BEFORE YOU ASK: LEARNING TO PROACTIVELY INITIATE INTERACTION IN STREAMING VIDEO

Anonymous authors
Paper under double-blind review

ABSTRACT

Streaming video assistants require model to provide timely, context-aware support while remaining silent when intervention is unnecessary. Existing methods remain fundamentally reactive, relying on explicit user queries to trigger responses, thus failing to autonomously initiate assistance in dynamic real-world scenarios. To bridge this gap, we propose a zero-trigger paradigm that shifts the objective from merely deciding *when to respond* to autonomously determining *whether, when, and how to intervene*. Specifically, we replace the conventional binary speak/silence actions with four functional tokens: `inquiry`, `response`, `notify`, and `silence`. To realize this, we introduce **SAIWen**, a streaming multimodal model post-trained via a two-stage framework on **SAIWen-Interact**, a dataset of 26,549 training and 2,000 test videos with temporally aligned dialogues. During Supervised Fine-Tuning (SFT), our a Hierarchical Class-Balanced Loss decouples macro-level intervention triggering from micro-level functional routing, mitigating extreme functional token imbalance without degenerating into trivial always-silent policies. Subsequently, Reinforcement Learning (RL) with a Curriculum Advantage Annealing directly optimizes temporal alignment against our MII-Score, which rigorously evaluates proactive inquiry quality and notification precision while penalizing passive abstention. Experiments demonstrate that **SAIWen** significantly advances autonomous video assistance, improving intervention precision by 30.32% and notification timeliness by 14.90%, while strictly minimizing interruption costs.

1 INTRODUCTION

The rapid advancement of multimodal large language models (MLLMs) (Seed, 2026; Team, 2026) has fundamentally transformed video understanding, evolving from offline clip-level classification toward real-time, streaming comprehension of continuous visual input (Chen et al., 2024; Ding et al., 2025). In this streaming paradigm, a model ingests frames as they arrive and responds to user queries at arbitrary moments, mirroring human perceptual dynamics. Building upon this, recent research (Li et al., 2026; Xia et al., 2026; Fu et al., 2025) has shifted attention from *what a model can understand* to *when to respond*, giving rise to the concept of *proactive agency* (Li et al., 2026), wherein the model autonomously determines both the timing and content of its responses in a live video stream. For instance, once instructed (e.g., to monitor a cooking procedure), a model autonomously determines the timing and content of its interventions. Despite this progress, existing proactive approaches (Hurst et al., 2024; Yao et al., 2026) share a critical limitation: their autonomy is strictly anchored to an initial explicit user request. Regardless of subsequent behavior, these models remain reactive at the outset, waiting for a command before monitoring the stream. We argue that true proactive agency in ambient intelligence necessitates a qualitatively new capability: **zero-trigger, model-initiated interaction**. Rather than waiting for instructions, the model must continuously interpret unfolding activities, anticipate potential needs, and autonomously initiate dialogue. For instance, by observing a chef struggling with a recipe, the model could proactively issue an `<inquiry>` to offer guidance, as illustrated in Figure 1.

Realizing this zero-trigger paradigm introduces three fundamental challenges. First, the model must transition from passive response generation to autonomous intervention triggering, grounding its decisions in sufficient visual evidence to balance anticipation against premature intrusion. Second, unlike prior binary silence-speak formulations (Xia et al., 2026; Li et al., 2026), model-initiated interaction demands decomposing the "speak" action into fine-grained functional speech acts to

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.

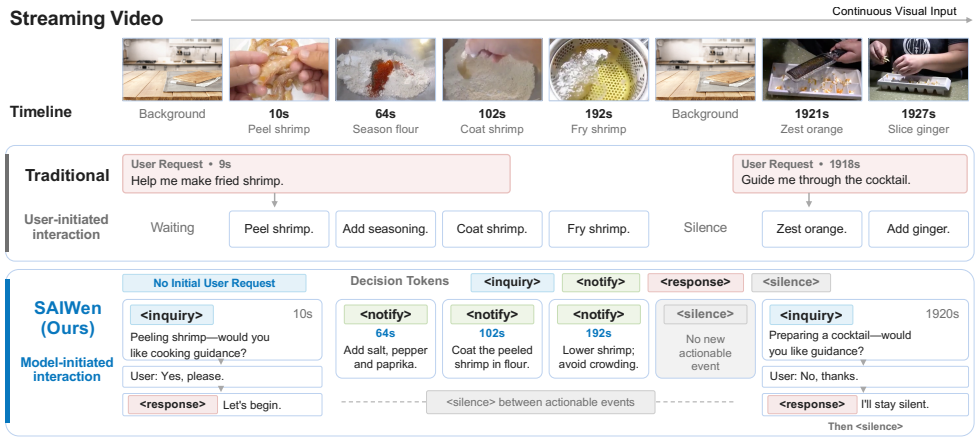


Figure 1: **Traditional user-initiated vs. our model-initiated interaction in streaming video. Top:** Traditional pipelines remain reactive, requiring user requests to trigger guidance. **Bottom:** SAIWen operates without initial prompts, autonomously emitting functional tokens based on continuous visual input. It proactively issues an `<inquiry>` to confirm cooking intent, subsequently delivering timely `<notify>` tokens aligned with unfolding steps. Crucially, the model maintains `<silence>` between non-actionable intervals and gracefully handles rejected assistance via `<response>`, demonstrating precise temporal alignment and minimal interruption costs.

differentiate proactive inquiries from task-level notifications. Third, achieving optimal interaction requires jointly optimizing semantic accuracy, temporal precision, and minimal interruption costs, which is a multi-objective problem that existing evaluation metrics fail to capture.

To address these challenges, we propose a novel task formulation where the model autonomously decides *whether, when, and how* to intervene. Specifically, we replace the conventional binary decision space with four functional tokens: `inquiry` (autonomously soliciting user intent), `response` (reactively answering user requests), `notify` (delivering task-oriented assistance), and `silence` (continuous passive observation). Building upon a general-purpose MLLM, we introduce **SAIWen**, a streaming video model post-trained via a two-stage framework. To support this, we establish a data pipeline synthesizing an autonomous assistance dataset of 26,549 temporally aligned video dialogues (over 660 hours). During Supervised Fine-Tuning (SFT), we propose a Hierarchical Class-Balanced Loss decoupling macro-level intervention triggering from micro-level functional routing, mitigating extreme token imbalance. Subsequently, Reinforcement Learning (RL) with Curriculum Advantage Annealing directly optimizes temporal alignment against our **MII-Score**. This metric rigorously evaluates proactive inquiry quality and event notification precision while eliminating the reward for degenerate always-silent policies. To summarize, our contributions are as follows:

- We propose a model-initiated interaction paradigm for streaming video understanding, replacing binary speak/silence actions with four functional tokens (`inquiry`, `response`, `notify`, `silence`). This shifts the objective from reactive timing to autonomous determination of *whether, when, and how to intervene*.
- We introduce a unified post-training framework featuring Hierarchical Class-Balanced Loss to resolve token imbalance and Curriculum Advantage Annealing for robust policy optimization. Additionally, we release **SAIWen-Interact**, a systematic dataset with dedicated splits, alongside the **MII-Score** for rigorous proactive evaluation.
- Extensive experiments demonstrate that **SAIWen** significantly advances autonomous assistance, improving intervention precision by 30.32% and notification timeliness by 14.90%, while strictly minimizing interruption costs.

2 Related Work

Streaming Video Understanding. Driven by foundation models' growing capabilities in offline video understanding (Zhang et al., 2023; Xue et al., 2022; Hurst et al., 2024; Team, 2026), the field is shifting towards streaming video understanding. Building upon this, proactive streaming under-

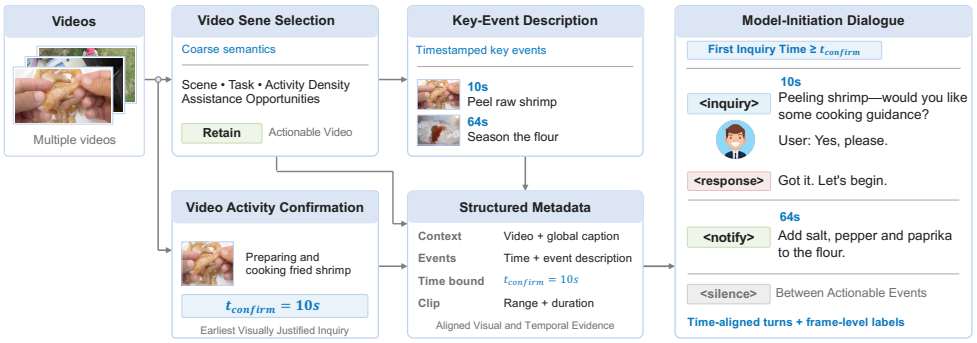


Figure 2: **Overview of data construction pipeline.**

standing (Chen et al., 2024; Xia et al., 2026; Li et al., 2026) has emerged, aiming to enable models to autonomously determine optimal response timing based on incoming streams. Current approaches broadly fall into two paradigms. (1) Token-based proactive models (Xia et al., 2026; Li et al., 2026) empower models to decide silence or response timing via token generation. For instance, videollm-online Chen et al. (2024) generates response tokens on the fly within continuous streams, while StreamO (Xia et al., 2026) introduces a "standby" mechanism allocating preparation time for versatile real-time interactions. (2) Module-based proactive frameworks (Qian et al., 2025; Yao et al., 2026; Ding et al., 2025) incorporate auxiliary components into unified architectures. Dispider (Qian et al., 2025), for example, employs a decoupled design where a lightweight asynchronous module independently monitors streams to trigger interactions. However, we rethink this paradigm to propose a more proactive Vision-Language Model.

