, {Joon Young} Lee, and Bohyung Han. Urvos: Unified referring video object segmentation network with a large-scale benchmark. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (eds.), *Computer Vision – ECCV 2020 - 16th European Conference, 2020, Proceedings*, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), pp. 208–223, Germany, 2020. Springer Science and Business Media Deutschland GmbH. ISBN 9783030585549. doi: 10.1007/978-3-030-58555-6_13. Publisher Copyright: © 2020, Springer Nature Switzerland AG.; 16th European Conference on Computer Vision, ECCV 2020 ; Conference date: 23-08-2020 Through 28-08-2020.

Yansong Shi, Qingsong Zhao, Tianxiang Jiang, Xiangyu Zeng, Yi Wang, and Limin Wang. River: A real-time interaction benchmark for video llms, 2026. URL https://arxiv.org/abs/2603.03985.

Gunnar A. Sigurdsson, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, and Karteek Alahari. Charades-ego: A large-scale dataset of paired third and first person videos, 2018. URL https://arxiv.org/abs/1804.09626.

Yansong Tang, Dajun Ding, Yongming Rao, Yu Zheng, Danyang Zhang, Lili Zhao, Jiwen Lu, and Jie Zhou. Coin: A large-scale dataset for comprehensive instructional video analysis, 2019. URL https://arxiv.org/abs/1903.02874.

Haibo Wang, Bo Feng, Zhengfeng Lai, Mingze Xu, Shiyu Li, Weifeng Ge, Afshin Dehghan, Meng Cao, and Ping Huang. Streambridge: Turning your offline video large language model into a proactive streaming assistant, 2025a. URL https://arxiv.org/abs/2505.05467.

Haochen Wang, Qirui Chen, Cilin Yan, Jiayin Cai, Xiaolong Jiang, Yao Hu, Weidi Xie, and Stratis Gavves. Object-centric video question answering with visual grounding and referring, 2025b. URL https://arxiv.org/abs/2507.19599.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, Guanzhou Chen, Zichen Ding, Changyao Tian, Zhenyu Wu, Jingjing Xie, Zehao Li, Bowen Yang, Yuchen Duan, Xuehui Wang, Zhi Hou, Haoran Hao, Tianyi Zhang, Songze Li, Xiangyu Zhao, Haodong Duan, Nianchen Deng, Bin Fu, Yinan He, Yi Wang, Conghui He, Botian Shi, Junjun He, Yingtong Xiong, Han Lv, Lijun Wu, Wenqi Shao, Kaipeng Zhang, Huipeng Deng, Biqing Qi, Jiaye Ge, Qipeng Guo, Wenwei Zhang, Songyang Zhang, Maosong Cao, Junyao Lin, Kexian Tang, Jianfei Gao, Haian Huang, Yuzhe Gu, Chengqi Lyu, Huanze Tang, Rui Wang, Haijun Lv, Wanli Ouyang, Limin Wang, Min Dou, Xizhou Zhu, Tong Lu, Dahua Lin, Jifeng Dai, Weijie Su, Bowen Zhou, Kai Chen, Yu Qiao, Wenhai Wang, and Gen Luo. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency, 2025c. URL https://arxiv.org/abs/2508.18265.

Xin Wang, Taein Kwon, Mahdi Rad, Bowen Pan, Ishani Chakraborty, Sean Andrist, Dan Bohus, Ashley Feniello, Bugra Tekin, Felipe Vieira Frujeri, Neel Joshi, and Marc Pollefeys. Holoassist: an egocentric human interaction dataset for interactive ai assistants in the real world, 2023. URL https://arxiv.org/abs/2309.17024.

Yi Wang, Xinhao Li, Ziang Yan, Yinan He, Jiashuo Yu, Xiangyu Zeng, Chenting Wang, Changlian Ma, Haian Huang, Jianfei Gao, Min Dou, Kai Chen, Wenhai Wang, Yu Qiao, Yali Wang, and Limin Wang. Internvideo2.5: Empowering video mllms with long and rich context modeling, 2025d. URL https://arxiv.org/abs/2501.12386.

