BLV-GATE: BENCHMARKING PROACTIVE COMMUNICATION IN BLIND AND LOW-VISION ASSISTANCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Continuous visual assistance for blind and low-vision (BLV) users requires an assistant to decide whether to communicate, when to respond, and what information is useful, while avoiding unnecessary output. Existing evaluations focus on isolated questions or predefined event recall, providing limited evidence about this sequence-level communication problem. We introduce BLV-GATE, a benchmark built from 499 verified egocentric video sessions across five everyday scene categories, containing 5,863 temporally localized communication events with grounded visual facts and reference responses. BLV-GATE evaluates complete causal response sequences across intervention regulation, timing and latency, grounding and response quality, and session-level information burden. We also introduce SenseBridge Agent, a companion baseline with causal visual reasoning and memory. Experiments with ten representative baselines reveal clear trade-offs among event coverage, communication burden, response quality, and latency: periodic systems achieve higher recall at the cost of frequent output, conservative streaming systems miss necessary events, proactive assistants show stronger potential, and stronger backbones improve grounding and response quality without resolving timely delivery.[1]

1 INTRODUCTION

Large vision-language models (LVLMs) have substantially improved AI systems' ability to perceive visual content and communicate it through natural language (Liu et al., 2023; Wang et al., 2024a). Building on this progress, video-language models extend visual understanding from static images to dynamic scenes by modeling activities, environmental changes, and temporal context (Maaz et al., 2024; Lin et al., 2024a). More recently, streaming video-language models have enabled systems to process visual inputs causally as they arrive, supporting online question answering and continuous commentary (Chen et al., 2024a; Wang et al., 2025c; Lin et al., 2024b). Going beyond continuous perception, proactive video-language models explicitly decide whether to respond or remain silent until sufficient evidence is available (Video Understanding Team of JoyAI-VL, 2026; Wang et al., 2025b). Together, these advances make it increasingly feasible to build systems that continuously observe and communicate about the visual world.

This capability is particularly promising for blind and low-vision (BLV) users, who may benefit from assistance in understanding their surroundings, identifying objects and signs, recognizing potential hazards, and accessing contextual information during everyday activities (Gurari et al., 2018; Massiceti et al., 2021; Xiao et al., 2025; Yuan et al., 2025). When paired with wearable cameras or smart glasses, video-language models could serve as always-on assistants that continuously observe the environment, answer user questions, and selectively communicate relevant changes. However, this setting introduces a challenge beyond visual understanding: the assistant must determine whether communication is useful, when it should occur, and when silence is preferable. Existing benchmarks mainly evaluate predefined tasks, annotated events, or individual responses (Lin et al., 2024b; Gao et al., 2026; Wang et al., 2025b; Xiao et al., 2025; Liu et al., 2026). As illustrated in Figure 1, such evaluations provide limited evidence about whether an assistant speaks unnecessarily or repeatedly, misses important information, or responds too late.

[1] All experimental data and code are available at: https://anonymous.4open.science/r/BLV-Gate-7A03.

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.

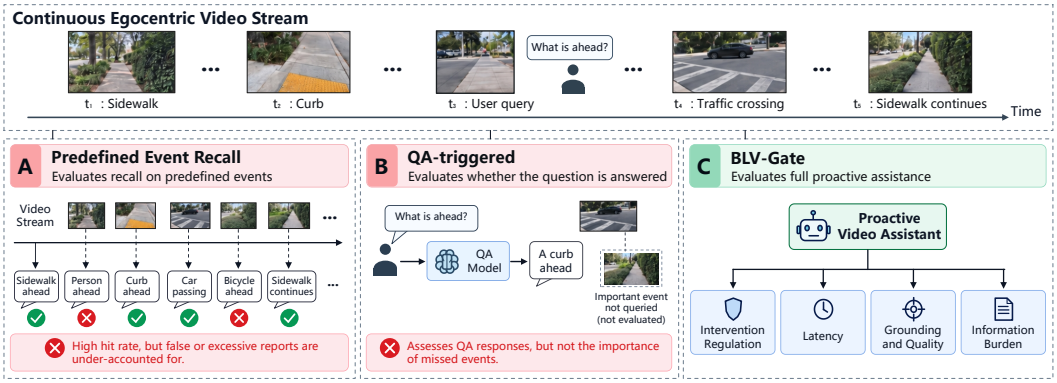


Figure 1: Existing evaluations assess predefined events or user-triggered questions in isolation. BLV-GATE evaluates complete causal response sequences for continuous BLV assistance across intervention regulation, latency, grounding and response quality, and information burden.

To bridge this gap, we introduce BLV-GATE, a benchmark for continuous visual assistance for blind and low-vision users. BLV-GATE is built from egocentric videos spanning five everyday settings: *Home*, *Shopping & Service*, *Transportation*, *Outdoor Walking*, and *Public Facilities*. It contains 499 verified sessions and 5,863 communication events, including `INFORM`, `WARN`, and `VQA`, with response windows, grounded visual facts, reference responses, and user queries. Unlike benchmarks centered on isolated questions or predefined clips, BLV-GATE evaluates complete response sequences under a causal protocol, capturing useful interventions and natural periods of silence.

BLV-GATE provides a sequence-level evaluation framework that separately measures intervention regulation, response timing and latency, visual grounding, response quality, and session-level communication burden. We also introduce SenseBridge Agent, a companion baseline combining causal visual reasoning with short-term visual memory and response history. Experiments across ten representative offline, streaming, and proactive video-language baselines reveal clear trade-offs. SenseBridge Agent with GPT-5.6-Luna achieves the highest intervention recall (81.5%) and F1 (0.390), yet produces 8.47 false interventions per minute and misses 99.8% of response deadlines due to high output latency. In contrast, SimpleStream attains higher precision (0.294) with lower communication volume, but recalls only 50.8% of annotated events. Rate-controlled experiments further show that increasing decision frequency generally improves recall at the cost of more false interventions, while the effect of input frame rate varies across architectures. Overall, event coverage, communication efficiency, response quality, and timely delivery remain difficult to optimize jointly.

In summary, the key contributions of this paper are threefold: (1) we formulate continuous BLV assistance as a communication-regulation problem that requires models to determine whether, when, and what to communicate; (2) we introduce BLV-GATE, together with a human-verified construction pipeline and a fine-grained evaluation framework for complete causal response sequences; and (3) we provide SenseBridge Agent and systematically evaluate existing video-language systems, revealing capability boundaries and trade-offs among event coverage, intervention precision, response usefulness, latency, and communication volume. We hope BLV-GATE can facilitate future research on reliable and efficient continuous visual assistance for blind and low-vision users.

2 Related Work

LVLMs. LVLMs connect pretrained vision encoders with large language models (LLMs) and use visual instruction tuning to support general vision-language reasoning (Radford et al., 2021; Li et al., 2022; Liu et al., 2023; Wang et al., 2024a), including visual question answering (Antol et al., 2015), image captioning (Cheng et al., 2025), and visual reasoning (Thawakar et al., 2025). Subsequent Video-Language Models, such as Video-ChatGPT (Maaz et al., 2024), Video-LLaMA (Zhang et al., 2023), and Video-LLaVA (Lin et al., 2024a), adapt these capabilities to temporally ordered visual inputs, while ShareGPT4Video (Chen et al., 2024b) provides video-instruction data for model adaptation. More specialized video-language models, including TimeChat (Ren et al., 2024), MovieChat (Song et al., 2024), and VideoAgent (Wang et al., 2024b), further introduce temporal reasoning, long-video memory, retrieval, or video-agent planning. In parallel, benchmarks such as MVBench (Li et al., 2024), Video-MME (Fu et al., 2025), and LongVideoBench (Wu et al., 2024)

Table 1: Comparison of related visual assistance benchmarks for BLV users.

Resource	Data Scope			Interaction		Evaluation Dimensions			
	BLV Focus	Ego. Video	Continuous Session	User Query	Proactive Intervention	Timing & Silence	Grounded Facts	Response Quality	Session Burden
VizWiz (Gurari et al., 2018)	✔	✘	✘	✔	✘	✘	✓	✓	✘
ObjectFinder (Liu et al., 2024)	✔	✓	✘	✔	✘	✘	✔	✓	✘
WalkVLM (Yuan et al., 2025)	✔	✔	✓	✓	✓	✓	✔	✓	✘
EgoBlind (Xiao et al., 2025)	✔	✔	✔	✓	✘	✘	✓	✘	✘
VisAssist (Gao et al., 2026)	✔	✔	✓	✔	✘	✓	✓	✘	✘
VIABench (Liu et al., 2026)	✔	✔	✓	✔	✓	✓	✔	✔	✘
BLV-GATE (Ours)	✔	✔	✔	✔	✔	✔	✔	✔	✔

evaluate video understanding from action recognition to long-context temporal reasoning. However, these models and benchmarks are mainly designed for offline or question-driven inference and do not require the model to regulate unsolicited interventions during continuous video observation.

Streaming and Proactive Video-Language Models. Recent work has moved beyond offline video understanding toward models that process causally available video streams. Within streaming video-language understanding, existing studies mainly consider two interaction settings. The first is streaming video question answering, where the model waits for sufficient evidence before answering, as explored by VideoLLM-online (Chen et al., 2024a), StreamingBench (Lin et al., 2024b), OVO-Bench (Niu et al., 2025), and MMDuet (Wang et al., 2025c). The second is continuous video commentary or online description, where the model generates temporally grounded descriptions with the video streaming (Yang et al., 2026; Wang et al., 2025c). Building on these causal streaming settings, proactive video-language models further allow the system to explicitly decide whether to respond, wait for more evidence, or remain silent, as in DisPider (Qian et al., 2025), ProactiveVideoQA (Wang et al., 2025b), MMDuet2 (Wang et al., 2025a), JoyAI-VL-Interaction (Video Understanding Team of JoyAI-VL, 2026), and Vinci2 (Gong et al., 2026). Although these systems target continuous and low-latency interaction, they are primarily developed for general-purpose video scenarios and do not account for the safety requirements, information needs, or interruption costs of blind and low-vision users.

Visual Assistance for Blind and Low-Vision Users. Visual assistance for blind and low-vision (BLV) users has evolved from user-captured image question answering to personalized object recognition, interactive guidance, and egocentric video assistance. Early resources such as VizWiz (Gurari et al., 2018) established the importance of user-captured images and naturally occurring visual questions, while ORBIT (Massiceti et al., 2021), GuideDog (Kim et al., 2026), and ObjectFinder (Liu et al., 2024) extended BLV assistance to personalized object recognition, accessibility-aware guidance, and interactive object search. EgoBlind and VisAssist (Xiao et al., 2025; Gao et al., 2026) further introduced first-person video question answering, and WalkVLM (Yuan et al., 2025) explored proactive navigation assistance. More recently, VIABench (Liu et al., 2026) unified proactive reminders, visual question answering, and vision-guided interaction in a video benchmark collected from or associated with visually impaired users. Despite these advances, existing methods and benchmarks remain largely organized around predefined tasks or individual responses. Their evaluations therefore emphasize task completion, event detection, or answer quality, while providing limited analysis of the complete interaction sequence.