Streaming Video Benchmark. Several specialized benchmarks have recently been introduced to systematically evaluate streaming video understanding. StreamingBench (Lin et al., 2026) pioneers comprehensive evaluation by simulating continuous streams with timestamped questions to assess real-time, omni-source, and contextual understanding. OVBench (Huang et al., 2025) evaluates perception, memorization, and reasoning across past, current, and future temporal contexts. ODV-Bench (Zeng et al., 2026) targets domain-specific real-time comprehension in autonomous driving. Furthermore, OVO-Bench (Niu et al., 2025) emphasizes temporal awareness, assessing dynamic response adaptation based on query timestamps, including forward active responding (delaying responses until sufficient future context is available). We use these benchmarks to validate basic streaming comprehension, alongside our own to assess proactive interaction.

3 Model-Initiated Interaction in Streaming Video

3.1 Problem Statement

Traditional video understanding operates under an offline, single-turn paradigm, mapping a complete video $V = \{v_1, \ldots, v_T\}$ and a user query Q directly to an answer A. Recent advances in streaming video understanding mitigate latency by enabling models to dynamically determine *when* ($t \leq T$) to intervene. Nevertheless, these approaches remain fundamentally reactive, as their interventions are strictly conditioned on explicit initial user queries.

To transcend this limitation, we propose **Model-Initiated Interaction**, a zero-trigger paradigm where the agency to initiate dialogue is driven entirely by the model. Operating continuously over the visual stream, the model autonomously interprets unfolding events based on partial observations $V_{\leq t}$. Upon recognizing an opportunity for assistance, it issues an `inquiry` token. If affirmed, it sequentially emits a `response` token to confirm engagement and a `notify` token to deliver guidance before reverting to `silence`; if declined, it immediately suppresses intervention to minimize interruption costs. Figure 1 illustrates this workflow.

3.2 Data Construction

Data Construction Pipeline. Models require training data comprising temporally aligned video-dialogue pairs. As the research community currently lacks publicly available datasets tailored for model-initiated streaming interactions, we establish a systematic data construction pipeline (Fig-

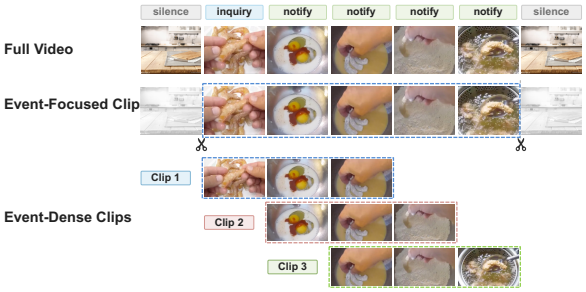


Figure 3: Temporal windowing Cropping.

Table 1: Dialouge-based temporal windowing cropping mitigates decision-token class imbalance.

View composition	Silence:Speak
Full only—*no windowing*	**20.51**
Event-centric views added by windowing	
Event-focused	18.91
Event-dense	15.78
Event-focused + Event-dense	15.90
Full + event-centric views (ours)	**17.72**

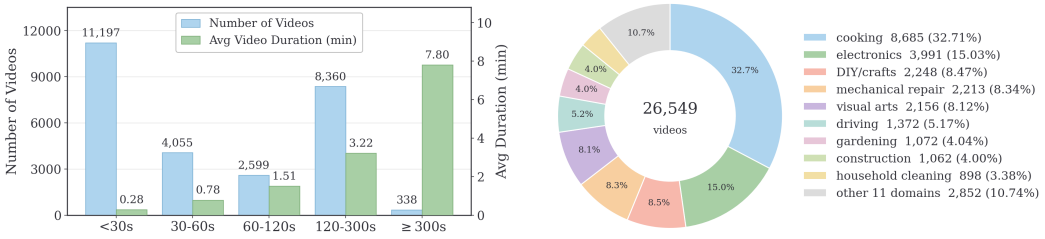
ure 2). Given an input video, our objective is to densely annotate interaction triggers and responses. The pipeline comprises five steps: (1) **Video Scene Selection.** We employ a Vision-Language Model (VLM) to generate global descriptions and perform semantic analysis over the video collection, identifying videos with clear tasks and high activity density suitable for interaction, filtering out static demonstrations and monologues. (2) **Key Event Description.** The selected videos undergo fine-grained analysis. Guided by the global descriptions, the VLM identifies specific key events (e.g., peeling shrimp, adding seasonings) and generates detailed annotations for frames. These descriptions provide the necessary context to ground subsequent notifications. (3) **Video Activity Confirmation.** Concurrently with Step (2), we prompt the VLM to determine the earliest timestamp, t_{confirm}, at which the core activity can be reliably confirmed. This timestamp defines the optimal temporal boundary for initiating user engagement. (4) **Interactive Dialogue Generation.** Integrating the global description, activity theme, and key event annotations, we synthesize structured metadata into sequence: opening inquiry, simulated user response, the model's acknowledgment, and task-related notifications. (5) **Dialogue Data Validation.** Dialogues are verified for temporal accuracy and role validity. Utterances are assigned functional decision tokens (`inquiry`, `response`, `notify`), while non-speaking frames are labeled as `silence`.

Dialogue-based Temporal Window Cropping. In long videos, salient events are sparse, causing `silence` frames to overwhelmingly outnumber active tokens. This severe class imbalance risks diluting the dialogue signals, thereby hindering the model's ability to learn proactive initiation. To mitigate this class imbalance, we apply temporal window cropping to generate **Event-Focused Clips** (Figure 3), which spans from 5 seconds before the `inquiry` event to 20 seconds after the final `notify` event. Additionally, to preserve the model's general video understanding capabilities, we synthesize **Event-Dense Clips** from the original dialogues. Each such clip encompasses three consecutive `notify` events, spanning from 10 seconds before the first event to 15 seconds after the third, as detailed in Figure 3. As shown in Table 1, this strategy effectively alleviates the imbalance between `silence` and active speaking labels.

SAIWen-Interact Dataset. We introduce **SAIWen-Interact**, a dataset for model-initiated streaming video interaction comprising 26,549 training and 2,000 test videos, paired with temporally aligned dialogues and frame-level supervision for four actions: `inquiry`, `response`, `notify`, and `silence`. Figure 4(a) illustrates the duration distribution of the training set (662.7 hours). We deliberately integrate short online clips with longer procedural and egocentric recordings to cultivate distinct capabilities. Short clips enable computationally efficient, dense coverage of diverse triggers, while long videos simulate the sparse interaction density of real-world proactive assistants, capturing extended inactivity and multi-task tracking that short clips cannot represent. Furthermore, we categorize the training split into 20 domains (Figure 4(b)). *Cooking* constitutes the largest domain (8,685 videos, 32.71%), followed by *Electronics* (15.03%), *DIY/Crafts* (8.47%), *Mechanical Repair* (8.34%), and *Visual Arts* (8.12%). The remaining 15 domains account for 7,256 videos (27.33%). Notably, this distribution is driven strictly by each video's suitability for proactive conversation initiation, rather than raw source material volume.

3.3 Evaluation Metric

MII-Score. Conventional evaluation protocols are inherently reactive, assessing only whether a model responds correctly to explicit requests. For instance, PAUC-based metrics (Wang et al., 2025) measure temporal alignment within target events but fail to penalize excessive interventions, nor evaluate initial zero-trigger inquiries. We introduce the **Model-Initiated Interaction Score (MII-**



(a) Distribution of Video Duration (b) Distribution of Data Scenarios

Figure 4: **Statistics of constructed the training split of SAIWen-Interact Dataset.**

Score) to assess two aspects of the model: inquiry quality and notification utility. Specifically, Q_{inquiry} evaluates whether the initial intervention identifies ongoing activities and offers relevant assistance, while U_{notify} combines timeliness-aware semantic accuracy with precision to penalize redundant emissions. MII-Score is their arithmetic mean:

$$\text{MII-Score} = \frac{Q_{\text{inquiry}} + U_{\text{notify}}}{2}. \tag{1}$$

Inquiry Score Q_{inquiry}. Q_{inquiry} measures autonomous dialogue initiation quality. Given the first `inquiry` token and ground-truth clip context (detailed in Appendix A.1), a judge model rates on a five-point scale whether it accurately identifies ongoing activities and offers relevant assistance, normalized to $[0, 1]$. If no valid `inquire` token is produced, $Q_{\text{inquiry}} = 0$. We impose no strict temporal constraint on the inquiry, provided it precedes the first `notify` action.

Notification Score U_{notify}. For event evaluation, given N ground-truth events annotated at timestamps $\{\tau_j\}_{j=1}^{N}$, a `Notify` token emitted by the model at time t_i is admissible for event j only if it falls within the tolerance window:

$$W_j = [\tau_j - \delta_{\text{pre}},\ \tau_j + \delta_{\text{post}}], \tag{2}$$

where δ_{pre} and δ_{post} define the maximum permissible anticipation and delay margins, respectively. Let $q_{ij} \in [0, 1]$ denote the semantic matching quality between the i-th notification and the j-th event. The instantaneous quality attained at time t for event j is defined as the prefix maximum over all admissible notifications emitted up to time t:

$$g_j(t) = \max_{\{\, i\,:\, t_i \in W_j,\ t_i \le t \,\}} q_{ij}, \qquad \max \emptyset \triangleq 0. \tag{3}$$

This formulation ensures that once a high-quality notification is emitted, its credit is preserved for the window's remainder; subsequent redundant or lower-quality notifications cannot degrade it. The event notification score U_{notify} decomposes interaction quality into timeliness-aware accuracy (E_{PAUC}) and intervention precision (P_{notify}). E_{PAUC} integrates the best semantic quality achieved so far as a non-decreasing, right-continuous step function. Averaging this time-integrated quality over the window yields the score for event j's score, and averaging across all N events gives:

$$E_{\text{PAUC}}^{(j)} = \frac{1}{\delta_{\text{pre}} + \delta_{\text{post}}} \int_{\tau_j - \delta_{\text{pre}}}^{\tau_j + \delta_{\text{post}}} g_j(t)\, \mathrm{d}t, \qquad E_{\text{PAUC}} = \frac{1}{N} \sum_{j=1}^{N} E_{\text{PAUC}}^{(j)}, \tag{4}$$

with $E_{\text{PAUC}} = 0$ when $N = 0$. Equation 4 enforces a linear decay in accuracy with respect to latency. A fully correct notification ($q_{ij} = 1$) achieves a perfect score at the window onset and decays proportionally to the delay. Anticipation is rewarded strictly within δ_{pre}; notifications outside W_j are inadmissible, preventing indiscriminate early guessing from being scored as foresight.