Yueqian Wang, Songxiang Liu, Disong Wang, Nuo Xu, Guanglu Wan, Huishuai Zhang, and Dongyan Zhao. Mmduet2: Enhancing proactive interaction of video mllms with multi-turn reinforcement learning, 2025e. URL https://arxiv.org/abs/2512.06810.

Yueqian Wang, Xiaojun Meng, Yifan Wang, Huishuai Zhang, and Dongyan Zhao. Proactivevideoqa: A comprehensive benchmark evaluating proactive interactions in video large language models, 2025f. URL https://arxiv.org/abs/2507.09313.

Yuxuan Wang, Yueqian Wang, Bo Chen, Tong Wu, Dongyan Zhao, and Zilong Zheng. Omnimmi: A comprehensive multi-modal interaction benchmark in streaming video contexts, 2025g. URL https://arxiv.org/abs/2503.22952.

Zikang Wang, Boyu Chen, Zhengrong Yue, Yi Wang, Yu Qiao, Limin Wang, and Yali Wang. Videochat-a1: Thinking with long videos by chain-of-shot reasoning, 2026. URL https://arxiv.org/abs/2506.06097.

Bo Wu, Shoubin Yu, Zhenfang Chen, Joshua B Tenenbaum, and Chuang Gan. Star: A benchmark for situated reasoning in real-world videos. *arXiv preprint arXiv:2405.09711*, 2024.

Weijia Wu, Yuzhong Zhao, Zhuang Li, Jiahong Li, Hong Zhou, Mike Zheng Shou, and Xiang Bai. A large cross-modal video retrieval dataset with reading comprehension, 2023. URL https://arxiv.org/abs/2305.03347.

Jiaer Xia, Peixian Chen, Mengdan Zhang, Xing Sun, and Kaiyang Zhou. Streaming video instruction tuning, 2026. URL https://arxiv.org/abs/2512.21334.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa:next phase of question-answering to explaining temporal actions, 2021. URL https://arxiv.org/abs/2105.08276.

Yuan Xie, Tianshui Chen, Zheng Ge, and Lionel Ni. Video-mtr: Reinforced multi-turn reasoning for long video understanding, 2026. URL https://arxiv.org/abs/2508.20478.

Haomiao Xiong, Zongxin Yang, Jiazuo Yu, Yunzhi Zhuge, Lu Zhang, Jiawen Zhu, and Huchuan Lu. Streaming video understanding and multi-round interaction with memory-enhanced knowledge, 2025. URL https://arxiv.org/abs/2501.13468.

Ruyi Xu, Guangxuan Xiao, Yukang Chen, Liuning He, Yao Lu, and Song Han. Streamingvlm: Real-time understanding for infinite video streams, 2026. URL https://arxiv.org/abs/2510.09608.

Biao Yang, Bin Wen, Boyang Ding, Changyi Liu, Chenglong Chu, Chengru Song, Chongling Rao, Chuan Yi, Da Li, Dunju Zang, Fan Yang, Guorui Zhou, Guowang Zhang, Han Shen, Hao Peng, Haojie Ding, Hao Wang, Haonan Fan, Hengrui Ju, Jiaming Huang, Jiangxia Cao, Jiankang Chen, Jingyun Hua, Kaibing Chen, Kaiyu Jiang, Kaiyu Tang, Kun Gai, Muhao Wei, Qiang Wang, Ruitao Wang, Sen Na, Shengnan Zhang, Siyang Mao, Sui Huang, Tianke Zhang, Tingting Gao, Wei Chen, Wei Yuan, Xiangyu Wu, Xiao Hu, Xingyu Lu, Yi-Fan Zhang, Yiping Yang, Yulong Chen, Zeyi Lu, Zhenhua Wu, Zhixin Ling, Zhuoran Yang, Ziming Li, Di Xu, Haixuan Gao, Hang Li, Jing Wang, Lejian Ren, Qigen Hu, Qianqian Wang, Shiyao Wang, Xinchen Luo, Yan Li, Yuhang Hu, and Zixing Zhang. Kwai keye-vl 1.5 technical report, 2025a. URL https://arxiv.org/abs/2509.01563.