3 THE BLV-GATE BENCHMARK

To evaluate continuous visual assistance for blind and low-vision users, we introduce BLV-GATE, a benchmark of complete egocentric video sessions. It evaluates when assistants should intervene, whether responses are timely and useful, and how much they communicate over a session.

3.1 BENCHMARK CONSTRUCTION

As illustrated in Figure 2, we construct BLV-GATE from continuous egocentric videos collected from EgoBlind (Xiao et al., 2025) and VIABench (Liu et al., 2026). We retain videos from five everyday scene categories: *Home*, *Shopping & Service*, *Transportation*, *Outdoor Walking*, and *Public Facilities*. Corrupted, unreadable, severely occluded, and very short videos are removed during initial screening. The remaining videos are divided into sessions with consistent physical settings.

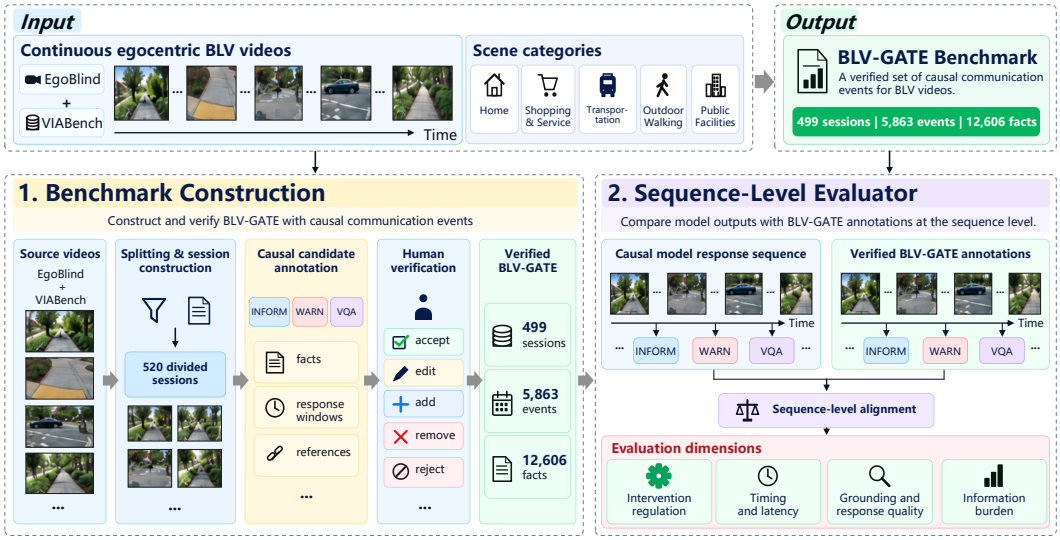


Figure 2: BLV-GATE construction framework.

To avoid spurious visual cues, we remove burned-in captions and watermarks while preserving text that naturally appears in the scene, such as signs and labels. An annotation agent then processes each session causally, generating a visual timeline and proposing `INFORM`, `WARN`, and `VQA` events. Each event is associated with an ideal response time, an acceptable response window, atomic visual facts, a reference response, and a decision rationale.

Three human reviewers verify the complete session through an interactive interface. They may accept, edit, add, or remove event annotations, and reject sessions whose visual content cannot be reliably verified. Starting from 520 screened sessions, automatic annotation produces 7,466 candidate event records. After session- and event-level verification, 499 sessions and 5,863 communication events remain. Detailed annotation instructions, prompts, and quality-control procedures are provided in Appendix A.

3.2 Dataset Statistics

Table 2 summarizes the distribution across scene categories. BLV-GATE contains 499 continuous egocentric video sessions with a total duration of 6.34 hours. It includes 5,863 communication events: 4,422 `INFORM`, 560 `WARN`, and 881 `VQA` events. The `VQA` subset contains 770 present-evidence questions and 111 delayed-evidence questions. Across all events, the annotations contain 12,606 atomic visual facts, categorized as critical, general, or low-value.

3.3 Evaluation Framework

BLV-GATE evaluates the complete response sequence generated during each session rather than isolated answers. An independent language-model judge aligns each non-empty response with the verified event using the response, the causal visual evidence, the active user query, and the corresponding event annotations. The earliest response attributed to an event within its acceptable response window is treated as the primary response and receives event credit. Responses that are unmatched, outside the acceptable window, or redundant are retained for diagnostic analysis rather than discarded. This sequence-level protocol allows us to evaluate not only what the assistant says, but also whether it speaks at the appropriate time and how frequently it interrupts the user. Detailed matching rules and judge prompts are provided in Appendix C.

Intervention Regulation. We treat every non-empty response as an intervention and report intervention precision, recall, and F1. Precision measures how often interventions can be attributed to verified events within their acceptable windows, whereas recall measures how many verified events receive at least one timely response. Silence is not scored as an event.

Table 2: Statistics of the current verified BLV-GATE dataset.

Scene	Sessions	Duration (h)	Inform	Warn	VQA	Total
Home	80	0.59	364	25	106	495
Shopping & Service	76	1.11	325	36	147	508
Transportation	84	0.84	314	93	125	532
Outdoor Walking	160	2.40	2,163	354	311	2,828
Public Facilities	99	1.40	1,256	52	192	1,500
Total	499	6.34	4,422	560	881	5,863

Timing and Latency. For each attributed response, we determine whether it falls within the event's acceptable response window. We further distinguish decision-point latency from model-output latency. Decision-point latency measures how long the system takes to make an intervention decision after the annotated ideal response time, while model-output latency measures the time required to produce the response after the decision trigger. We report latency distributions and the deadline violation rate, which captures responses that become available after the end of the acceptable window.

Grounding and Response Quality. For successful primary responses, weighted Fact Coverage (FC) measures the proportion of verified visual information correctly expressed by the response. Facts are weighted by their importance, so missing a critical safety fact has a larger effect than omitting optional context. Response quality is evaluated separately on four five-point dimensions: conciseness, understandability, value, and actionability.

Session-level Information Burden. We characterize communication burden across the complete session using false interventions per minute and output characters per minute. False interventions per minute counts non-empty responses that cannot be attributed to any verified event, whereas output characters per minute captures the volume of all generated communication, including user-query answers and responses that receive no event credit.

3.4 LLM-AS-A-JUDGE

We use LLM-as-a-Judge for event attribution, fact coverage, and response-quality assessment. We select GPT-5.6-Luna through a human-agreement study on 1,135 candidate responses, where it achieves the highest Spearman correlation and lowest mean absolute error among the evaluated judges. This study calibrates response-quality scoring rather than directly validating event attribution or fact coverage. The human-rating protocol, judge comparison, and prompts are provided in Appendix C.5.

4 SENSEBRIDGE AGENT

SenseBridge Agent is a companion baseline designed for the continuous interaction setting of BLV-GATE. It combines causal reasoning with two forms of short-term memory. The first is short-term visual memory: at each decision point, the agent receives the current frame batch together with representative frames from the previous two decision rounds, allowing it to reason about recently observed scene changes. The second is longer-horizon textual memory: the agent receives the responses generated in previous rounds and uses them to track information that has already been communicated, thereby reducing repetition over the session. Active user queries are also included in the context when answering questions.

Based on this visual and textual context, the agent reasons about new scenes, spatial changes, important objects, visible dangers, ongoing actions, and evidence needed for user questions. It returns structured reasoning, a binary `intervene`/`silence` decision, and a user-facing response. The agent is prompted to intervene when the current observation provides useful new information and to remain silent when the scene is unchanged or the information has already been conveyed. Non-empty responses are normalized as interventions, while empty responses are treated as silence. More implementation details are provided in Appendix B.

Table 3: Main baseline results on the complete BLV-GATE benchmark.

Category	Model	Intervention Reg. P(↑)	R(↑)	F1(↑)	Timing and Latency Deci.(↓)	Out.(↓)	DVR(%)(↓)	Quality FC(↑)	Qual.(↑)	Info. Burden False/min(↓)	Chars/min(↓)
Video LMs	JoyAI	0.278	0.221	0.246	0.983	4.843	69.13	10.16	3.40	6.27	954
	LiveStar-8B	0.229	0.097	0.136	0.763	–	–	5.19	2.72	3.92	1,047
	MMDuet2	0.186	0.280	0.224	0.705	–	–	11.18	2.95	10.61	9,774
	VideoLLM-online	**0.360**	0.088	0.141	0.455	**0.665**	**8.99**	12.75	3.36	**2.28**	**152**
	SimpleStream	0.294	0.508	0.373	0.606	0.799	12.64	28.51	3.50	5.41	6,664
Periodic VLM	GPT-5.6-Luna	0.227	0.732	0.347	0.502	7.738	99.87	42.19	**3.85**	9.99	10,674
	Qwen3-VL-4B	0.214	0.688	0.326	0.543	0.923	9.52	22.95	3.38	13.15	16,145
	Qwen3-VL-8B	0.198	0.639	0.303	0.547	0.983	13.32	23.76	3.51	14.23	9,063
SenseBridge Agent	GPT-5.6-Luna	0.257	**0.815**	**0.390**	**0.367**	12.145	99.79	**51.78**	3.81	8.47	8,969
	Qwen3-VL-8B	0.276	0.451	0.343	0.715	1.399	29.02	23.49	3.37	6.85	4,133

P, R, and F1 denote intervention precision, recall, and F1, respectively. Decision and Output denote the mean decision-point latency and model-output latency in seconds, respectively. DVR denotes the deadline violation rate in percent. FC is weighted Fact Coverage, and Quality is the mean response quality score on a one-to-five scale. False/min denotes false interventions per minute, and Chars/min denotes the number of model-output characters per minute over the complete session. "–" indicates that the required wall-clock timestamps are unavailable for the corresponding baseline.

5 EXPERIMENTS AND ANALYSIS

5.1 EXPERIMENTAL SETUP

Evaluation Setup. We evaluate ten main baselines representing three interaction strategies. Video-language models include JoyAI (Video Understanding Team of JoyAI-VL, 2026), LiveStar (Yang et al., 2026), MMDuet2 (Wang et al., 2025a), VideoLLM-online (Chen et al., 2024a), and SimpleStream (Shen et al., 2026). Periodic VLM serves as an always-response baseline that generates a description at every decision point. SenseBridge Agent, described in Section 4, uses causal reasoning and visual and textual memory to decide whether to respond. All systems are tested on the same 499 verified sessions.

Unless otherwise specified, models receive two input frames per second (2 FPS) and make one decision per second (1 Hz). Before inference, all videos are preprocessed into frame sequences with a fixed spatial resolution of 780 pixels, ensuring a consistent visual input format across baselines. Each session is processed causally to produce a complete response sequence. For locally hosted models, inference uses vLLM on a single NVIDIA A800 GPU with 80 GB memory, bfloat16 precision, and a temperature of 0.8.

Evaluation Metrics. We assess model performance along four complementary dimensions. Intervention regulation is measured by intervention precision, recall, and F1. Timing and latency are assessed using decision-point latency, model-output latency, and deadline violation rate. Grounding and response quality are measured by weighted Fact Coverage and five-point scores for conciseness, understandability, value, and actionability. Session-level information burden is measured by false interventions per minute and output characters per minute. We report these metrics separately without combining them into a single overall score.