To penalize verbosity and minimize interruption costs, P_{notify} evaluates the efficiency of the model's actual emissions. Given M total `notify` tokens emitted across the video, we compute the optimal one-to-one bipartite matching $\mathcal{A}$ between notifications and events to prevent repeated notifications for the same event from accumulating credit:

$$Q^* = \max_{\mathcal{A}} \sum_{(i,j)\in\mathcal{A}} q_{ij}, \qquad P_{\text{notify}} = \begin{cases} Q^*/M, & M > 0, \\ 0, & M = 0. \end{cases} \tag{5}$$

Every unmatched or low-quality notification enlarges the denominator M without raising Q^*, thereby heavily penalizing redundancy. Finally, U_{notify} is defined as the harmonic mean of E_{PAUC} and P_{notify}, ensuring a high score requires both timely accuracy and non-redundancy:

$$U_{\text{notify}} = \begin{cases} \dfrac{2\,E_{\text{PAUC}}\,P_{\text{notify}}}{E_{\text{PAUC}} + P_{\text{notify}}}, & E_{\text{PAUC}} + P_{\text{notify}} > 0, \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

Crucially, unlike prior formulations, no reward is assigned solely for remaining silent throughout the video. A fully passive policy yields $M = 0$, resulting in $U_{\text{notify}} = P_{\text{notify}} = 0$. Thus, the model cannot achieve a high MII-Score by abstaining from proactive engagement.

4 Training Framework

We develop **SAIWen** by post-training a general-purpose VLM for model-initiated streaming video interaction. This paradigm shift entails three challenges: (1) transitioning from passive response to autonomous initiation; (2) expanding fine-grained actions; and (3) ensuring both content accuracy and temporal precision. We address these via a two-stage training: SFT with Hierarchical Class-Balanced Loss injects new token representations, while RL optimizes proactive interaction quality.

4.1 Supervised Fine-Tuning with Hierarchical Class-balanced Loss

Training Data Composition. To equip the model with proactive interaction capabilities while preserving offline and streaming video understanding, we organize the training data into three main components. First, we incorporate 9,645 offline samples for standard video QA and captioning without functional-token supervision, preserving general vision-language representations. Second, we include 7,731 streaming instances from StreamO (Xia et al., 2026) to retain streaming video understanding capabilities. Third, we use 26,549 model-initiated interaction dialogues from the training split of SAIWen-Interact, providing dense supervision across all four functional tokens (`inquiry`, `notify`, `response`, and `silence`). We additionally include 922 user-rejection dialogues to teach the model to return to `silence` when assistance is declined. Together, these components form an SFT mixture of 44,847 training instances.

Hierarchical Class-balanced Loss. Expanding the conventional binary functional space into fine-grained space $\mathcal{S} = \{\texttt{silence}\} \cup \mathcal{S}_{\text{act}}$ transforms the learning objective into a hierarchical problem: macro-level intervention triggering (`silence` vs. $\mathcal{S}_{\text{act}}$) and micro-level functional routing within $\mathcal{S}_{\text{act}} = \{\texttt{inquiry}, \texttt{notify}, \texttt{response}\}$. Treating the three active tokens as independently rare classes would grant them a disproportionate aggregate gradient budget, destabilizing optimization. To address this, we propose a Hierarchical Class-Balanced Loss ($\mathcal{L}_{HCB}$). We first decouple the gradient contributions by assigning a macro-level prior $1 - \rho$ to passive observation and ρ to all active interventions. The effective hierarchical weight $\hat{w}_k$ for token k is formulated as:

$$\beta_{\text{silence}} = 1 - \rho, \qquad \beta_k = \rho \cdot \frac{n_k}{n_{\text{act}}} \;\; (k \in \mathcal{S}_{\text{act}}), \qquad \hat{w}_k = \frac{N \cdot \beta_k}{n_k}, \tag{7}$$

where n_k is the corpus-level frequency of token k, $N = \sum_{k \in \mathcal{S}} n_k$, and $n_{\text{act}} = \sum_{k \in \mathcal{S}_{\text{act}}} n_k$. Since $n_k \hat{w}_k = N \beta_k$, hyperparameter ρ strictly governs the gradient budget proportion for active speech acts. Crucially, all active tokens share the per-token weight $N\rho / n_{\text{act}}$ rather than individual amplification, preventing rare tokens from dominating optimization. To further suppress the overwhelming gradient contribution from easily classified `silence` frames, we integrate a focal modulation term directly into the autoregressive training objective:

$$\mathcal{L}_{\text{HCB}} = -\frac{1}{M} \sum_{i=1}^{M} w_i \cdot \underbrace{(1 - p_\theta(y_i \mid y_{<i}, x))^\gamma}_{\text{focal modulation}} \cdot \log p_\theta(y_i \mid y_{<i}, x), \tag{8}$$

where the base weight w_i is dynamically assigned based on the token type:

$$w_i = \begin{cases} \hat{w}_{y_i}, & y_i \in \mathcal{S}, \\ \lambda_{\text{text}}, & y_i \in \mathcal{T}, \end{cases} \tag{9}$$

with $\mathcal{T}$ denoting the natural language vocabulary, λ_{text} balancing decision learning against text generation, and $\gamma = 2$ serving as the focal exponent. This unified formulation ensures that the policy concentrates learning capacity exclusively on hard, temporally critical decisions, effectively preventing the model from degenerating into a trivial always-silent policy.

4.2 Reinforcement Learning with Advantage Annealing

To transition from the foundational policy established by $\mathcal{L}_{\text{HCB}}$ to optimal temporal alignment, we employ RL to directly optimize the quality and precision metrics underlying our MII-Score.

Reward Formulation. We define the scalar reward as the unnormalized sum of the three measures our MII-Score, rather than the composite index: $R = Q_{\text{inquiry}} + E_{\text{PAUC}} + P_{\text{notify}}$. Crucially, silence-region reliability and false-`notify` counts are logged strictly for diagnostic purposes and are excluded from R. By ensuring the reward never compensates the model for abstaining from interaction, we eliminate the degenerate always-silent optimum.

Curriculum Advantage Annealing. In the early stages of RL, valid `notify` actions are exceedingly sparse. A static reward dominated by E_{PAUC} and P_{notify} yields near-zero within-group variance, providing insufficient learning signals. To address this, we propose a **Curriculum Advantage Annealing** strategy that smoothly shifts the optimization focus from functional coverage to temporal utility. For a group of K rollouts sampled from the same video context, we compute the advantage A_i for the i-th rollout via within-group standardization ($z(\cdot)$) across each metric axis:

$$A_i = \begin{cases} 0.2\, z(Q_{\text{inquiry},i}) + 0.6\, z(P_{\text{notify},i}) + 0.2\, z(E_{\text{PAUC},i}), & \text{Phase I (Exploration)}, \\ 0.2\, z(Q_{\text{inquiry},i}) + 0.3\, z(P_{\text{notify},i}) + 0.5\, z(E_{\text{PAUC},i}), & \text{Phase II (Transition)}, \\ 0.2\, z(Q_{\text{inquiry},i}) + 0.8\, z(E_{\text{PAUC},i}), & \text{Phase III (Exploitation)}. \end{cases} \tag{10}$$

This phased weighting anneals the objective: Phase I prioritizes notification precision (P_{notify}) to discourage redundant notifications; Phase II balances precision with timeliness; and Phase III rewards early, accurate event confirmation (E_{PAUC}). Metric axes exhibiting zero within-group variance (e.g., absent ground-truth events exist in a window) naturally yield $z(\cdot) = 0$ and are implicitly masked without disrupting gradient flow. The policy is updated using this advantage with a learning rate scheduler comprising a 5% linear warmup followed by cosine decay to 10% of the peak rate.

5 Experiments

SAIWen is initialized from Qwen3.5-4B and post-trained via SFT and RL on four NVIDIA H20 GPUs. SFT uses `ms-swift` with BF16 precision, gradient checkpointing, and DeepSpeed ZeRO-2. We freeze the vision encoder while applying LoRA ($r = 64$, $\alpha = 128$) to remaining parameters. Videos are sampled at 2 FPS into sliding windows of up to 200 frames at 384×384 resolution, concatenated with timestamps and dialogue history. The learning rate is 5×10^{-5} with a per-device batch size of 1 and 4 gradient accumulation steps. For decision-token supervision, we employ $\mathcal{L}_{\text{HCB}}$, setting the macro-level speak-to-silence weight ratio to 7.3 and focal exponent $\gamma = 2$.

5.1 Main Results

Table 2 evaluates models on the 2,000-video SAIWen-Interact test set, where SAIWen achieves an MII-Score of 36.37, substantially outperforming the strongest baseline JoyAI-VL-Interaction (Yao et al., 2026) (14.83). This superiority is driven by a markedly higher Q_{inquiry} (56.68 vs. 28.00), reflecting robust dialogue initiation, alongside consistent E_{PAUC} gains of 7.65, 6.81, and 10.39 points across progress monitoring, technical guidance, and safety alerting. These improvements confirm that SAIWen more effectively converts continuous visual evidence into timely, task-relevant interventions under our protocol. Notably, while Omni models demonstrate competitive E_{PAUC} in task assistance, it lags in response accuracy P_{notify} and overall interaction. This divergence highlights distinct positioning of the respective models.