Dongjie Yang, Suyuan Huang, Chengqiang Lu, Xiaodong Han, Haoxin Zhang, Yan Gao, Yao Hu, and Hai Zhao. Vript: A video is worth thousands of words, 2024. URL https://arxiv.org/abs/2406.06040.

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, Yuanhan Zhang, Peiyuan Zhang, Fangzhou Hong, Joerg Widmer, Francesco Gringoli, Lei Yang, Bo Li, and Ziwei Liu. Egolife: Towards egocentric life assistant, 2026. URL https://arxiv.org/abs/2503.03803.

Zhenyu Yang, Yuhang Hu, Zemin Du, Dizhan Xue, Shengsheng Qian, Jiahong Wu, Fan Yang, Weiming Dong, and Changsheng Xu. Svbench: A benchmark with temporal multi-turn dialogues for streaming video understanding, 2025b. URL https://arxiv.org/abs/2502.10810.

Dingyu Yao, Junhao Zhou, Chenxu Yang, Chuanyu Qin, Haowen Hou, Zheming Liang, Congcong Wang, Yuhang Cao, Shenglong Ye, Shuai Xie, Shuhuan Gu, Haoyang Huang, Qingyi Si, Nan Duan, and Jiaqi Wang. Joyai-vl-interaction: Real-time vision-language interaction intelligence, 2026. URL https://arxiv.org/abs/2606.14777.

Suyu Ye, Haojun Shi, Darren Shih, Hyokun Yun, Tanya Roosta, and Tianmin Shu. Realwebassist: A benchmark for long-horizon web assistance with real-world users, 2025. URL https://arxiv.org/abs/2504.10445.

Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes, 2023. URL https://arxiv.org/abs/2308.11417.

Yuqian Yuan, Hang Zhang, Wentong Li, Zesen Cheng, Boqiang Zhang, Long Li, Xin Li, Deli Zhao, Wenqiao Zhang, Yueting Zhuang, Jianke Zhu, and Lidong Bing. Videorefer suite: Advancing spatial-temporal object understanding with video llm, 2025. URL https://arxiv.org/abs/2501.00599.

Jun Zhang, Teng Wang, Yuying Ge, Yixiao Ge, Xinhao Li, Ying Shan, and Limin Wang. Timelens: Rethinking video temporal grounding with multimodal llms, 2026. URL https://arxiv.org/abs/2512.14698.

Yichi Zhang, Xin Luna Dong, Zhaojiang Lin, Andrea Madotto, Anuj Kumar, Babak Damavandi, Joyce Chai, and Seungwhan Moon. Proactive assistant dialogue generation from streaming egocentric videos. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 12044–12068. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.605. URL https://aclanthology.org/2025.emnlp-main.605/.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Llava-video: Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024.

Hang Zhao, Antonio Torralba, Lorenzo Torresani, and Zhicheng Yan. Hacs: Human action clips and segments dataset for recognition and temporal localization, 2019. URL https://arxiv.org/abs/1712.09374.

Yuhan Zhu, Changlian Ma, Xiangyu Zeng, Xinhao Li, Zhiqiu Zhang, Songze Li, Jun Zhang, Tianxiang Jiang, Yuandong Yang, Ziang Yan, Zikang Wang, Xinyu Chen, Haoran Chen, Shaowei Zhang, and Limin Wang. Timelens2: Generalist video temporal grounding with multimodal llms, 2026. URL https://arxiv.org/abs/2607.17423.