5.2 CAN MODELS BALANCE INTERVENTION ACCURACY WITH INFORMATION BURDEN?

We compare intervention performance and communication volume across ten baselines to examine whether broad event coverage can be achieved without excessive output. Table 3 presents a comprehensive performance comparison of 10 representative baselines.

Coverage and communication burden remain difficult to balance. Periodic VLM with GPT-5.6-Luna achieves 0.732 recall but produces 10,674 characters and 9.99 false interventions per minute. At the other extreme, VideoLLM-online generates only 152 characters per minute but recalls just 0.088 of annotated events. SimpleStream offers a more balanced operating point, achieving 0.294 precision and 0.508 recall.

Higher recall does not eliminate unnecessary interventions. SenseBridge Agent with GPT-5.6-Luna achieves the highest F1 (0.390) and recall (0.815) among the evaluated configurations, while reducing output volume relative to Periodic VLM with the same backbone. Nevertheless, its precision remains 0.257, with 8.47 false interventions per minute, indicating substantial room for more selective communication.

Table 4: Performance across three assistance tasks. For each task, we report recall (R), weighted fact coverage (FC), and response quality (Q).

Category	Model	Inform			Warn			VQA		
		R (%)	FC	Q	R (%)	FC	Q	R (%)	FC	Q
Video-language models	JoyAI	18.46	12.75	3.31	18.28	6.60	3.36	47.89	0.88	3.93
	LiveStar-8B	9.84	4.98	2.71	5.04	4.17	2.70	11.22	6.39	2.77
	MMDuet2	25.26	12.81	2.85	25.21	12.56	2.96	47.56	0.59	3.70
	VideoLLM-online	2.93	9.27	3.09	8.19	12.71	3.36	47.00	13.93	3.49
	SimpleStream	46.16	25.38	3.44	46.85	21.04	3.24	83.33	39.69	3.79
Periodic VLM	GPT-5.6-Luna	70.13	37.39	3.75	69.54	47.99	3.96	95.11	57.52	4.29
	Qwen3-VL-4B	66.15	20.02	3.29	54.41	16.30	3.22	93.56	35.61	3.86
	Qwen3-VL-8B	61.41	20.04	3.43	49.37	20.15	3.32	87.67	38.13	3.96
SenseBridge Agent	GPT-5.6-Luna	86.52	49.55	3.72	87.14	58.58	4.09	52.67	58.77	4.27
	Qwen3-VL-8B	45.45	21.68	3.27	43.75	16.47	3.27	44.15	35.97	4.01
Average		41.33	21.03	3.30	39.18	21.48	3.35	62.52	26.37	3.79

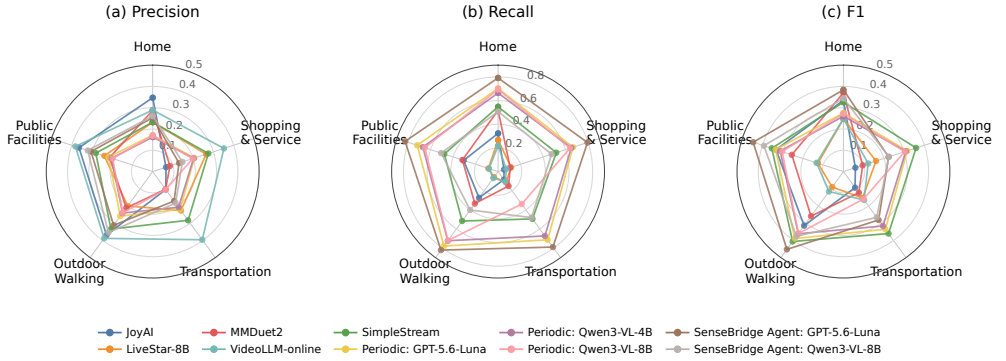


Figure 3: Intervention performance across scene categories.

5.3 Do models generalize across scene categories and assistance tasks?

We examine performance across assistance tasks and scene categories to identify where intervention and response quality vary.

Strong query answering does not imply strong proactive intervention. Table 4 reports recall, weighted Fact Coverage (FC), and response quality for the three non-silent Gold event types: `INFORM`, `WARN`, and `VQA`. `VQA` achieves the highest average Fact Coverage (26.37) and response quality (3.79). However, recall varies across policies: Periodic VLM with GPT-5.6-Luna reaches 95.1% VQA recall, whereas SenseBridge Agent with the same backbone reaches 86.5% `INFORM` and 87.1% `WARN` recall but only 52.7% `VQA` recall. Thus, explicit queries often facilitate high-quality responses, whereas the proactive policy preferentially covers environmental events.

Intervention performance varies across everyday environments. Figure 3 reports intervention precision, recall, and F1 across five scene categories. Outdoor Walking and Home achieve the highest average F1 values (0.312 and 0.306), whereas *Shopping & Service* and *Transportation* remain the most challenging scenes, at 0.217 and 0.225. SenseBridge Agent with GPT-5.6-Luna illustrates this variation: its F1 reaches 0.452 in *Outdoor Walking* but falls to 0.223 in *Shopping & Service*, where precision is only 0.130 despite 0.795 recall. These results show that scene context affects not only event coverage but also the risk of unnecessary intervention. More results with detailed metrics are shown in Appendix D.1.

5.4 Can models respond at the appropriate time?

We compare decision-point and model-output latency to examine whether responses can be delivered within the acceptable time windows.

Generation latency limits timely delivery. For successful primary responses, mean decision-point latency remains below one second across all baselines with available records. Output latency, how-

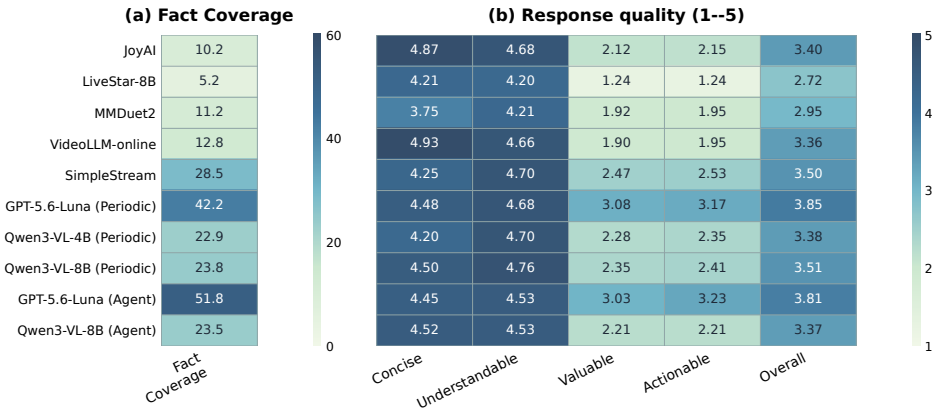


Figure 4: Quality heatmap on the complete BLV-GATE benchmark, averaged across correctly attributed and timed primary responses, on a 1–5 scale.

ever, varies substantially. Despite achieving the highest intervention F1, SenseBridge Agent with GPT-5.6-Luna requires 12.145 seconds per output and misses 99.79% of response deadlines. Periodic VLM with the same backbone also incurs 7.738 seconds of latency. These delays may reflect API overhead and, for SenseBridge Agent, additional complicated framework processing. In contrast, locally hosted VideoLLM-online and SimpleStream produce outputs in 0.665 and 0.799 seconds, respectively, highlighting the latency advantage of simpler inference pipelines.

5.5 ARE MODEL RESPONSES REALLY HELPFUL?

We assess grounding and four response-quality dimensions on successful responses to distinguish readable language from useful assistance. Figure 4 reports weighted Fact Coverage and quality scores across four dimensions.

Readability exceeds practical usefulness. VideoLLM-online scores highly on conciseness (4.93) and understandability (4.66), but much lower on value and actionability. This pattern holds broadly across baselines, showing that clear, concise responses can still provide little assistance.

Stronger backbones improve fact coverage and response quality. Within Periodic VLM, GPT-5.6-Luna achieves higher FC (42.2) than Qwen3-VL-8B-Instruct (23.8) and a higher overall quality score (3.85 versus 3.51). The gains are concentrated in value and actionability, which is similar to SenseBridge Agent. These results suggest that stronger backbones contribute more useful, grounded content, not merely more fluent responses.

5.6 HOW DO FRAME RATE AND DECISION FREQUENCY AFFECT CONTINUOUS ASSISTANCE?

We select JoyAI, SimpleStream, Periodic VLM and SenseBridge Agent with Qwen3-VL-8B-Instruct to run rate experiments, including 1 FPS/1 Hz, 2 FPS/2 Hz, 4 FPS/1 Hz, and 4 FPS/2 Hz with their main 2 FPS/1 Hz baselines. More detailed results are included in Appendix D.2. Figure 5 shows how intervention behavior changes with the input frame rate under different decision frequencies. The horizontal axis denotes the input frame rate, while solid and dashed lines represent 1 Hz and 2 Hz decision frequencies, respectively. Increasing the decision frequency generally raises recall but also increases false interventions, with the effect being particularly pronounced for Periodic VLM. In contrast, changing the input frame rate produces model-dependent trends: JoyAI benefits from higher frame rates in precision, whereas SimpleStream and SenseBridge Agent show no consistent monotonic improvement in recall. These results demonstrate that input FPS and decision Hz affect different parts of the interaction loop: decision frequency primarily controls how often the model can act, while input FPS changes the visual evidence available at each decision.

5.7 WHAT FAILURE MODES EMERGE IN COMPLETE INTERACTION SESSIONS?

Figure 6 illustrates three representative session-level behaviors. Periodic VLM correctly identifies a descending escalator at 8.0 seconds, but repeatedly restates it through 47.0 seconds and then incor-

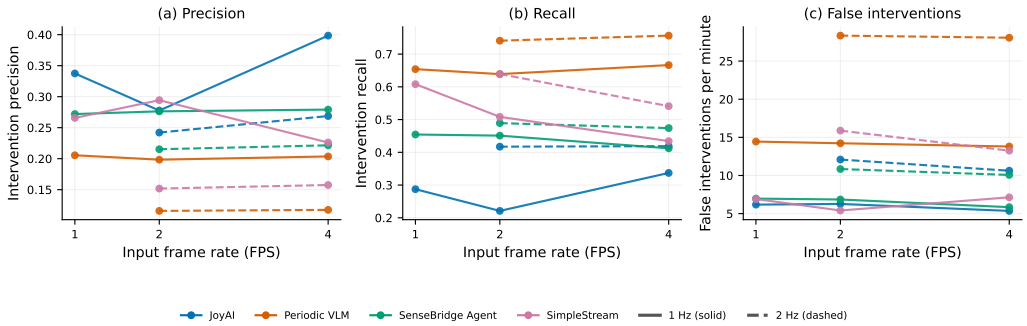


Figure 5: Effects of input frame rate and decision frequency on continuous assistance.