SAIWen's proactive initiation is grounded in its general video understanding; consequently, we evaluate the model on four query-based streaming benchmarks (Table 3). Relative to the Qwen3.5-4B baseline, SAIWen achieves substantial gains of 14.26, 17.58, 13.13, and 23.01 on OVOBench (Niu et al., 2025), ODVBench (Zeng et al., 2026), OVBench (Huang et al., 2025), and

Table 2: **Comparison of models on the SAIWen-Interact test set.**

Method	Params	Assistance by Category ($E_{\text{PAUC}}/P_{\text{notify}}$)				Overall Interaction		
		Progress Monitoring	Technical Guidance	Safety Alerting	Reminder	Q_{query}	U_{notify}	MII-Score
Offline Models + Proactive Prompt								
Qwen3.5	4B	1.32 / 1.06	1.03 / 1.00	1.11 / 1.17	1.03 / 0.55	10.95	1.15	6.05
Qwen3.5	9B	0.93 / 0.80	1.23 / 0.79	1.35 / 1.08	0.42 / 0.44	8.60	1.17	4.89
Qwen2.5-Omni (Xu et al., 2025)	7B	10.68 / 1.70	10.04 / 1.56	8.36 / 1.35	5.60 / 0.85	23.77	2.17	12.97
Qwen3-Omni	30B-A3B	34.26 / 3.88	34.01 / 3.37	30.07 / 3.41	29.38 / 3.00	21.93	4.87	13.40
Open-source Streaming Models								
JoyAI-VL-Interaction (Yao et al., 2026)	8B	14.67 / 1.83	14.79 / 1.67	10.22 / 1.51	18.13 / 1.85	28.00	1.65	14.83
Model-Initiated Interaction Model								
SAIWen (Ours)	4B	22.32 / 19.86	21.60 / 19.72	20.61 / 18.32	15.42 / 14.74	**56.68**	**16.05**	**36.37**

Table 3: **Comparison of models across streaming video understanding benchmarks.**

Model	Params	Train Data	Frames	OVOBench	ODVBench	OVBench	StreamingBench$_{\text{RT}}$
Closed-source Streaming Models							
Gemini 1.5 Pro (Team et al., 2024)	–	–	1fps	63.00	–	–	75.69
GPT-4o (Islam & Moushi, 2025)	200B	–	64	59.54	–	–	73.28
Offline Models							
Qwen3.5	4B	–	1fps/64	38.31	35.73	35.11	53.90
Qwen2.5-Omni (Xu et al., 2025)	7B	–	1fps/64	54.81	57.48	49.77	81.14
Qwen3-Omni-30B-A3B-Instruct	30B	–	1fps/64	61.44	57.83	54.50	81.47
Streaming Models							
VideoLLM-online (Chen et al., 2024)	8B	134K	2fps	12.84	–	9.61	35.99
Dispider (Qian et al., 2025)	7B	122K	1fps	44.76	46.01	42.20	69.94
VideoChat-Online (Huang et al., 2025)	4B	96K	1fps	43.97	53.65	55.20	66.76
StreamForest (Zeng et al., 2026)	7B	491K	1fps	53.09	60.22	60.50	77.81
JoyAI-VL-Interaction (Yao et al., 2026)	8B	4M	1fps/64	54.68	71.22	59.51	79.58
SAIWen (Ours)	4B	45K	1fps	52.56	53.28	48.23	76.91
Δ vs. Qwen3.5-4B				+14.25	+17.55	+13.12	+23.01

Table 4: **Ablation study of the Hierarchical Class-Balanced Loss.**

$w_{\text{spk}}/w_{\text{sil}}$	Focal	Q_{query}	E_{PAUC}	P_{notify}	MII-Score
5.00	✓	46.75	9.94	**8.16**	27.42
10.00	✓	47.00	**13.65**	8.08	28.03
7.26	×	30.25	2.07	3.47	16.26
7.26	✓	**55.30**	6.97	5.65	**30.05**

Table 5: **Ablation study of the temporal matching-window.**

Setting	Temporal Window (s)		Evaluation Metrics			
	Δ_{pre}	Δ_{post}	Q_{query}	E_{PAUC}	P_{notify}	MII-Score
W1	−2	+3	54.00	7.28	5.78	29.81
W2	−2	+5	54.00	10.09	7.96	30.66
W3	−2	+8	54.01	13.37	9.26	31.68

StreamingBench$_{\text{RT}}$ (Lin et al., 2026), respectively. These results confirm that mastering proactive interaction patterns does not degrade explicit query-based performance; rather, it enhances temporal reasoning. Notably, despite being post-trained on only 45K dialogues, significantly fewer than the corpora for closed-source or offline streaming models, SAIWen remains highly competitive. This demonstrates that our data pipeline and training framework effectively support both model-initiated interaction and traditional query-driven video understanding. Details are in Appendix A.3.

5.2 Analysis

All ablations are evaluated on an identical 100-video subset of the SAIWen-Interact test set.

Ablation Study of Hierarchical Class-Balanced Loss. Table 4 evaluates the Hierarchical Class-Balanced objective on the shared test subset. The default speak-to-silence per-token loss-weight ratio $w_{\text{spk}}/w_{\text{sil}}$ is derived from $\rho = 0.25$ using the training-token counts and rounded to 7.26 for display; we directly ablate this ratio rather than ρ. At the default ratio, adding focal modulation increases MII-Score from 16.26 to 30.05 (+13.79 points), with the largest gain in inquiry quality Q_{inquiry} (+25.05 points). Among focal-enabled settings, the default ratio achieves the highest MII-Score, exceeding ratios of 5 and 10 by 2.63 and 2.02 points, respectively. Although these alternatives achieve higher notification precision, and a ratio of 10 yields the highest E_{PAUC}, the default ratio provides the strongest inquiry quality and the best overall MII-Score among the evaluated settings.

Effect of the Temporal Window. To quantify the temporal latency of SAIWen's interventions relative to ground-truth events, Table 5 presents an ablation on the post-event tolerance margin δ_{post}. A smaller δ_{post} imposes stricter constraints on notification timeliness. Empirically, tightening δ_{post} to +3s degrades E_{PAUC} from 13.37 to 7.28, accompanied by a concurrent drop in P_{notify} from 9.26 to 5.78. This reduction is attributed to the narrowed temporal window, which restricted the pool of

Table 6: **Reinforcement learning ablation on the shared subset of the SAIWen-Interact test set.**

Setting	Initialization	Reward Curriculum	Q_{query}	E_{PAUC}	P_{notify}	MII-Score
SFT	SFT checkpoint	None	**55.30**	6.97	5.65	30.05
GRPO	Identical SFT checkpoint	Fixed reward weights	51.25	**14.47**	7.46	30.54
GRPO + Annealing (Ours)	Identical SFT checkpoint	Three-stage curriculum	54.01	13.37	**9.26**	**31.68**

eligible notifications for matching, thereby diminishing the aggregate quality score Q. This reveals a persistent temporal lag between the model's responses and the ground-truth timing. Consequently, we select Setting W3 as the default temporal window parameters.

Ablation on Reinforcement Learning with Advantage Annealing. Table 6 compares GRPO variants initialized from the identical SFT checkpoint and trained for equivalent steps. Relative to SFT, fixed-weight GRPO improves E_{PAUC} (from 6.97 to 14.47) and P_{notify} (from 5.65 to 7.46), yet degrades inquiry quality Q_{inquiry} (from 55.30 to 51.25). In contrast, Curriculum Advantage Annealing preserves a higher Q_{inquiry} (54.01) and further elevates P_{notify} to 9.26, while maintaining an E_{PAUC} of 13.37. Because P_{notify} penalizes redundant emissions via bipartite matching, this increase confirms that event notification is delivered more efficiently with strictly minimized interruption costs. Together, these dynamics yield the highest MII-Score (31.68), outperforming both fixed-weight GRPO (30.54) and SFT (30.05). Thus, GRPO strengthens temporal alignment, reward annealing optimally balances proactive inquiry quality against notification precision.

Case Study. Figure 5 illustrates how SAIWen provides zero-trigger assistance from a continuous visual stream without requiring an explicit user prompt. Our model recognizes visual cues consistent with making vegan mushroom and onion gravy, spontaneously emitting an `inquiry` to ask user intent. Upon receiving an affirmative reply, SAIWen transitions from `inquiry` to functional assistance by sequentially emitting `response` and `notify` tokens to deliver step-by-step guidance.

6 Conclusion

In this work, we first propose the paradigm of proactively initiated interaction in streaming video understanding, shifting the model capabilities from reactive response to autonomous intervention. The proposed **SAIWen** model autonomously determines *whether*, *when*, and *how* to intervene by decomposing actions into four functional tokens (`inquiry`, `notify`, `response`, and `silence`). Trained on the SAIWen-Interact training split, our framework employs HCB Loss to resolve extreme token imbalance during SFT, while Curriculum Advantage Annealing in RL maximizes temporal alignment against the MII-Score. Experiments on the test set demonstrate that SAIWen improves intervention precision by 30.32% and notification timeliness by 14.90%, while preserving natural silence. Ultimately, this blueprint advances next-generation streaming video interaction.