A Additional Discussion on Related Works

Turn-based Video LLMs. Recent offline video LLMs increasingly focus on effectively modeling long and structured visual contexts. InternVideo2.5 and Keye-VL-1.5 adapt the amount and granularity of visual evidence, whereas Qwen3-VL strengthens temporal grounding and InternVL3.5 advances general multimodal reasoning and efficiency (Wang et al., 2025d; Yang et al., 2025a; Bai et al., 2025; Wang et al., 2025c). VideoChat-A1 and Video-MTR further formulate evidence acquisition as an iterative process, selecting question-relevant shots or temporal segments over multiple reasoning steps (Wang et al., 2026; Xie et al., 2026). Nevertheless, these methods assume that the complete video is available before inference: context construction, evidence selection, and reasoning operate over a fixed input, while any multi-turn loop remains internal to the model or its tools. Consequently, they are not directly designed for environments that require immediate responses to a continuously arriving visual stream.

Streaming Video Understanding. Streaming video LLMs extend video understanding from fixed-input inference to causal processing over continuously arriving observations. One line of work

Table 5: Ablation study on IPIBench and FSI-Bench.

Configuration	IPIBench	FSI-Bench	
	Overall	D	C
Base data	21.20	62.01	35.47
Base data + contrastive data	31.2	88.08	48.56
Base data + contrastive data + CGBO	33.0	91.49	50.81

formulates streaming interaction primarily as a response-scheduling problem: determining whether the current observation should trigger language generation. VideoLLM-online introduces LIVE, which combines streaming supervision, interleaved video–text data, Streaming EOS prediction, and continuous KV-cache inference (Chen et al., 2024). DisPider decouples perception, decision, and reaction, using a lightweight monitoring module to trigger an asynchronous response model (Qian et al., 2025). MMDuet2 represents each turn as either a textual reply or `NO REPLY`, and applies multi-turn reinforcement learning with a PAUC-inspired reward to learn timely responses without requiring exact reply timestamps (Wang et al., 2025e). Another line treats response timing as one component of a broader interaction policy, extending the action space beyond simple response generation. Streamo learns `Silence`, `Standby`, and `Response` states together with multiple streaming instruction-following tasks (Xia et al., 2026). AURA combines chunk-wise interaction, dual sliding windows, a Silent-Speech Balanced Loss, and KV-cache reuse to support real-time, proactive, and multi-response question answering over long streams (Lu et al., 2026). JoyAI-VL-Interaction makes per-second decisions among silence, response, and delegation, coupling this policy with hierarchical memory and an asynchronous background model (Yao et al., 2026). ProAssist further constructs proactive assistant–user dialogues from streaming egocentric videos, providing supervision for proactive assistance in continuously observed environments (Zhang et al., 2025).

Complementary work improves sustained processing over long streams. StreamBridge uses compressed interaction memory and lightweight activation, and StreamingVLM emphasizes compact KV-cache management and training–inference alignment (Wang et al., 2025a; Xu et al., 2026). However, they do not jointly model the user-facing resolution of missing information: a model may answer, remain silent, trigger processing, or delegate, but it does not explicitly decide whether to request clarification, ask the user to perform a concrete action, or seek permission to continue monitoring, and then incorporate the returned feedback into subsequent decisions.

B Ablation Study

- **Group A** (Baseline): The base model trained on standard interaction data without contrastive samples or CGBO.
- **Group B** (+ Contrastive Data): Adds both action-strategy contrasts (§4) and response-gate contrasts (§4) to the training data, using standard cross-entropy loss.
- **Group C** (+ CGBO): The full model, adding CGBO conditional balancing (§4) on top of Group B.