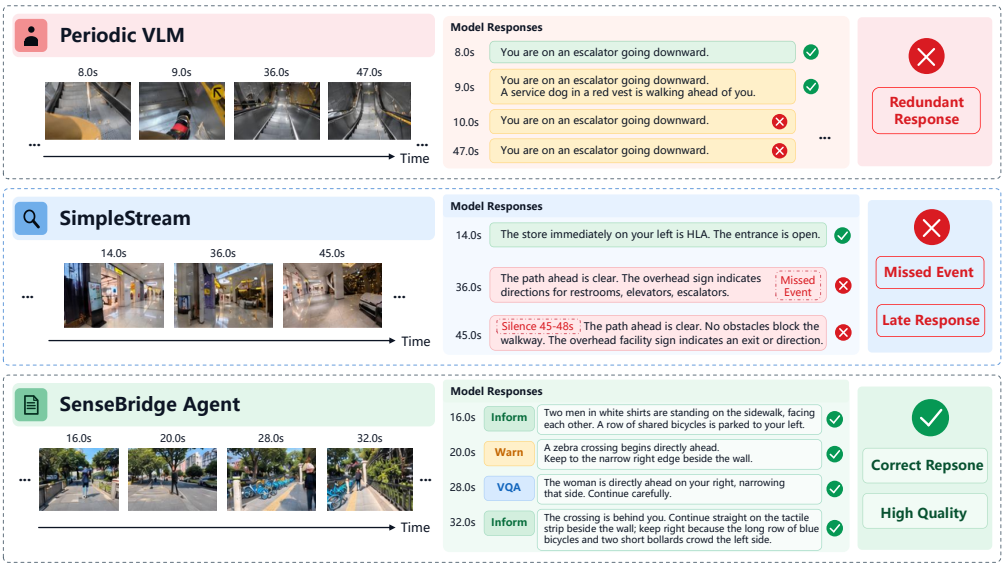


Figure 6: Case study highlighting the differences between models' different response policy.

rectly reports upward movement. SimpleStream correctly identifies an open store at 14.0 seconds, but misses the later navigation-relevant overhead sign and produces a false generic description. In contrast, SenseBridge Agent provides four correct, temporally localized updates that track the evolving sidewalk layout, crossing, and obstacles. These examples show that effective continuous assistance requires both selective intervention and context-sensitive guidance.

6 Discussion and Conclusion

This paper introduces BLV-Gate, a benchmark for evaluating communication regulation in continuous visual assistance for blind and low-vision users. Built from 499 verified egocentric video sessions and 5,863 temporally localized communication events, BLV-Gate evaluates complete causal response sequences across intervention regulation, timing and latency, grounding and response quality, and session-level information burden. Our results reveal a central challenge: stronger visual understanding alone does not yield effective continuous assistance. Systems that improve event coverage often incur more unnecessary interventions, while selective systems may miss important events; similarly, stronger backbones improve grounding and response quality but do not resolve delivery latency. These trade-offs suggest that continuous assistance requires joint optimization of perception, intervention policy, and system efficiency rather than improvements to any single component. BLV-Gate remains a controlled benchmark and does not fully capture real-world user–assistant collaboration, deployment constraints, or end-to-end task success. Future work should therefore extend evaluation toward interactive, task-oriented settings while developing assistants that jointly regulate communication, maintain grounded context, and operate under practical latency and resource constraints.

ETHICS STATEMENT

The BLV-GATE benchmark presented in this work aims to support research on continuous visual assistance for blind and low-vision users. We believe that the benchmark itself will not produce direct negative social impacts, while recognizing that assistive systems developed with it should be used with caution and appropriate ethical oversight in real-world settings. Although BLV-GATE evaluates communication regulation for BLV assistance, the source datasets, base models, and annotation process may inadvertently reflect biases or may not fully represent the needs and preferences of all users. Future work could examine the fairness and accessibility of these evaluation dimensions across diverse users, environments, and interaction contexts. We particularly consider the following ethical aspects: (1) compliance with the usage conditions of the source video resources; (2) exclusion of sessions that cannot be reliably reviewed because of privacy or content-quality concerns; (3) avoiding unsupported or stereotypical assumptions in event and response annotations; and (4) releasing the benchmark to promote community research on accessible visual assistance. Real-world applications should preserve privacy, user control, and appropriate ethical oversight. All research was conducted in strict compliance with ICLR ethics guidelines.

REPRODUCIBILITY STATEMENT

To support reproducibility, we provide detailed documentation of the benchmark construction, evaluation protocol, and experimental setup. Appendix A describes the session construction, causal candidate annotation, and human verification procedures. Appendix B provides the implementation details and policy prompt for SenseBridge Agent. Appendix C specifies the common causal evaluation protocol, response normalization, event attribution, metric computation, and LLM-judge prompts, while Appendix D reports more detailed results. We also release the verified BLV-GATE annotations, evaluation scripts, baseline experiments, prompts, and experiment configurations at https://anonymous.4open.science/r/BLV-Gate-7A03 to enable reproduction and extension of our results.

AI USE STATEMENT

We used GPT-5.6-Sol to assist with drafting and polishing portions of this paper, including English grammar or paragraph architecture. During dataset construction, we used GPT-5.6-Luna to generate candidate communication events, visual facts, reference responses, VQA pairs, and session-level keywords; the full procedure is described in the paper. All generated annotations were subsequently reviewed and verified by human reviewers. All code was written, reviewed, and verified by the authors. All prompts contained no private or sensitive data. Large language models did not provide any novel algorithmic ideas or academic claims; the authors take full responsibility for the content. Large language models are not authors of this paper.

REFERENCES

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pp. 2425–2433, 2015.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18407–18418, 2024a.

Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Bin Lin, Zhenyu Tang, et al. Sharegpt4video: Improving video understanding and generation with better captions. *Advances in Neural Information Processing Systems*, 37:19472–19495, 2024b.

Kanzhi Cheng, Wenpo Song, Jiaxin Fan, Zheng Ma, Qiushi Sun, Fangzhi Xu, Chenyang Yan, Nuo Chen, Jianbing Zhang, and Jiajun Chen. Caparena: Benchmarking and analyzing detailed image

captioning in the llm era. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 14077–14094, 2025.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24108–24118. IEEE, 2025.

Qi Gao, Heng Li, Yixin Zhou, Meixuan Zhou, Jieqiong Chen, and Xinyu Chai. Visassist: A visually impaired-captured video question answering benchmark for assistive systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 4149–4157, 2026.

Sitong Gong, Tianyu Yan, Caixin Kang, Bo Zheng, Xiang Ruan, Huchuan Lu, Kaipeng Zhang, Yoichi Sato, and Yifei Huang. Vinci2: Providing proactive assistance in continuous egocentric videos, 2026. URL https://arxiv.org/abs/2607.11523.

Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3608–3617, 2018.

Junhyeok Kim, Jaewoo Park, Junhee Park, Sangeyl Lee, Jiwan Chung, Jisung Kim, Ji Hoon Joung, and Youngjae Yu. Guidedog: A real-world egocentric multimodal dataset for blind and low-vision accessibility-aware guidance. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5545–5574. Association for Computational Linguistics, 2026.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pp. 12888–12900. PMLR, 2022.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22195–22206, 2024.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pp. 5971–5984, 2024a.

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. *arXiv preprint arXiv:2411.03628*, 2024b.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Ruiping Liu, Jiaming Zhang, Angela Schön, Karin Müller, Junwei Zheng, Kailun Yang, Kathrin Gerling, and Rainer Stiefelhagen. Objectfinder: Open-vocabulary assistive system for interactive object search by blind people. *arXiv preprint arXiv:2412.03118*, 2024.

Yunfeng Liu, Yuandong Yang, Jiarui Han, Zhenpeng Huang, Yuqing Tang, Xiangyu Zeng, Gangshan Wu, and Limin Wang. Viabench: A comprehensive video benchmark collected from blind individuals for visual impairment assistance. *arXiv preprint arXiv:2607.14660*, 2026.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12585–12602, 2024.

Daniela Massiceti, Luisa Zintgraf, John Bronskill, Lida Theodorou, Matthew Tobias Harris, Edward Cutrell, Cecily Morrison, Katja Hofmann, and Simone Stumpf. Orbit: A real-world few-shot dataset for teachable object recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10818–10828, 2021.

Junbo Niu, Yifei Li, 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? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18902–18913, 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24045–24055, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.

Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. Timechat: A time-sensitive multimodal large language model for long video understanding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14313–14323. IEEE, 2024.

Yujiao Shen, Shulin Tian, Jingkang Yang, and Ziwei Liu. A simple baseline for streaming video understanding. *arXiv preprint arXiv:2604.02317*, 2026.

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, et al. Moviechat: From dense token to sparse memory for long video understanding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18221–18232. IEEE, 2024.

Omkar Thawakar, Dinura Dissanayake, Ketan Pravin More, Ritesh Thawkar, Ahmed Heakl, Noor Ahsan, Yuhao Li, Ilmuz Zaman Mohammed Zumri, Jean Lahoud, Rao Muhammad Anwer, et al. Llamav-o1: Rethinking step-by-step visual reasoning in llms. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 24290–24315, 2025.

Video Understanding Team of JoyAI-VL. Joyai-vl-interaction: Real-time vision-language interaction intelligence, 2026.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024a.

Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. Videoagent: Long-form video understanding with large language model as agent. In *European Conference on Computer Vision*, pp. 58–76. Springer, 2024b.

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, 2025a. 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. *arXiv preprint arXiv:2507.09313*, 2025b.

Yueqian Wang, Xiaojun Meng, Yuxuan Wang, Jianxin Liang, Jiansheng Wei, Huishuai Zhang, and Dongyan Zhao. Videollm knows when to speak: Enhancing time-sensitive video comprehension with video-text duet interaction format. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 6338–6359. Association for Computational Linguistics, 2025c.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *Advances in Neural Information Processing Systems*, 37:28828–28857, 2024.

Junbin Xiao, Nanxin Huang, Hao Qiu, Zhulin Tao, Xun Yang, Richang Hong, Meng Wang, and Angela Yao. Egoblind: Towards egocentric visual assistance for the blind. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Zhenyu Yang, Kairui Zhang, Yuhang Hu, Bing Wang, Shengsheng Qian, Bin Wen, Fan Yang, Tingting Gao, Weiming Dong, and Changsheng Xu. Livestar: Live streaming assistant for real-world online video understanding. *Advances in Neural Information Processing Systems*, 38:31266–31304, 2026.

Zhiqiang Yuan, Ting Zhang, Yeshuang Zhu, Jiapei Zhang, Ying Deng, Zexi Jia, Peixiang Luo, Xiaoyue Duan, Jie Zhou, and Jinchao Zhang. Walkvlm: Aid visually impaired people walking by vision language model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9845–9854, 2025.

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.

Part I

Appendix

Table of Contents

A Annotation Pipeline **15**
- A.1 Session Split . . . 15
- A.2 Candidate Event Generation . . . 16
- A.3 Human Review . . . 17

B Details of SenseBridge Agent **18**
- B.1 Causal Visual Context . . . 19
- B.2 Textual Memory . . . 19
- B.3 User Query Injection . . . 19
- B.4 Structured Decision and Output Normalization . . . 20
- B.5 Agent Policy Prompt . . . 20

C Detailed Experiment Setup **20**
- C.1 Common Causal Evaluation Protocol . . . 20
- C.2 Evaluation Baselines . . . 21
- C.3 Common BLV Policy Prompt . . . 22
- C.4 Detailed Evaluation Metrics . . . 23
- C.5 Judge Model Selection . . . 24

D Detailed Experiment Results **24**
- D.1 Scene Breakdown . . . 24
- D.2 Effect of Input Frame Rate And Decision Frequency . . . 24

A Annotation Pipeline

This section details the annotation pipeline used to transform continuous first-person videos into verified evaluation sessions. The pipeline has three stages: (1) splitting source videos into coherent sessions, (2) generating causally grounded candidate events, VQA pairs, and keywords, and (3) human review of all automatically proposed annotations.