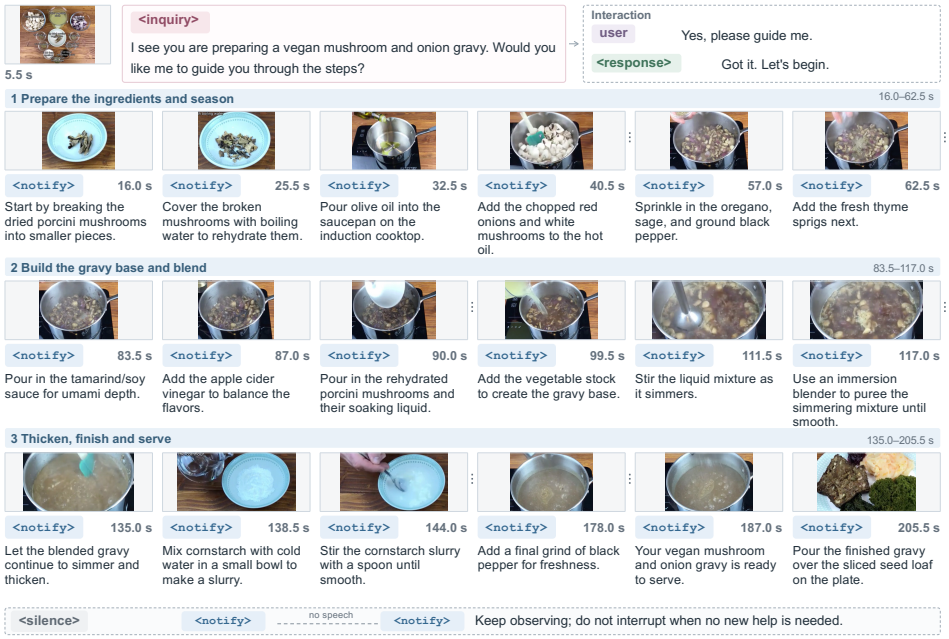


Figure 5: **Example of model-initiated interaction in mushroom and onion gravy preparation.**

AI USE STATEMENT

In this work, we used generative AI tools for coding assistance and text polishing. We have not used generative AI tools for drafting the manuscript, experimental design, or theoretical derivations. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

The dataset utilized in this study is primarily curated from publicly available open-source datasets. Furthermore, the interaction scenarios are strictly confined to routine daily assistance. Consequently, this work does not raise any ethical concerns.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we provide comprehensive details regarding our data sources, the data construction pipeline, model hyperparameters, and the specific prompts employed throughout the process. Upon acceptance, we will release additional reproducibility resources and further refine our materials based on reviewer feedback.

REFERENCES

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. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18407–18418. IEEE, 2024.

Xin Ding, Hao Wu, Yifan Yang, Shiqi Jiang, Qianxi Zhang, Donglin Bai, Zhibo Chen, and Ting Cao. Streammind: Unlocking full frame rate streaming video dialogue through event-gated cognition. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13448–13459. IEEE, 2025.

Shenghao Fu, Qize Yang, Yuan-Ming Li, Yi-Xing Peng, Kun-Yu Lin, Xihan Wei, Jian-Fang Hu, Xiaohua Xie, and Wei-Shi Zheng. Vispeak: Visual instruction feedback in streaming videos. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21778–21788. IEEE, 2025.

Zhenpeng Huang, Xinhao Li, Jiaqi Li, Jing Wang, Xiangyu Zeng, Cheng Liang, Tao Wu, Xi Chen, Liang Li, and Limin Wang. Online video understanding: Ovbench and videochat-online. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3328–3338. IEEE, 2025.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Raisa Islam and Owana Marzia Moushi. Gpt-4o: The cutting-edge advancement in multimodal llm. In *Intelligent Computing-Proceedings of the Computing Conference*, pp. 47–60. Springer, 2025.

Ao Li, Zihan Xiao, Zihao Yue, Boshen Xu, Linli Yao, Jiaze Li, Pei Fu, Jianzhong Ju, Jian Luan, and Qin Jin. Streampro: From reactive perception to proactive decision-making in streaming video. *arXiv preprint arXiv:2605.16381*, 2026.

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. In *2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 12147–12151, Barcelona, Spain, May 2026. IEEE.

Junbo Niu, Yifei Li, Ziyang Miao, Chunjiang Ge, Yuanhang Zhou, Qihao He, Xiaoyi Dong, Haodong Duan, Shuangrui Ding, Rui Qian, et al. Ovo-bench: How far is your video-llms from real-world online video understanding? In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18902–18913. IEEE, 2025.

Rui Qian, Shuangrui Ding, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Dahua Lin, and Jiaqi Wang. Dispider: Enabling video llms with active real-time interaction via disentangled perception, decision, and reaction. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24045–24055. IEEE, 2025.

Bytedance Seed. Seed2. 0 model card: Towards intelligence frontier for real-world complexity. *arXiv preprint arXiv:2607.00248*, 2026.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

Qwen Team. Qwen3. 5-omni technical report. *arXiv preprint arXiv:2604.15804*, 2026.

Yueqian Wang, Xiaojun Meng, Yifan Wang, Huishuai Zhang, and Dongyan Zhao. Proactivevideoqa: A comprehensive benchmark evaluating proactive interactions in video large language models. *arXiv preprint arXiv:2507.09313*, 2025.

Jiaer Xia, Peixian Chen, Mengdan Zhang, Xing Sun, and Kaiyang Zhou. Streaming video instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 31219–31229, 2026.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-omni technical report. *arXiv preprint arXiv:2503.20215*, 2025.

Xiaojun Xue, Chunxia Zhang, Zhendong Niu, and Xindong Wu. Multi-level attention map network for multimodal sentiment analysis. *IEEE Transactions on Knowledge and Data Engineering*, 35 (5):5105–5118, 2022.

Dingyu Yao, Junhao Zhou, Chenxu Yang, Chuanyu Qin, Haowen Hou, Zheming Liang, Congcong Wang, Yuhang Cao, Shenglong Ye, Shuai Xie, et al. Joyai-vl-interaction: Real-time vision-language interaction intelligence. *arXiv preprint arXiv:2606.14777*, 2026.

Xiangyu Zeng, Kefan Qiu, Qingyu Zhang, Xinhao Li, Jing Wang, Jiaxin Li, Ziang Yan, Kun Tian, Meng Tian, Xinhai Zhao, et al. Streamforest: Efficient online video understanding with persistent event memory. *Advances in Neural Information Processing Systems*, 38:75804–75835, 2026.

Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. In *Proceedings of the 2023 conference on empirical methods in natural language processing: system demonstrations*, pp. 543–553, 2023.

A Appendix

A.1 Prompt Design

Dialogue Generation (Assistant)

You are an expert at designing IDEAL proactive AI assistant dialogues for first−person (egocentric) video . Goal: produce a multi−turn conversation that a smart assistant would naturally have with the user in real time, time−aligned with the video. Output ONLY valid JSON.

The three semantic tags −− every assistant turn MUST start with exactly one
− [Notify] = the assistant SPONTANEOUSLY tells the user something. Declarative.
− [Inquiry] = the assistant ASKS the user something. ANY question MUST be [Inquiry].
− [Response] = the assistant REPLIES to a preceding user turn. Declarative.

Evidence threshold before speaking Speak only when the event is specific, visually grounded, and helpful. Do NOT invent goals, risks, or next steps unsupported by the event list or caption. If intent is uncertain, prefer a cautious [Inquiry] over a confident [Notify].

Opening timing
The assistant watches in real time and cannot see the future. It stays SILENT until the user's overall activity is visually clear. The opening [Inquiry] MUST NOT be earlier than ' earliest_allowed_opening_time'; events before it are background only. Opening at t=0 is almost always wrong.

Required opening pattern
The FIRST task MUST open with a cautious [Inquiry] (grounded in the first recognizable activity), then a user reply, then a [Response] with a short plan.

Dialogue Generation (USER)

You are given an egocentric video summary and a Stage−2 event list. Convert them into a natural, time−aligned proactive assistant conversation.

Video Info
− video_id: {video_id}
− duration: {duration:.1f}s
− caption: {caption}

Overall Activity Confirmation (teacher−distilled)
{inquiry_time_text}

Stage−2 Events
{events_text}

Output JSON Schema
{ "tasks": [{ "task_id": <int>, "task_desc": "<3−8 words>",
"turns": [{ "role": "assistant"|"user", "content": "<text>",
"time": <float>, "start_time": <float>, "end_time": <float> }] }] }

Key Rules
1. Cover important events with assistant turns anchored near their timestamps, only when concrete evidence makes speaking useful.
2. FIRST task opening: [Inquiry] −> user reply −> [Response]; the [Inquiry] time MUST be >= earliest allowed opening time.
3. ANY assistant question MUST be tagged [Inquiry]; [Response] MUST directly follow a role=user turn; never place two user turns consecutively.
4. Preserve numbers, tools, object names, safety and technique details.
5. Keep each message to 1−3 sentences; times strictly increasing within a task.

Now output STRICT JSON only.

Inquiry-Time Grounding

You are a precise temporal−grounding teacher for egocentric video. Output ONLY valid JSON. Watch the video and identify the earliest reliable timestamp at which the user's overall activity can be inferred from visual evidence. Do NOT generate dialogue or Inquiry/Notify/Response labels.

Prefer the first moment where a specific action/tool/object combination makes the activity clear; do not use the first visible object or generic hand motion. The timestamp answers: "From this point onward, the user is doing <activity>."

Output JSON Schema
{ "video_id": "{video_id}",
"overall_activity": "<short phrase>",
"activity_confirmed_time": <float seconds or null>,
"confidence": "high"|"medium"|"low",
"status": "confirmed"|"ambiguous"|"no_clear_activity",
"evidence": [{ "time": <float>, "description": "<visible evidence>" }],
"reason": "<why this is the earliest reliable confirmation point>" }

Reward – Query

You are judging whether a proactive assistant's inquiry is reasonable for the activity shown in the reference inquiry. Judge semantic content only; ignore timing. Score the MODEL INQUIRY on 0−4:
4 = identifies the specific activity and asks for fully appropriate help.
3 = correct action/domain and key object or goal; useful even if a detail differs.
2 = only the correct setting or a very broad activity; limited but not misleading.
1 = weak neighbouring context with almost no actionable match.
0 = wrong activity, contradictory, unrelated, or not an inquiry.
Return JSON only: {"score":0,"reason":"short reason"}.