Tab. 5 studies the contributions of contrastive training data and CGBO on both IPIBench and FSI-Bench. Starting from the base interaction data, adding action-strategy and response-gate contrastive examples yields a large improvement across both benchmarks, increasing IPIBench from 21.2 to 31.2 and substantially improving FSI-Bench decision and completion scores. This shows that explicitly contrasting when to answer, wait, or request feedback is critical for learning interactive streaming behavior. Adding CGBO further delivers the best overall results, improving IPIBench to 33.0 and achieving better performance on FSI-Bench, demonstrating its effectiveness in optimizing interactive streaming behaviors.

C Dataset Details

Source video datasets. ALIVE-234K draws streaming interaction trajectories from 41 publicly available video datasets spanning egocentric, third-person, and mixed domains. Table 6 lists each

source dataset together with its trajectory count. In total, 234,371 trajectories are constructed from these sources.

Table 6: Source video datasets used for constructing ALIVE-234K, sorted by trajectory count in descending order.

Dataset	# Trajectories
LLaVA-Video (Zhang et al., 2024)	31,364
TimeLens2 (Zhu et al., 2026)	30,499
TimeLens (Zhang et al., 2026)	22,158
Nymeria (Ma et al., 2024)	15,112
DenseStep2M (Ge et al., 2026)	14,046
VideoRefer-700K (Yuan et al., 2025)	13,576
VideoGPT+ (Maaz et al., 2024)	13,355
COIN (Tang et al., 2019)	12,069
EPIC-KITCHENS (Damen et al., 2020)	7,906
Vript (Yang et al., 2024)	7,131
HACS (Zhao et al., 2019)	6,900
QVHighlights (Lei et al., 2021)	6,019
Shot2Story (Han et al., 2025b)	5,958
Charades-Ego (Sigurdsson et al., 2018)	5,376
WebVid-2M (Bain et al., 2022)	4,789
Ego4D (Grauman et al., 2022)	4,620
Charades-STA (Gao et al., 2017)	4,088
HoloAssist (Wang et al., 2023)	4,082
DiDeMo (Hendricks et al., 2017)	3,751
Oops (Epstein et al., 2019)	3,278
Ego-Exo4D (Grauman et al., 2024)	3,036
Ref-YouTube-VOS (Seo et al., 2020)	2,960
RoomTour3D (Han et al., 2025a)	2,163
AVA (Gu et al., 2018)	2,115
MeViS (Ding et al., 2023)	1,800
ActivityNet (Caba Heilbron et al., 2015)	1,284
IndEgo (Chavan et al., 2025)	890
VideoInfer (Wang et al., 2025b)	665
QuerYD (Oncescu et al., 2021)	567
TextVR (Wu et al., 2023)	545
PerceptionTest (Pătrăucean et al., 2023)	480
EgoLife (Yang et al., 2026)	392
STAR (Wu et al., 2024)	330
NExT-QA (Xiao et al., 2021)	260
ScanNet++ (Yeshwanth et al., 2023)	228
EgoMask (Liang et al., 2025)	155
WebVid-10M (Bain et al., 2022)	148
EgoIT-99K (Yang et al., 2026)	148
TACoS (Regneri et al., 2013)	128
Total	**234,371**

D Implementation Details

We build two model variants, ALIVE-4B and ALIVE-8B, on top of Qwen3-VL-4B-Instruct and Qwen3-VL-8B-Instruct, respectively. The training data combines ALIVE-234K with 153K samples from open-source offline video instruction datasets. We detail the full data composition in Appendix C. During training, the vision encoder and the connector are frozen, and only the language model is updated. We train for a single epoch with a batch size of 256 and a learning rate of 1×10^{-5}. Input videos are sampled at 1 FPS. Following the instance construction in §4, we set the context window length to N=30 frames and the maximum post-anchor offset to M=4 positions.

E Additional Experimental Results

F FSI-Bench Details

This appendix documents the composition of FSI-Bench, the rollout protocol (§5), the judging procedure, and the intermediate metrics behind the success rate reported in §5.1. Table **??** reports the full breakdown.