A.1 Session Split

Scene Segmentation Instruction

Task: Review each complete source video and divide it into temporally contiguous sessions. Each session should correspond to one primary real-world scene.

Segmentation principles:

- Watch the complete video before defining session boundaries.
- Mark the start and end time of each session according to meaningful transitions between primary scenes.
- Assign exactly one primary scene label to every session: Home, Shopping & Service, Transportation, Outdoor Walking, Public Facilities.
- Use the scene that best characterizes the main environment during the segment.
- Consecutive sessions must not overlap, and each session end time must be later than its start time.

Do not create a new session solely because of:

- a brief pause, camera movement, or minor visual change;
- an isolated hazard;
- a user question or answer;
- a short-lived object that does not indicate a transition to a different primary environment.

Add optional notes only when they clarify an important transition or ambiguity. Save the current video after checking all session boundaries and scene labels.

Figure 7: Instructions for human annotators in splitting sessions

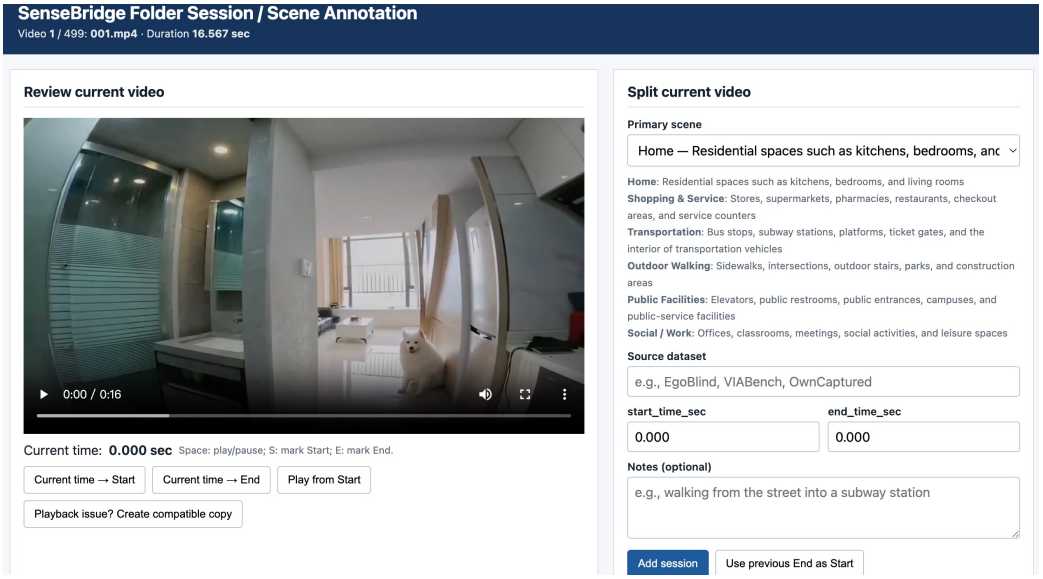


Figure 8: screenshot of human annotation process in session split

We first divide each source video into temporally contiguous sessions before generating event-level annotations. This step prevents unrelated physical environments from being merged into one interaction sequence and establishes the environmental context for subsequent intervention decisions. Annotators review the complete source video, identify meaningful transitions between primary environments, and assign one scene category to each resulting session.

A new session is created only when the primary real-world setting changes. Short camera movements, temporary occlusions, minor object changes, isolated hazards, and individual user questions do not by themselves constitute a session boundary. Consecutive sessions are non-overlapping and temporally ordered. Each session is assigned one primary category from *Home*, *Shopping & Service*, *Transportation*, *Outdoor Walking*, and *Public Facilities*. The detailed annotation instruction is shown in Figure 7, and the session-splitting interface is illustrated in Figure 8.

A.2 Candidate Event Generation

After session segmentation, candidate annotations are generated under a causal protocol. The annotation agent processes short overlapping frame windows and produces one grounded observation for each sampled frame. These observations are merged into a chronological visual timeline. Based only on information available up to each decision point, the temporal labeler proposes candidate communication events; it does not access future frames when generating observations, selecting event times, or writing supporting facts.

The labeler proposes three non-silent event types: `INFORM`, `WARN`, and `VQA`. `INFORM` represents newly available environmental information that can improve orientation or independent decision-making. `WARN` is reserved for immediate and visually supported hazards requiring timely attention. `VQA` is associated with a user question and is created only when sufficient evidence is available. Periods without a candidate event are retained as natural silence rather than exhaustively annotated as negative segments.

Each candidate event contains a response time, supporting frame indices, atomic visual facts, a reference response, and a short rationale. Facts are assigned one of three importance levels: `critical`, `general`, and `low_value`. A separate pass generates candidate `VQA` pairs and session-level keywords. The `VQA` generator supports present-evidence and delayed-evidence questions, while keywords summarize important places, objects, visible text, transportation elements, and safety-relevant content. The causal observation, temporal labeling, and VQA-generation prompts are provided below.

Prompt template for causal frame observation

You are a conservative BLV video observation annotator. This window contains an ordered sequence of first-person images. For EVERY image, write EXACTLY ONE short English sentence describing only what is visibly grounded in that image. Return exactly one description for every supplied frame index.
Write from the lived perspective of a blind or low-vision user: include details that could improve orientation, navigation, environmental awareness, independent decisions, or readiness to act. Attend to path and accessibility cues, useful object identity, location and state, meaningful social changes, readable real-world text, and temporal changes. Do not infer information outside the images, use future images to describe an earlier image, or state uncertain content as fact. Do not produce actions, event labels, response timing, or user answers.
Return strict JSON only: ”frame_descriptions”:[”frame_index”:0,”description”:”...”], ”window_summary”:”...”, ”global_facts”:[”...”], ”changes”:[”...”], ”uncertainty”:”...”.

Prompt template for temporal interaction labeling

You are the temporal interaction labeler. Use only the supplied causal frame descriptions to decide when an assistant should act. Actions may only be INFORM, WARN, or ANSWER.

Proactively INFORM when newly available grounded information materially improves a blind or low-vision user's environmental awareness, orientation, independent decision-making, or readiness to act. Do not repeat unchanged facts or narrate irrelevant clutter. Use WARN only for a clearly visible, immediate safety hazard or collision risk; a nearby object is not a warning unless it blocks the usable path or requires action now. An initial INFORM may establish the setting, usable path or layout, and the most important grounded landmark. Later INFORMs must be concise state updates. Keep facts atomic and observable. Use critical only for imminent danger or an immediate necessary action; otherwise use general or low_value.
Select the earliest eligible response frame supported by evidence available at or before that frame. Do not use future evidence. Write concise direct English suitable for spoken delivery; lead with the immediate hazard for WARN. Return strict JSON only: "events": ["event_id": "E001", "action": "INFORM/WARN/ANSWER", "response_frame_index": 0, "facts":["text": "Atomic grounded fact", "level" : "critical/general/low_value"], "response":"...", "reasoning": "observation": "...", "decision": "...", "confidence": 0.0, "evidence_frame_indices":[0]].

Prompt template for VQA candidate generation

You are designing a small synthetic VQA set, a continuously running BLV assistant. Review the frame-indexed causal timeline and representative first-person frames. Generate grounded questions a blind or low-vision user could naturally ask at a decision point. Prefer object identity, state, location, text, access, or an environmentally meaningful change. Do not ask generic questions or invent unseen details.
Use exactly one mode: present (current visual evidence answers the question), or delayed_evidence (evidence arrives naturally later in the stream). For delayed_evidence, do not predict the future or require navigation, search, tracking, tools, or a multi-step task. Do not generate long-horizon monitoring requests. For path safety, distinguish an object near a route from one that visibly blocks or makes the route unsafe.
Use only evidence available at or before the response frame. Return strict JSON only: "qa_pairs":["qa_id":"QA001", "question_frame_index":0, "response_frame_index": 0, "question": "...", "answer": "...", "facts": ["text":"...", "level": "critical/general/low_value"], "mode": "present/delayed_evidence", "confidence":0.0,"evidence_frame_indices":[0]]. Return an empty list if no grounded question is appropriate. These are proposals, not ground truth.

A.3 Human Review

All model-generated annotations are treated as editable proposals rather than Gold annotations. Human reviewers inspect the complete original session together with proposed events, facts, responses, VQA pairs, and keywords. The review instructions are shown in Figure 10. The event-review interface is illustrated in Figure 11, and the interface for reviewing VQA candidates and session-level keywords is shown in Figure 12. Figure 9 visualizes the session-level keywords.

For each candidate event, reviewers verify whether communication is useful at the proposed time, whether the event type is appropriate, and whether every fact is visually supported. They remove repetitive, unsupported, or low-value interventions; revise response timing when evidence is unavailable or the response would be too late; and edit reference responses to ensure that they are accurate, concise, and appropriate for spoken delivery. For VQA annotations, reviewers additionally verify the question, answer, evidence mode, supporting facts, and temporal alignment. A session is rejected only when reliable

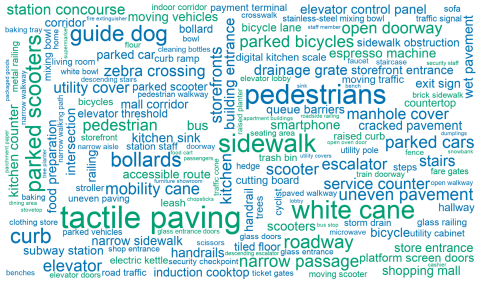


Figure 9: Word cloud of verified session-level keywords.

review is impossible, for example because the video is corrupted, persistently obscured, or otherwise unsuitable for benchmark annotation.

Human Verification and Keyword Annotation Instruction

Task: Review the complete video and refine the automatically generated event, VQA, and keyword annotations. The goal is to produce reliable annotations of information that would be useful for a blind or low-vision user.

Event verification:

- Check the event action: `INFORM`, `WARN`, or `ANSWER`.
- Keep `INFORM` only for new and useful environmental information.
- Use `WARN` only for immediate, clearly visible safety risks that require timely attention.
- Use `ANSWER` only when it directly answers a relevant user question with sufficient visual evidence.
- Remove repetitive, unsupported, or low-value events.
- Adjust the response time and acceptable time window so that the response is neither before the evidence appears nor too late to be useful.
- Ensure that facts are visually grounded. Mark immediate safety facts as `critical`; use `general` or `low_value` for other facts.
- Responses should be accurate, concise, understandable, and written in English.

VQA and keywords:

- Check each VQA question, answer, mode, timing, and supporting facts.
- Remove questions that are irrelevant, unsupported, or cannot be answered from available visual evidence.
- Keep representative keywords for important objects, places, text, transportation elements, or safety events.
- Remove duplicated, hallucinated, or uninformative keywords.