REFERENCE INQUIRY:
{gold}

MODEL INQUIRY:
{pred}

Reward – Response

You are strictly auditing a proactive assistant after a user reply. The GT inquiry and GT response define the true activity. The simulated user may incorrectly accept a hallucinated inquiry; never use that to legitimise a wrong activity. Score MODEL RESPONSE on 0−2:
2 = same underlying activity and an equally valid, specific response.
1 = same activity and broadly useful, but generic, partial, or missing a key instruction.
0 = wrong activity, unrelated, materially incorrect, unsafe, or missing.
Set authorization_violation=true if the user refuses/corrects but the model continues the rejected task.
Return JSON only: {"score":0,"authorization_violation":false,"reason":"short reason"}.

GT INQUIRY: {gt_inquiry}
GT USER: {gt_user}
GT RESPONSE: {gt_response}
MODEL INQUIRY: {model_inquiry}
SIMULATED USER: {sim_user}
MODEL RESPONSE: {model_response}

Reward – Notify

You are judging one proactive notification against one ground−truth event. The attached image is the video frame at the event's exact timestamp. Judge using BOTH the ground−truth text and this image; do not invent unsupported details. Score 0−2:

2 = fully conveys the same event or an equally correct, useful instruction.
1 = partially conveys it: right object/action but incomplete or partly off.
0 = unrelated, contradictory, or materially wrong.
Return JSON only: {"score":0,"reason":"short reason"}.

GROUND TRUTH EVENT:
{gold}

MODEL NOTIFICATION:
{pred}

User Simulator – Caption-grounded reply

You are the human user wearing a head−mounted camera. The full−video caption below is authoritative factual background: {factual_background} The concise activity label is: "{scene}". An AI assistant is watching your video and just spoke to you. Reply briefly as the user: ONE short natural sentence (<=20 words), no tags, no prefixes.

Policy (strict_scene): Treat the stated activity as ground truth. If the activity and requested help are correct , accept naturally. If a detail is wrong, correct it and then accept. If the assistant describes a different activity, explicitly reject the false premise and state the true activity.

Do NOT pretend to see your own hands; speak as the user wearing the camera.

Eval – Stage-C proactive assistant

You are a proactive real−time visual assistant. You continuously watch a first−person video stream frame by frame. At every frame you must decide whether to stay silent and keep watching, or to proactively speak up (notify, inquire, or respond) when it is genuinely helpful. Only speak when the visual context warrants it; otherwise remain silent.

A.2 SAIWEN-INTERACT DATASET DETAILS

Analysis of Proactive-Opening Density. To comprehensively characterize the distribution of proactive interactions within the training split of SAIWen-Interact, we systematically analyze the density of proactive openings (defined as the sum of *inquiry* and *notify* tokens). As summarized in Table 7, the corpus comprises 26,549 annotated videos (662.7 hours in total), yielding 179,053 proactive openings. We observe a natural discrepancy between the per-video mean opening rate (micro-average, 8.87/min) and the dataset-wide macro rate (4.50/min). To provide a robust representation of typical single-video behavior that is less sensitive to length variations, we also report the median rate of 7.82/min.

This difference between micro and macro averages primarily reflects the systematic variation of interaction density across different video lengths, as detailed in Table 8. Shorter clips ($<$ 30s), which constitute a substantial portion of the corpus ($\sim$42%), naturally exhibit a higher concentration of interaction nodes, resulting in a mean density of 14.15 openings/min. Conversely, long-form videos ($\geq$ 300s) feature a more gradual and sparse interaction rhythm, with the density adjusting to 1.88 openings/min. This inverse correlation between duration and opening density is an inherent characteristic of diverse video content, where shorter segments often capture focused, event-driven moments.

Recognizing this duration-dependent density variation is crucial for stable model training. If left unaddressed, the model might inadvertently bias its turn-initiation policy toward the high-frequency patterns of short clips or the prolonged silence of longer videos. This motivates our proposed multi-view temporal windowing and the class-balanced focal objective. By explicitly normalizing the learning signal across different temporal scales, our framework mitigates the length-induced distribution shift, ensuring the model learns robust, context-aware turn-initiation behaviors that generalize well across videos of varying durations.

Table 7: Summary of proactive-opening density in the training split of SAIWen-Interact.

Statistic	Value
#Videos	26,549
Total duration (h)	662.7
Total proactive openings (Notify+Inquiry)	179,053
Mean openings per video	6.74
Openings/min, per-video mean (micro)	8.87
Openings/min, per-video std	5.76
Openings/min, dataset macro	4.50
Openings/min, median (p50)	7.82
Openings/min, p10 / p90	2.51 / 17.04
Openings/min, min / max	0.00 / 35.84

Table 8: Proactive-opening density by video-duration bucket in the training split of SAIWen-Interact.

Duration (s)	#Videos	Mean openings/min	Macro openings/min
< 30	11,197	14.15	13.30
30–60	4,055	8.56	8.37
60–120	2,599	5.31	5.15
120–300	8,360	3.35	3.20
≥ 300	338	1.88	1.80

Summary of SAIWen-Interact Test Set. Figure 6 summarises the functional composition of the 2,000 videos in the SAIWen-Interact test set along the four function actions defined in Sec. 3. The labels yield 4,810 behaviour annotations in total: Progress Monitoring constitutes the largest share (49.77%), followed by Technical Guidance (27.52%), Safety Alerting (19.86%), and Reminder (2.85%). Figure 7 details the domain composition across the same 20 scene domains as those in SAIWen-Interact's training split. Cooking is the largest domain (39.05%), followed by Electronics (13.80%) and Mechanical Repair (9.30%), while the remaining seventeen domains each account for at most 5.95%.

A.3 Detailed Results

Tables 9, 10, 11, 12 detail the fine-grained performance of SAIWen on conventional streaming evaluation benchmarks. Specifically, SAIWen exhibits suboptimal performance in object grounding. This limitation stems primarily from our current focus on the turn-initiation training paradigm, which necessitates training predominantly on question-answering datasets. Future work will augment the training corpus with grounding-specific data to further enhance the model's proficiency in this area.

Inference efficiency. A proactive assistant must run continuously alongside an open-ended video stream, so its dominant cost is not the occasional response but the watch-and-decide loop executed at every frame, whether or not the model ultimately speaks. We therefore isolate this per-frame decision cost under an identical Qwen3.5-4B backbone and an identical incremental KV-cache streaming frontend (Table 13). A natural training-free approach is to prompt the model at every frame and let it autoregressively verbalize whether to act; since the backbone dedicates no symbol to this choice, each frame costs $\sim$8 forward passes (2,213 per 278-frame clip) merely to produce a decision that is almost always "remain silent." Our model is instead trained to express the per-frame choice as one of four dedicated decision tokens (<|*silence*|>, <|*notify*|>, <|*inquiry*|>, <|*response*|>). Each new frame block is consumed by a single incremental forward, and the decision is read directly from that forward's logits via an argmax over the decision tokens, incurring no additional forward, while the full generator is invoked only on the rare frames where a speak token is emitted. Always-on monitoring therefore costs exactly one forward per frame (277 per clip), an $\sim 8\times$ reduction. Because this

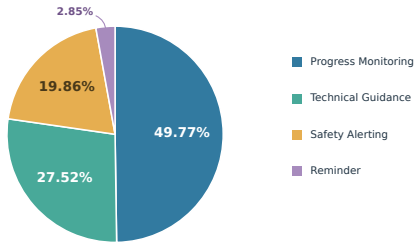


Figure 6: Distribution of assistance categories in the SAIWen-Interact test set.

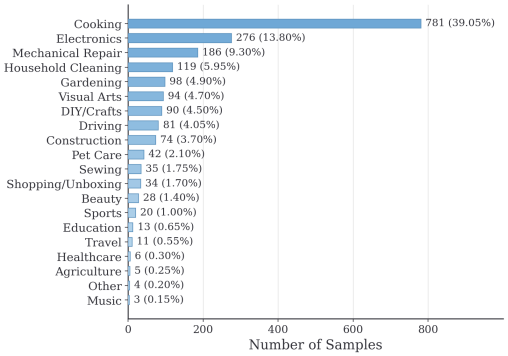


Figure 7: Distribution of scene domains in the SAIWen-Interact test set.

Table 9: Per-task results on OVO-Bench. **Real-Time Visual Perception**: OCR (Optical Character Recognition), ACR (Action Recognition), ATR (Attribute Recognition), STU (Spatial Understanding), FPD (Future Prediction), OJR (Object Recognition); **Backward Tracing**: EPM (Episodic Memory), ASI (Action Sequence Identification), HLD (Hallucination Detection); **Forward Active Responding**: REC (Repetition Event Count), SSR (Sequential Steps Recognition), CRR (Clues Reveal Responding). All entries use the strict unique-option scoring protocol; AVG is the official group-macro score. **Bold** denotes the best in each column.