F.1 Composition

Each episode is a causal trajectory: a video prefix, a query Q_{t_0} issued at the decision time t_0, the feedback that resolves it, an answer window $[t_s, t_e]$, and a reference answer. Frames are indexed by integer sampling steps. The prefix starts at the first frame shown to the annotator and never extends past t_0; every episode runs until $t_e + 3$ so that responses after the answer are also observed.

Table 7: Composition of FSI-Bench. t_1 denotes the step at which the user's clarification arrives.

Task	Episodes	Reference decision at t_0	Feedback in the stream	Answer window
Ask-for-Monitor	165	Wait	later visual evidence	annotated $[t_s, t_e]$
Ask-for-Information	450	Request (clarification)	user clarification at t_1	$[t_1, t_1]$
Ask-for-Execution	412	Request (action)	requested action performed in the video	annotated $[t_s, t_e]$
Answer (control)	200	Answer	none	$[t_0, t_0]$
Total	1,282			

Ask-for-Monitor. The queried information is not yet observable, and no immediate user action can reveal it. The reference decision is to wait and keep watching; the annotated window marks when the evidence first appears in the video. The prefix starts five steps before t_0.

Ask-for-Information. Several visible instances make the reference of the question ambiguous, and different instances yield different answers. The reference decision is a clarification request. The user's clarification, which identifies the intended target without revealing the answer, arrives at t_1, and the model must answer in that same step, i.e., $[t_s, t_e] = [t_1, t_1]$.

Ask-for-Execution. The target or region is identifiable, but the required visual evidence is hidden and a concrete user action can reveal it, such as opening a container. The reference decision is an execution request; the recorded action then unfolds in the video, and the annotated window marks when its outcome supports the answer.

Answer (control). Direct-answer controls are derived from the feedback-seeking cases so that the same kinds of questions are asked when the evidence is already visible: 100 from Ask-for-Information cases, rewritten to remove the ambiguity and asked at the frame of the clarification; 50 from Ask-for-Execution cases (10 per action category); and 50 from paired Ask-for-Execution and Ask-for-Monitor cases, asked once the action outcome is visible. The reference decision is to answer at t_0, and the episode runs until $t_0 + 3$. The control prefix starts at the same frame as its source case.

Verification. Ask-for-Execution, Ask-for-Monitor, and Answer episodes are converted from cases approved by human review; the unambiguous rewrites of the Ask-for-Information questions are generated automatically from human-approved question–answer pairs. Only episodes whose frames are complete from the prefix start to $t_e + 3$ are retained; missing frames are never padded with duplicates.

F.2 Rollout Protocol

Decision at t_0. The prefix frames and the query are given in each model's native input format, and the query turn is prefixed with the following instruction, identical for all models:

```
Use the video observed so far to decide what to do with the
user's question.  Choose exactly one:
```

```
A.Answer: Give the answer only when the visible evidence
already supports it. Never guess hidden contents.
B.Wait: Use only when further observation may help and no
specific user action or clarification can reveal the answer
now. Briefly say what is missing and ask whether the user
wants you to keep watching.
C.Request: Ask for the missing information or a specific
action that would reveal the evidence.
Output only A.Answer: <answer>, B.Wait: <brief monitoring
offer>, or C.Request: <request>.
Question: {query}
```

A reply that starts with one of the three labels is parsed directly; an empty reply or a bare silence token counts as a wait; any other reply is mapped to one of the three decisions by the LLM judge from the question and the reply text. The resulting decision determines how the rollout continues and is reused at scoring time.

Continuation. The instruction and the label are removed from the history: the query is kept verbatim, an answer, a request, or a wait with a monitoring offer is kept as the model's own utterance in its native response format, and a bare wait or silence is recorded as native silence. The rollout then advances one frame per step in the model's native streaming format, with its own system prompt and visual budget, and every output is recorded; the standby token of VideoChat3 is treated as silence. User feedback enters at its annotated step: the clarification as a user turn for Ask-for-Information, the performed action as video for Ask-for-Execution, and the later visual evidence for Ask-for-Monitor. For the Answer control, a decision other than answering ends the episode immediately with zero scores.