Reject a session only when the complete video cannot be reliably annotated, for example because it is corrupted, substantially obscured, unusable, or unsuitable for the benchmark.

Figure 10: Instructions for human annotators verify auto annotated events and keywords.

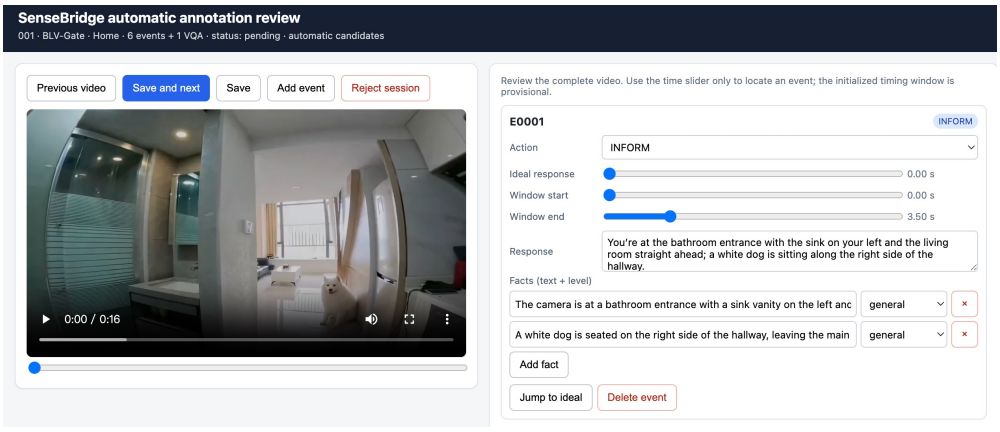


Figure 11: screenshot of human review in auto annotation events

B Details of SenseBridge Agent

This section provides implementation details of SenseBridge Agent beyond the high-level description in the main paper. The agent is implemented as a causal frame-sequence policy that processes

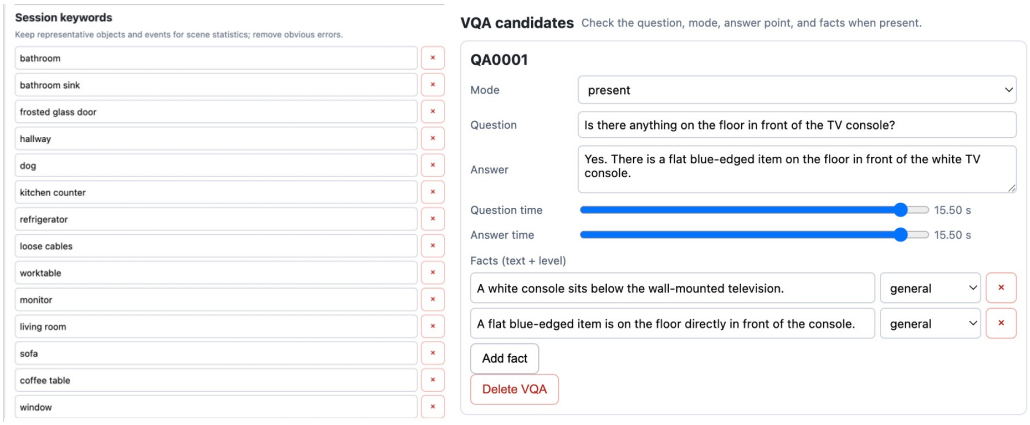


Figure 12: screenshot of human review in auto annotation keywords and VQA candidate

a session one decision round at a time. We describe its visual context construction, textual memory, query injection mechanism, and structured output policy below.

B.1 Causal Visual Context

At decision round t, the agent receives only frames that are causally available at that time. The visual input consists of the current frame batch sampled at the configured input rate, together with one representative frame from each of the two preceding decision rounds. The historical frames provide short-term temporal context for recognizing whether the scene has changed, while the current frame batch provides the evidence used for the present decision.

Historical frames and current-round frames are deliberately represented differently in the model request. The two historical frames are labeled as historical context and provided at low visual detail, whereas current-round frames are labeled as current observations and provided at high detail. This emphasizes currently available visual evidence while retaining sufficient recent context to distinguish a persistent object or layout from a newly emerging change. No future frame, future annotation, or future user query is included in the input.

B.2 Textual Memory

The agent maintains a lightweight textual memory containing only prior non-empty user-facing responses. Each memory entry records the decision index and the response text that was previously communicated. Silent rounds, raw model generations, and internal reasoning are not retained in the memory context.

This design makes previous communication explicit without maintaining an unrestricted dialogue trace. In particular, the agent can compare the current visual evidence with information that has already been delivered to the user and suppress repeated descriptions of unchanged objects, layouts, or activities. The textual memory therefore functions primarily as a repetition-control mechanism rather than as a source of additional visual evidence.

B.3 User Query Injection

User queries are treated as one-shot temporal inputs. At every decision round, the runner identifies queries whose timestamps have become available and whose identifiers have not been injected previously. Newly available queries are included both as explicit user messages and in the structured context payload sent to the model.

Once a query has been injected, its identifier is recorded and it is not reintroduced in subsequent rounds. This design prevents a query that has already been answered from acting as a persistent instruction that repeatedly triggers responses. It also preserves the causal interaction setting: a question can influence the model only from its arrival time onward, and the model must rely on visual evidence available at the current decision round when producing an answer.

B.4 Structured Decision and Output Normalization

At each round, the model is required to return a JSON object with three fields: `reasoning`, `action`, and `response`. The action space is binary: `intervene` or `silence`. When the action is `silence`, the response is required to be null. The reasoning field is retained for auditing but is not exposed as a user-facing output.

The inference runner normalizes outputs before evaluation. If the action field is invalid, the runner falls back to the presence of a non-empty response: an output with a response is treated as `intervene`, while an output without one is treated as `silence`. An intervention action without a valid response is converted to silence. Unparseable outputs and API failures are likewise recorded as silent rounds with a warning. This procedure yields a consistent binary response sequence in which every non-empty normalized response is evaluated as an intervention.

For every decision round, the runner records the decision timestamp, input-frame indices, newly injected user queries, normalized action, user-facing response, raw model output, response-ready time, token usage, and normalization warnings. Frames are resized and encoded before decision timing begins, so reported output latency measures the interval from submitting the prepared causal request to receiving the model output, without including one-time image-preparation overhead.

B.5 Agent Policy Prompt

Prompt template for SenseBridge Agent policy

You are a real-time visual assistant for a blind user. Whenever you encounter a new scene or useful new information, describe it. Give a comprehensive description when entering a new scene and briefly describe changes within the same scene. Warn about urgent visible dangers and answer questions. Do not repeat facts that have already been described. Keep your response concise.
Respond when the current images contain new information that is useful to a blind user and has not already been said. Useful information includes a new scene or clear transition, obstacles or urgent danger, directions and spatial layout, important objects, ongoing actions, meaningful environmental changes, or evidence needed to answer an active user question. Choose silence when the current round adds no such information and no danger needs reporting.
Return JSON only with keys reasoning, action, response. action must be "intervene" or "silence". If action is silence, response must be null. Keep reasoning brief and grounded.

C Detailed Experiment Setup

This section provides implementation-level details for the evaluation protocol, baseline adaptation, metric computation, and judge-model selection. All methods are converted into a common causal response-sequence format before evaluation. The evaluator receives only timestamped model responses and latency records; it does not require baselines to share the same internal action space, prompting format, or model architecture.

C.1 Common Causal Evaluation Protocol

Each baseline processes a session sequentially using only frames available up to the current decision time. The unified launcher prepares a causal frame sequence, passes the configured input frame rate and decision frequency to the corresponding adapter, and stores outputs in a shared response schema. Each record contains the decision video time, input-frame indices, user-facing response, and wall-clock timestamps for the decision trigger and response completion.

The common evaluation interface deliberately distinguishes a model's native output protocol from the benchmark-level intervention interpretation. Some baselines produce an explicit silence token, some emit structured actions, and others generate free-form text. After model-specific parsing, every non-empty user-facing response is treated as an intervention, whereas an empty response is

treated as silence. This normalization permits direct sequence-level comparison without modifying the original model architectures.

C.2 Evaluation Baselines

We evaluate baselines that span three distinct approaches to streaming visual interaction: trained proactive video-language models that explicitly decide when to respond, streaming video-language models designed for online dialogue, and controlled reference policies that isolate the role of recent context or intervention gating. All methods receive causally available frames under the common BLV-Gate frame-sequence protocol. We retain each method's original inference paradigm whenever possible and use adapters only to provide causal inputs, inject benchmark VQA turns at their annotated times, and serialize outputs into the common response-sequence schema.

JoyAI-VL-Interaction. JoyAI-VL-Interaction (Video Understanding Team of JoyAI-VL, 2026) is an 8B vision-first interaction model designed for real-time video interaction. Unlike conventional turn-based VLMs, it makes an internal decision at each time step to remain silent, produce a response, or delegate a difficult subtask to a background agent. Its interaction behavior is learned from time-aligned visual interaction data and further refined with reinforcement learning. The original system also uses a predictive video codec, AdaCodec, to allocate more visual tokens to changing content and fewer tokens to predictable frames.

For BLV-Gate, we use the released real-time interaction model in its vision-triggered response/silence mode. Delegation is disabled because it invokes an external background process rather than producing a directly comparable user-facing response sequence. The adapter streams causal frames to the official inference service and preserves the model's native decision behavior; a non-empty native output is recorded as an intervention and a silent output is recorded as null.

LiveStar. LiveStar (Yang et al., 2026) is a live streaming assistant that addresses online video understanding through adaptive streaming decoding. Its training procedure supports incremental video–language alignment over variable-length streams, and its response–silence decoding mechanism determines whether a response should be emitted through single-forward-pass verification. For long streaming inputs, LiveStar further uses peak-end memory compression together with a streaming key-value cache to reduce the cost of maintaining temporal context.

We evaluate the official LiveStar streaming inference implementation. At each causal decision tick, the model's SVeD-based response gate determines whether narration is generated. The adapter preserves this native gate: generated narrations are stored as responses, whereas decision ticks that do not trigger narration are stored as silence. The common BLV assistance policy is appended to the original task prompt without replacing LiveStar's streaming memory or response-silence mechanism.

MMDuet2. MMDuet2 (Wang et al., 2025a) is a proactive Video-MLLM that formulates response timing as a text-to-text decision problem. Given the dialogue history and visual context available up to the current streaming frame, the model autonomously determines whether to respond or remain silent. Rather than relying on manually selected decision thresholds or precisely annotated reply times, MMDuet2 is trained with multi-turn reinforcement learning to encourage timely and accurate proactive responses.

In BLV-Gate, we use the official streaming client, which maintains its conversation history and visual key-value cache across causal turns. The adapter supplies frames sequentially and injects visual questions only when they become active. MMDuet2's native `NO REPLY` output is normalized as silence; all other generated text is retained as a user-facing response. Thus, the benchmark evaluates its learned proactive response behavior rather than imposing an external intervention classifier.