Model	Params	Real-Time Visual Perception						Backward Tracing			Forward Active Responding			AVG
		OCR	ACR	ATR	STU	FPD	OJR	EPM	ASI	HLD	REC	SSR	CRR	
Qwen3.5	4B	47.0	33.9	40.5	33.7	32.7	33.7	33.3	37.8	26.9	29.9	63.4	42.5	38.3*
Qwen2.5-Omni	7B	**82.6**	65.1	**82.8**	57.9	**77.2**	**72.8**	62.6	61.5	18.8	25.2	48.5	57.5	54.8
Qwen3-Omni	30B	81.2	**66.1**	81.0	**61.2**	75.2	70.7	**64.6**	**66.2**	41.4	33.7	68.0	62.5	**61.4**
VideoLLM-online	8B	8.1	23.9	12.1	14.0	45.5	21.2	22.2	18.8	12.2	–	–	–	12.8*
Dispider	7B	57.0	48.6	62.1	44.9	61.4	51.6	49.5	54.7	4.3	25.2	53.7	52.1	44.7
VideoChat-Online	4B	62.4	58.7	56.9	47.8	48.5	53.8	36.7	50.0	10.2	21.6	65.2	47.9	44.0
StreamForest	7B	71.1	55.0	71.6	51.7	67.3	63.0	57.6	66.2	29.6	23.2	47.5	**64.2**	53.1*
JoyAI-VL-Interaction	8B	76.5	56.0	79.3	59.6	72.3	66.8	61.6	64.9	28.5	34.4	63.0	34.6	54.7
SAIWen (Ours)	4B	69.8	51.4	57.8	44.9	63.4	47.3	52.2	61.5	**44.1**	**34.8**	**68.2**	45.0	52.6

saving falls on the code path that runs at every frame rather than on the infrequent responses, it compounds with stream length and is precisely what makes always-on proactive assistance affordable in practice, independent of how much the model eventually says.

A.4 Evlation of SAIWen

Baseline Evaluation Protocol. None of the baselines was trained for our streaming decision head, so we equip each of them with a prompt-level surrogate and evaluate all systems on the SAIWen-Interact test set using an identical decision grid. For every clip we reuse the frame timestamps $\{t_1, \ldots, t_M\}$ of the corresponding stream frame that our model consumes (with no temporal truncation) and let every system decide whether to speak at each t_i; each frame is decoded at its nearest position and resized to 384×384. All systems thus share the same speaking opportunities at the same time points, which makes the per-category figures directly comparable. Moreover, SAIWen is run through its native streaming path with a rolling context of up to 200 frames, whereas the prompt-driven baselines re-encode only a short frame window at each step.

Traditional Models with a Proactive Prompt. Offline video-language models inherently generate responses only upon explicit prompting, lacking an intrinsic mechanism for proactive silence. To adapt them into proactive agents, we invoke them at every frame under a constrained *action-digit* protocol. At step i, the model receives: (i) a system prompt defining four admissible actions—0 (stay silent), 1 (ask a clarifying question), 2 (emit a proactive update or instruction), and 3 (answer the user's latest reply); (ii) a sliding visual context comprising the initial frame of the clip and the C most recent frames (i.e., the current frame and its $C-1$ predecessors, totaling at most $C+1$ frames, where $C = 4$); (iii) a textual timestamp formatted as *"Time: t_i seconds"*; and (iv) the preceding $H = 8$ dialogue turns. The model is constrained to output either the single token 0

Table 10: Per-subtask results on ODVBench. **Single Static**: RTP (Real-time Traffic Perception), PTM (Past Traffic Memory), TCD (Traffic Change Detection), KIE (Key Information Extraction), DDM (Driving Decision-Making), HD (Hallucination Detection); **Single Dynamic**: LP (Location Prediction), DP (Distance Prediction), AP (Action Prediction); **Multi-Object**: RP (Risk Prediction), RA (Risk Analysis), ARA (Accident Reason Answering). AVG is the official (micro) score. **Bold** denotes the best in each column.

Model	Params	Targeting Single Static Objects						Targeting Single Dynamic Objects			Targeting Multi-Object Interaction			AVG
		RTP	PTM	TCD	KIE	DDM	HD	LP	DP	AP	RP	RA	ARA	
Qwen3.5	4B	29.9	30.7	54.5	24.5	44.6	31.7	25.1	50.2	33.0	36.3	24.0	29.5	35.7*
Qwen2.5-Omni	7B	65.2	62.9	72.7	39.6	34.3	6.5	91.6	46.8	50.9	46.2	74.9	47.4	57.5
Qwen3-Omni	30B	74.0	72.0	**81.8**	64.2	**51.2**	22.8	82.7	48.3	50.9	33.0	73.4	**57.7**	57.8
Dispider	7B	31.5	36.4	43.6	34.0	35.6	2.4	74.9	46.0	53.8	33.4	56.8	36.5	46.0
VideoChat-Online	4B	33.7	43.2	47.3	58.5	20.8	0.0	88.4	46.8	71.5	50.5	64.3	46.8	53.7
StreamForest	7B	51.8	65.7	56.4	54.7	38.6	17.1	83.4	46.3	72.6	60.3	73.8	47.4	60.2*
JoyAI-VL-Interaction	8B	**79.8**	**85.2**	78.2	**86.8**	46.5	**44.7**	**92.0**	**54.0**	**81.8**	19.8	**88.6**	56.4	**71.2**
SAIWen (Ours)	4B	53.3	52.8	69.1	39.6	45.9	40.7	70.5	47.2	46.5	**62.2**	49.9	35.9	53.3

Table 11: Full per-subtask results on OVBench. Task groups: FP (Future Prediction), THV (Temporal Hallucination Verification), PM (Past Memory), SP (Spatial Perception), STP (Spatio-Temporal Perception), TP (Temporal Perception). Subtasks: AA=Action Anticipation, GSP=Goal/Step Prediction, MP=Movement Prediction, AP=Action Persistence, SV=Step Verification, OP[a]=Object Presence, AR=Action Retrieval, PR=Procedure Recall, TR=Trajectory Retrieval, AL=Action Location, OP[b]=Object Position, AT=Action Trajectory, OT=Object Trajectory, AS=Action Sequence, SL=Step Localization, OES=Object Existence State. **Bold** marks the best in each column.

		FP			THV			PM			SP		STP		TP			
Model	**Params**	AA	GSP	MP	AP	SV	OP[a]	AR	PR	TR	AL	OP[b]	AT	OT	AS	SL	OES	**AVG**
Offline Models																		
Qwen3.5	4B	24.4	32.7	25.5	50.6	56.3	43.7	26.3	39.2	30.1	24.4	27.6	34.3	21.7	27.9	27.1	23.1	35.1
Qwen2.5-Omni	7B	44.9	**70.8**	27.9	49.4	**61.1**	58.2	27.2	69.3	73.0	33.8	13.9	87.0	44.1	64.7	25.2	8.9	49.8
Qwen3-Omni-30B-A3B-Instruct	30B	51.3	67.7	**36.5**	53.6	55.9	70.2	28.4	**73.3**	79.1	30.9	43.3	73.9	47.2	**72.1**	27.7	57.5	54.5
Streaming Models																		
VideoLLM-online	8B	0.0	1.8	20.9	5.2	5.9	32.6	0.0	2.3	26.7	0.6	26.6	0.9	19.9	0.9	1.7	8.3	9.6
Dispider	7B	38.5	53.8	23.6	42.9	45.8	61.2	29.1	55.4	59.5	31.3	19.5	67.4	48.1	57.7	25.8	4.6	42.2
VideoChat-Online	4B	60.3	57.7	18.7	59.7	58.2	66.8	40.0	56.8	74.2	45.3	37.8	86.5	51.6	71.6	25.5	72.6	55.2
StreamForest	7B	**69.2**	60.0	34.4	**69.1**	54.0	**72.9**	**50.9**	64.9	82.2	**56.6**	**87.9**	95.2	**61.2**	64.2	30.6	**92.6**	**60.5**
JoyAI-VL-Interaction	8B	55.1	56.7	25.5	48.5	54.7	72.6	40.3	72.4	**85.0**	40.0	83.3	**97.8**	45.7	69.3	**31.6**	85.2	59.5
SAIWen (Ours)	4B	48.7	57.4	28.5	37.3	56.2	60.0	34.1	64.4	63.2	25.9	35.3	65.7	42.2	68.8	28.4	35.4	48.2

[a] Object Presence (THV). [b] Object Position (SP). † Evaluated under our harness (official micro-accuracy AVG). ‡ Per-subtask numbers from the original paper. § Native official-repository run; AVG is the macro average over the 16 subtasks (micro = 52.4).

or an action digit followed by a single concise sentence (e.g., 2: The pan is starting to smoke, lower the heat.). Text is generated via sampling at a temperature of 1.0 with a maximum of 64 new tokens. A lightweight state machine strictly enforces protocol validity: an *inquiry* may be triggered at most once per clip; a Response is admissible only when a user reply is pending; and no unsolicited *notify* is issued while awaiting user input. Whenever the agent emits an *inquiry*, the subsequent user turn is generated by an LLM-based user simulator. This simulator is conditioned on the clip's reference activity description and the current frame, producing a single first-person sentence. Qwen2.5-Omni (Xu et al., 2025) is evaluated under this exact configuration. Conversely, Qwen3.5 is evaluated via its native streaming decision path using the off-the-shelf base checkpoint and a natural-language decision prompt. This ensures that the performance gap between it and our model isolates the effect of our proposed training paradigm rather than differences in the inference interface.

Scoring. All systems are scored with the same protocol as our model (tolerance window $\delta_{\text{pre}} = 2\,\text{s}$, $\delta_{\text{post}} = 8\,\text{s}$, three judge votes per item). The reported MII-Score is the average of the Query and Notify-Utility directions, where Query is a timing-agnostic $0-4$ judge rating of the model's first inquiry (normalised to $[0, 1]$) and Notify Utility is the harmonic mean of the event PAUC and the one-to-one notification precision. Because the visual-context length C trades accuracy against a prohibitive per-frame inference cost, we fix $C = 4$ for both prompt-driven baselines; this is a deliberate protocol choice, applied identically to every non-native baseline to preserve comparability.

Table 12: Per-category results on the Real-Time Visual Understanding split of StreamingBench. OP (Object Perception), ACP (Action Perception), ATP (Attribute Perception), SU (Spatial Understanding), TRU (Text-Rich Understanding), EU (Event Understanding), CT (Counting), CS (Clips Summarize), CR (Causal Reasoning), PR (Prospective Reasoning). **Bold** denotes the best in each column.