Qwen3-VL. Qwen3-VL has no native streaming format. It uses a neutral system prompt ("You are a streaming video assistant. Frames and user messages arrive over time; the last frame is the current moment. Use only what has been provided so far. A user message without text is just a new frame."), and each continuation turn appends the following instruction:

```
Given this streaming turn and the interaction history,
output exactly one of:
<silence>
<attention> your answer
Output <silence> if there is no ongoing task (it has not
started or is already complete), or if the current visual
information and user input are not enough to answer the
question.
Output <attention> followed by the answer if the user
input and visual information so far are enough to answer
a question that has not been answered yet.
Once a question has been answered, do not answer it again.
```

F.3 Judging and Metrics

Decisions and answer content are judged by an LLM; reply timing and response counts are computed from the recorded trajectory without any model call. The judge never sees whether the reply was on time.

Decision (D). The model's decision at t_0 is compared with the reference decision as in Table 8. A request on a request episode is additionally reviewed: the judge receives up to 10 frames sampled uniformly from the prefix start to t_0, the question, the candidate request, and the annotated reference request as an illustration of one valid way to close the gap rather than an exact-match target. It judges five criteria independently: *actionable* (specifies an action the user can perform now, or information the user can provide now), *informative* (complying provides the missing identity, context, or visual evidence), *reasonable* (relevant, feasible, safe, and proportionate to the scene), *grounded* (does not invent objects or states absent from the frames), and *not answer delegation* (does not ask the user to solve the question or supply its answer). All five must pass for $\mathrm{D} = 1$.

Timing (T). The scored reply is the first non-silent output after t_0, at time t_r; malformed outputs count as replies. For the Answer control, the scored reply is the output at t_0. Then $\mathrm{T} = w \cdot \mathbf{1}[t_r \in$

Table 8: Decision scoring. Rows are the model's decision at t_0 and columns the reference decision. "Stop" sets all four metrics to zero; "continue" evaluates the continuation. Silence at t_0 counts as Wait. Waiting on a request episode continues with weight $w = 0.5$ on T, A, and C.

Model decision	Reference: Answer	Reference: Wait	Reference: Request
Answer	D $= 1$, continue	0, stop	0, stop
Wait	0, stop	D $= 1$, continue	D $= 0.5$, continue ($w = 0.5$)
Request	0, stop	0, stop	D $\in \{0, 1\}$ by review, continue

$[t_s, t_e]]$, so that answering before the evidence is available, answering after the window closes, and never answering all score zero.

Accuracy (A). The judge receives up to 12 frames, the question and t_0, the scored reply and t_r, the reference answer with its annotated window, and the user feedback available by t_r, and returns a boolean; semantically equivalent wording is accepted. Frames are selected per task: for Ask-for-Execution and Ask-for-Monitor, $t_0 - 1$, t_0, and the annotated window, subsampled uniformly when long; for Ask-for-Information, frames from the prefix start to t_r, always including t_0 and t_r; for the Answer control, the evidence interval of the source case together with t_0. Then A $= w \cdot \mathbf{1}[\text{T} \wedge \text{correct}]$.

Completion (C). Let N_{extra} be the number of non-silent outputs in $(t_r,\, t_e + 3]$. Then C $= w \cdot \mathbf{1}[\text{A} \wedge N_{\text{extra}} = 0]$: for the feedback-seeking tasks, exactly one utterance after t_0, the answer itself; for the control, no utterance after the answer at t_0. C is the success rate reported in Table 4. By construction T $\geq$ A $\geq$ C, whereas D is not an upper bound on T because a request that fails the review continues with D $= 0$.

Aggregation. Each metric is averaged over the episodes of a task, and Avg. is the unweighted mean over the four task types.