VideoLLM-online. VideoLLM-online (Chen et al., 2024a) is built on the Learning-In-Video-Stream (LIVE) framework for temporally aligned, long-context dialogue over continuous video. LIVE combines a streaming language-modeling objective, a procedure for converting offline temporal annotations into streaming dialogue data, and an optimized inference pipeline for online interaction. The resulting model processes frames incrementally and maintains state over the evolving video stream rather than repeatedly reasoning over a complete pre-collected clip.

We replay each session through the stateful official VideoLLM-online inference pipeline. Frames are appended to one causal stream, while VQA turns are injected at their corresponding decision times. The adapter preserves the native streaming state and records each emitted response or silent turn in the common schema. Video preprocessing is completed before inference timing begins, so reported latency reflects online model processing rather than one-time frame conversion.

SimpleStream. SimpleStream (Shen et al., 2026) is a training-free streaming-video baseline that challenges the assumption that complex long-term memory is always necessary. At each time step, it provides an off-the-shelf VLM with only the most recent N causal frames. It uses no learned memory bank, retrieval module, compressed summary, or access to earlier frames outside the sliding window. The original study shows that this simple recent-context strategy can match or exceed more complex streaming systems on several benchmarks, while revealing a trade-off between current-scene perception and long-range memory.

We use the published default setting of $N = 4$ recent frames. At every decision tick, the adapter sends the newest four frames from the causal stream to the VLM together with the current prompt and any newly active user query. The model has no retained visual or textual memory between decisions. A native `SILENCE` output is treated as silence; every other non-empty output is treated as an intervention. This baseline tests how far recent visual context alone can support continuous assistance.

Periodic VLM. Periodic VLM is a controlled always-response reference policy rather than a separately trained proactive model. At each decision tick, an underlying VLM receives the current causal frame batch and generates a response under the same BLV-oriented instruction. It has no visual memory, textual response memory, learned intervention gate, retrieval component, or explicit mechanism for suppressing outputs according to prior communication.

We instantiate Periodic VLM with GPT-5.6-Luna, Qwen3-VL-4B-Instruct, and Qwen3-VL-8B-Instruct. This family provides a high-frequency communication reference that separates the ability to describe current visual evidence from the ability to decide whether speaking is warranted. Because the policy attempts generation at every decision point, it is expected to reveal the information-burden cost of replacing selective intervention with regular visual description.

SenseBridge Agent. SenseBridge Agent is our companion agentic baseline. It uses a VLM backbone together with a short causal visual context, a memory of previously communicated responses, and one-shot injection of newly active user queries. At each decision round, it receives the current frame batch, one representative frame from each of the two preceding rounds, and all prior non-empty responses. It then returns structured reasoning, a binary `intervene`/`silence` decision, and an optional user-facing response.

We evaluate SenseBridge Agent with GPT-5.6-Luna and Qwen3-VL-8B-Instruct backbones. Unlike Periodic VLM, it explicitly conditions its communication decision on whether current evidence provides new user-relevant information that has not already been conveyed. Unlike SimpleStream, it retains limited visual context and textual response memory. Its implementation, query-handling mechanism, output normalization, and complete policy prompt are described in Appendix B.

C.3 Common BLV Policy Prompt

To make baseline outputs comparable in the BLV assistance setting, we provide every baseline with a common high-level communication policy. The policy specifies the intended user-facing behavior—describe newly useful information, warn about urgent visible hazards, answer user questions, avoid repetition, and remain concise—but does not replace a method's native streaming architecture, memory mechanism, or response-selection procedure.

For methods with an official system prompt, including VideoLLM-online, MMDuet2, and JoyAI-VL-Interaction, we append the common policy to the native prompt. The original prompt remains responsible for the model's expected input format and native response protocol. For SimpleStream, Periodic VLM, and SenseBridge Agent, the policy serves directly as the base instruction. LiveStar receives the policy through its native streaming task prompt while retaining its SVeD response–silence decoding mechanism.

Common BLV assistance policy

You are a real-time visual assistant for a blind user. Whenever you encounter a new scene or notice useful new information, describe it. Give a comprehensive description when entering a new scene, and briefly describe changes within the same scene. Warn about urgent visible dangers and answer questions. Do not repeat facts that have already been described. Your response should be concise. If there is nothing useful to say, output exactly "SILENCE".

The final silence instruction is removed when a baseline already defines an incompatible native silence token. In particular, MMDuet2 retains its original `NO REPLY` protocol, JoyAI-VL-Interaction retains its `</silence>` action token, and LiveStar retains its SVeD-controlled response–silence decision. Their native outputs are subsequently normalized into the common response-sequence schema: any non-empty user-facing text is treated as an intervention, while native silence outputs are represented as null responses.

C.4 Detailed Evaluation Metrics

Event attribution and intervention regulation. The evaluator first attributes every non-empty response to a temporally nearby Gold event or to no event. The attribution judge receives the candidate response, evidence available to the model, active user queries, and candidate event facts within a five-second temporal neighborhood. It does not receive the model action label or the Gold action label.

For each Gold event, only the earliest matched response within the acceptable response window receives correct-intervention credit. A matched response outside the window is marked as a *misaligned intervention*; a later response matched to an event that already has a primary response is *redundant*; and an unmatched response is a *false intervention*. Intervention precision, recall, and F1 are computed from timely primary matches.

Prompt template for event attribution judge

You are an independent event-attribution judge for an online visual assistant for a blind or low-vision user. Decide which, if any, candidate Gold event the model response is addressing. Use only the response, the evidence available to the model, the active user query, and the candidate event facts and times. Do not infer from or request any Action label.
Return strict JSON only: "attribution": "matched—no_match", "event_id": "E0001 or null", "confidence":0.0, "reason":"short evidence-grounded explanation". Choose matched only when the response semantically refers to one candidate event. If no candidate fits, return no_match and null event_id. If two candidates are indistinguishable, choose the best-supported event only when confidence is at least 0.5; otherwise return no_match.

Timing and latency. Timing is evaluated only for correctly matched primary responses. Decision-point latency is the non-negative difference between the annotated ideal response time and the model decision time. Model-output latency is the wall-clock interval from the recorded decision trigger to the response-ready timestamp. A deadline violation occurs when the decision time plus model-output latency exceeds the endpoint of the event's acceptable response window. The deadline violation rate is computed only for responses with valid wall-clock latency records; methods without such timestamps are not assigned a synthetic latency value.

Grounding and response quality. For every successful primary response, the grounding judge compares the response with the verified atomic facts of the matched Gold event. Each required fact is labeled as `hit`, `missed`, or `contradicted`. We report weighted Fact Coverage (FC) as the only grounding metric, assigning weights of 5, 3, and 1 to `critical`, `general`, and `low_value` facts, respectively. Consequently, omitting an immediate safety fact has a larger effect than omitting optional descriptive context. Response quality is scored on four separate five-point dimensions: conciseness, understandability, value, and actionability. These scores are computed only for successful

primary responses and are not allowed to compensate for an unnecessary, redundant, or delayed intervention.

Session-level information burden. At the session level, we report false interventions per minute and output characters per minute as observable indicators of information burden.

Prompt template for fact-coverage and response-quality judge

You are an evaluation judge for an online visual assistant for a blind or low-vision user. Use only the supplied causal evidence, Gold facts, and candidate response. Do not use future information. Judge required facts semantically rather than by exact string matching. Do not reward a claim merely because it is plausible.
Return strict JSON only: "fact_matches":["fact_id": "F001", "status": "hit/missed/contradicted"], "response_quality":"concise": 1, "understandable": 1, "valuable": 1, "actionable":1. Each response-quality score is an integer from 1 to 5. Do not duplicate Fact Coverage in these scores. A descriptive response can be actionable when it resolves a query or supports the user's next decision; an imperative is not required.

C.5 Judge Model Selection

The semantic evaluator is selected through a separate response-quality agreement study. The human reference consists of 1,135 correctly attributed candidate responses. Each response is rated on conciseness, understandability, value, and actionability using a five-point scale. The annotators view the candidate response and reference facts, but do not see model identities, system outputs, videos, or evaluator scores. The annotation interface and rating instructions are shown in Figure 13 and Figure 14.

Each candidate judge receives the same compact input: the candidate response, the active user query when present, and the reference facts. It does not receive video frames or source-model metadata. We compare Qwen3-8B, DeepSeek-V4-Flash, GPT-5.4, and GPT-5.6-Luna against the A01 human annotation pass using Pearson correlation, Spearman correlation, Kendall correlation, mean absolute error, exact agreement, and agreement within one rating point. Figures 15, 16, and 17 visualize the score distributions, item-level agreement, and dimension-level behavior, respectively.

GPT-5.6-Luna is selected as the semantic judge because it achieves the strongest overall Pearson correlation (0.619), Spearman correlation (0.592), and lowest mean absolute error (0.590). GPT-5.4 achieves the highest within-one-point agreement (0.844), indicating that model selection depends on the target agreement criterion. We therefore use GPT-5.6-Luna as the fixed benchmark judge while retaining the human-quality study as a calibration analysis rather than treating automatic scores as a replacement for human evaluation.

D Detailed Experiment Results

In this section,we also report the detailed results across different scenes and additional ablation results of input frame rate and decision frequency.

D.1 Scene Breakdown

Tables 5-9 report the detailed results for Home, Outdoor Walking, Shopping & Service, Public Facilities, and Transportation, respectively. Each table reports sessions, intervention regulation, false interventions, grounding, and response quality; timing and latency are omitted because they are measured per model configuration rather than per scene.

D.2 Effect of Input Frame Rate And Decision Frequency

Table 10 lists the performance of selected baselines including JoyAI, SimpleStream, Periodic VLM and SenseBridge Agent with Qwen3-VL-8B-Instruct under different input frame rate and decision

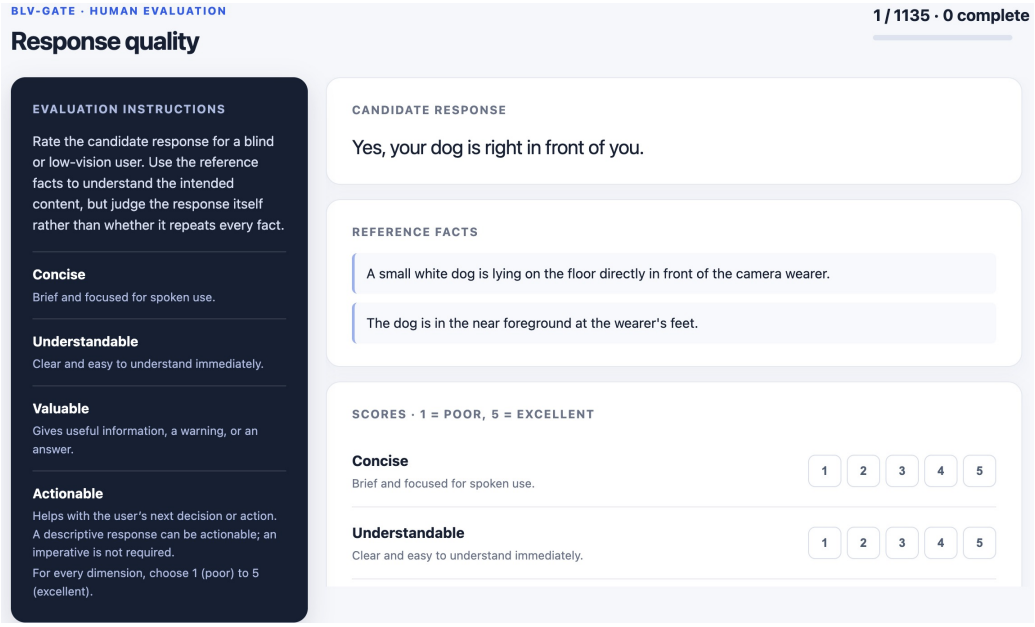


Figure 13: screenshot of human quality judge in the subset of 1135 response sequences

Response Quality Evaluation Instructions

Rate the candidate response for a blind or low-vision user. Use the reference facts to understand the intended content, but judge the response itself rather than whether it repeats every fact.