Model	Params	OP	ACP	ATP	SU	TRU	EU	CT	CS	CR	PR	AVG
Gemini 1.5 Pro	–	–	–	–	–	–	–	–	–	–	–	75.7
GPT-4o	–	–	–	–	–	–	–	–	–	–	–	73.3
Qwen3.5	4B	53.1	54.3	56.4	54.5	66.7	55.3	38.8	50.8	43.8	53.7	53.9*
Qwen2.5-Omni	7B	84.6	77.8	**87.2**	**74.8**	86.6	77.4	**60.1**	87.4	80.5	**84.3**	81.1
Qwen3-Omni	30B	**86.2**	**80.4**	86.2	70.7	**88.5**	78.6	56.4	**89.3**	81.2	82.4	**81.5**
VideoLLM-online	8B	39.1	42.5	31.1	34.2	32.4	46.0	27.9	34.5	40.1	31.5	36.0
Dispider	7B	75.9	72.7	74.0	61.8	65.7	72.3	33.5	80.8	**84.4**	69.4	69.9
VideoChat-Online	4B	71.3	63.4	78.2	65.9	62.0	71.7	44.1	73.5	59.4	66.7	66.8
StreamForest	7B	83.2	75.6	83.8	69.1	79.8	**79.0**	56.4	82.3	82.8	78.7	77.8*
JoyAI-VL-Interaction	8B	81.6	77.0	84.3	71.5	83.8	74.2	55.9	82.0	81.2	79.6	79.6
SAIWen (Ours)	4B	84.0	74.7	83.5	68.7	85.0	**79.0**	47.3	82.0	82.0	63.0	76.9

Table 13: **Per-frame monitoring cost of the inference mechanism.**

Inference mechanism	Per-frame decision	Fwd/frame ↓	Fwd/clip ↓
Per-frame prompt (training-free)	verbalized (generated)	7.95	2,213
Ours (SFT+RL, decision token)	single token (argmax)	**1.00**	**277**

A.5 Case Studies

Human Evaluation. To further validate the effectiveness of SAIWen, we conducted a manual evaluation on 50 instances sampled from the 2,000 videos in the SAIWen-Interact test set. Specifically, we manually assigned evaluation scores to both the *Inquiry* and the *notify* for each instance. As illustrated in Figure 8, SAIWen successfully detects the majority of opening events and demonstrates robust performance on specific scenarios, such as "cooking". However, certain misclassifications persist; for instance, the presence of a bicycle in a laundromat scene is erroneously interpreted as a user intent to wash the bicycle. In terms of event tracking, while SAIWen generates task-relevant notifications in most samples, the excessive frequency of these alerts adversely still impacts the overall evaluation score. Consequently, future work will focus on enhancing the precision of notification generation to mitigate user disruption and avoid excessive interruptions.

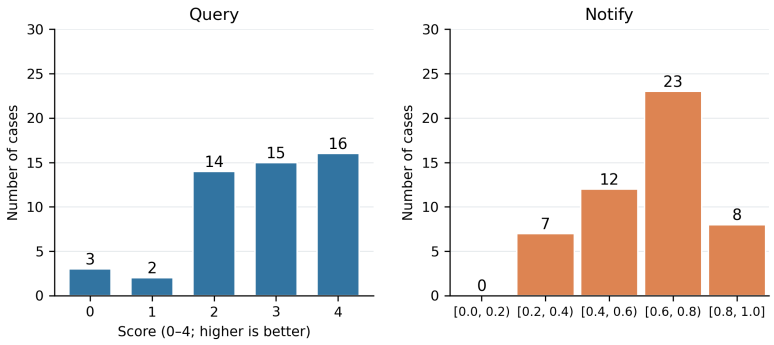


Figure 8: The illustration of human evulation.

Case Study. Figures 9 and 10 further illustrate the behavior of our model. As demonstrated in the case study, our model effectively initiates an *inquiry* at the onset of the video and provides continuous assistance upon receiving an affirmative response from the user. Notably, in Figure 8, when the user re-initiates a search operation for the second time, the model automatically recognizes

the new event intention and performs an *inquiry*. This indicates that, through training, the model has spontaneously acquired the capability to conduct *inquiry*, which fully validates the effectiveness of our proposed paradigm.

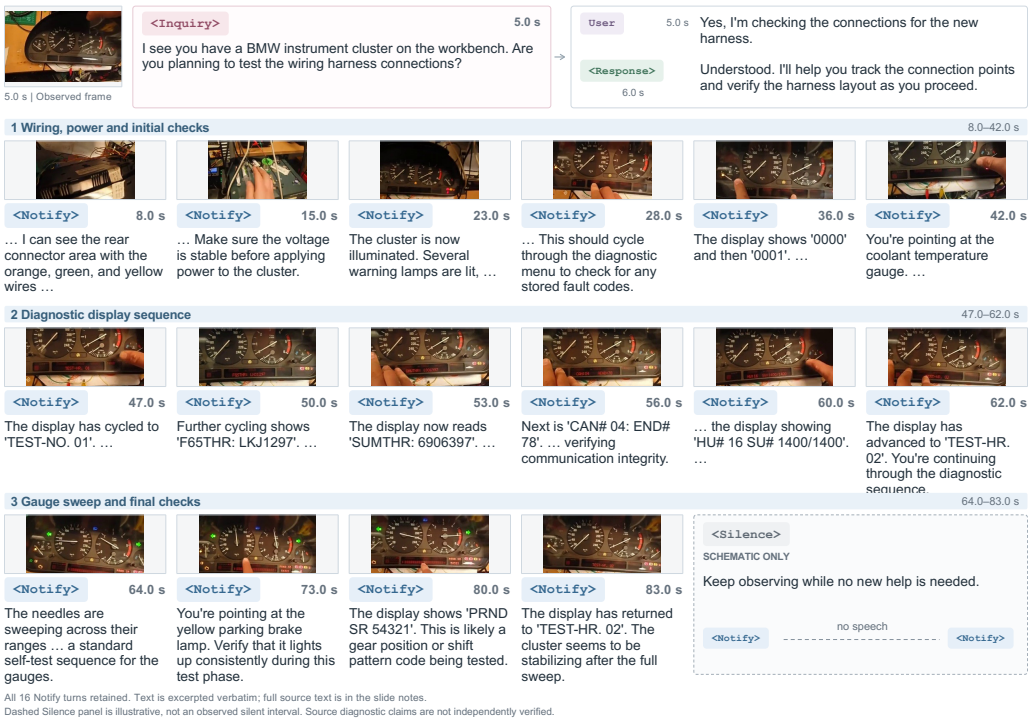


Figure 9: Case of a BMW instrument cluster on a workbench prepared for wiring harness connection testing.

A.6 Limitations and Future Work

Our current work focuses on autonomous, proactive initiation independent of explicit user queries, with empirical validation strictly confined to the streaming video understanding domain. However, the zero-trigger interaction paradigm fundamentally transcends this specific scope. Model-initiated interaction encompasses a broad spectrum of complex tasks and multimodal data distributions, including intricate spatial-grounding interactions and omni-modal active interventions augmented by continuous audio streams. Extending autonomous initiation to these scenarios presents substantial research and practical value utility. In future work, we plan to broaden our research scope by formalizing diverse downstream tasks and interaction types. Furthermore, we will explore the initiation of proactive interactions based on omni-modal data.

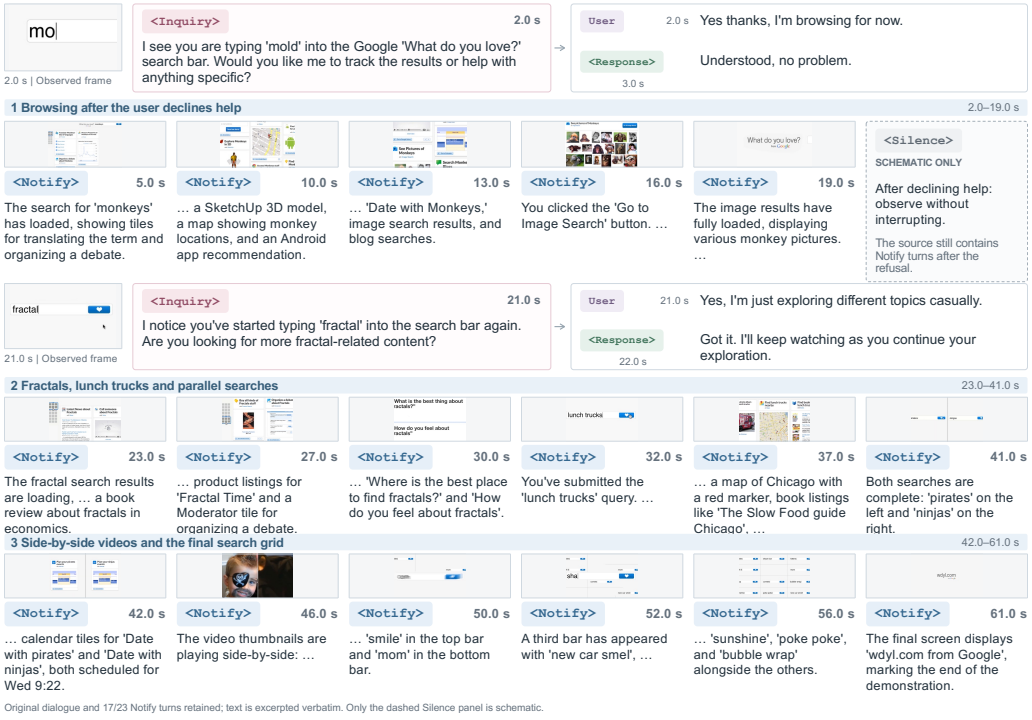


Figure 10: Case of proactive assistance offered to a user actively searching for “mold” on a Google interface.