Concise. Brief and focused for spoken use.

Understandable. Clear and easy to understand immediately.

Valuable. Gives useful information, a warning, or an answer.

Actionable. Helps with the user's next decision or action. A descriptive response can be actionable; an imperative is not required.

For every dimension, choose 1 (poor) to 5 (excellent).

Figure 14: Instructions for human annotators in quality judges

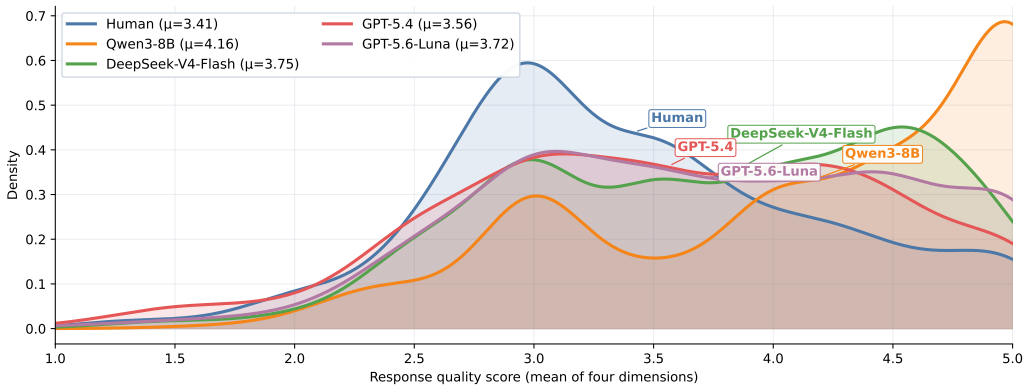


Figure 15: Distribution of response-quality scores assigned by the human annotator and candidate LLM judges. Each score is the mean of concise, understandable, valuable, and actionable ratings.

frequency configurations. Figure 18 visualizes all 20 measured operating points, including those four main-baseline reference points.

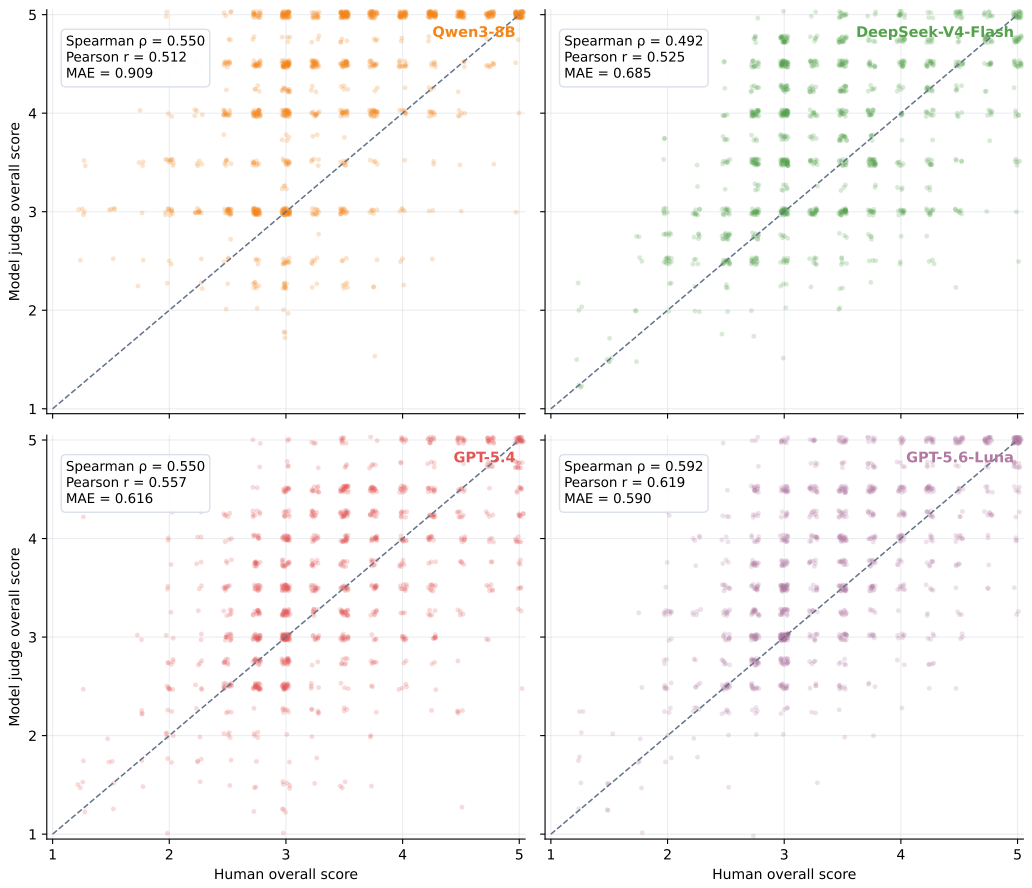


Figure 16: Item-level agreement between human and candidate LLM-judge response-quality scores. Panels, in reading order, correspond to Qwen3-8B, DeepSeek-V4-Flash, GPT-5.4, and GPT-5.6-Luna; the dashed line denotes exact agreement. GPT-5.6-Luna achieves the strongest overall agreement and lowest absolute error.

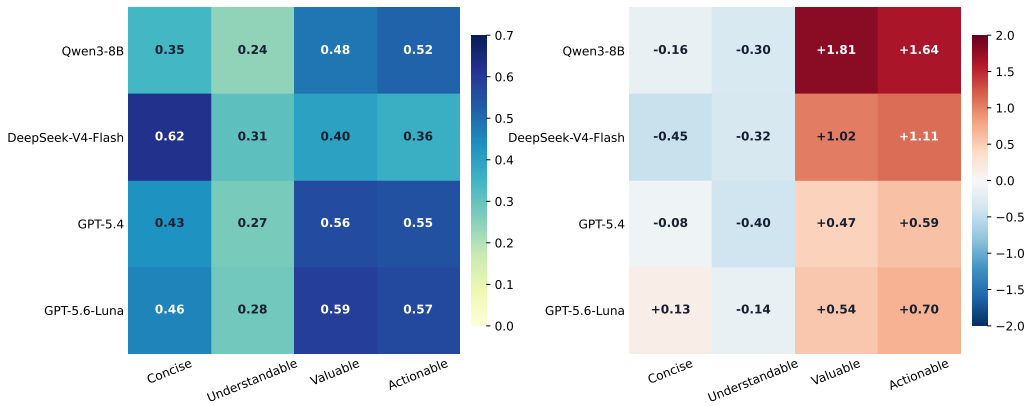


Figure 17: Dimension-level agreement between human and candidate LLM judges. The left heatmap shows Spearman rank correlation; the right heatmap shows mean score bias (model minus human). GPT-5.6-Luna shows the strongest agreement on valuable and actionable ratings, and is selected for the benchmark's semantic judgments.

Table 5: Detailed performance in *Home*.

Category	Model	Intervention Regulation P	R	F1	Grounding and Response Quality FC	Conci.	Understand.	Value.	Action.	Overall	Info. Burden False/min	Chars/min
Video-language models	JoyAI	0.3474	0.3234	0.3349	9.3797	4.93	4.81	2.02	2.07	3.46	5.382	771
	LiveStar-8B	0.2318	0.2673	0.2483	5.0691	4.32	4.30	1.24	1.24	2.78	7.1382	2043
	MMDuet2	0.2873	0.5178	0.3696	19.1097	3.29	3.98	2.23	2.23	2.93	8.3845	7157
	VideoLLM-online	0.2893	0.2168	0.2479	3.6093	4.89	4.49	1.59	1.61	3.15	5.9202	326
	SimpleStream	0.2318	0.5477	0.3257	36.6082	4.12	4.74	2.76	2.78	3.60	7.818	6627
Periodic VLM	GPT-5.6-Luna	0.171	0.7028	0.2751	39.834	4.55	4.69	2.98	3.04	3.82	15.9193	9282
	Qwen3-VL-4B	0.1619	0.6654	0.2604	34.1381	4.11	4.75	2.65	2.69	3.55	18.2703	14896
	Qwen3-VL-8B	0.1692	0.6953	0.2721	30.2705	4.45	4.77	2.64	2.68	3.64	16.2025	9197
SenseBridge Agent	GPT-5.6-Luna	0.187	0.6673	0.2921	41.5514	4.43	4.57	2.91	3.05	3.74	13.6815	7582
	Qwen3-VL-8B	0.2609	0.5071	0.3445	30.5656	4.58	4.61	2.43	2.44	3.51	5.4669	3718

Table 6: Detailed performance in *Outdoor Walking*.

Category	Model	Intervention Regulation P	R	F1	Grounding and Response Quality FC	Conci.	Understand.	Value.	Action.	Overall	Info. Burden False/min	Chars/min
Video-language models	JoyAI	0.3688	0.274	0.3144	9.089	4.89	4.66	1.79	1.85	3.30	5.055	978
	LiveStar-8B	0.2	0.0578	0.0897	3.0879	4.22	4.09	1.17	1.17	2.66	3.2983	750
	MMDuet2	0.2109	0.3339	0.2585	9.0522	3.83	4.14	1.75	1.79	2.88	9.6518	12388
	VideoLLM-online	0.3883	0.0684	0.1163	6.899	4.95	4.76	1.82	1.88	3.35	1.5554	123
	SimpleStream	0.3344	0.5173	0.4062	24.7698	4.27	4.74	2.30	2.34	3.41	4.0204	5661
Periodic VLM	GPT-5.6-Luna	0.2593	0.779	0.3891	47.1585	4.44	4.67	3.26	3.36	3.93	5.8605	11409
	Qwen3-VL-4B	0.2406	0.7229	0.361	19.5347	4.31	4.73	2.11	2.18	3.33	9.506	15026
	Qwen3-VL-8B	0.2394	0.7195	0.3593	22.2522	4.58	4.79	2.27	2.34	3.50	9.5129	8606
SenseBridge Agent	GPT-5.6-Luna	0.2677	0.7691	0.3971	49.9013	4.50	4.53	3.14	3.39	3.89	4.2079	9619
	Qwen3-VL-8B	0.3541	0.4017	0.3764	21.3761	4.52	4.54	2.15	2.15	3.34	4.3329	3715

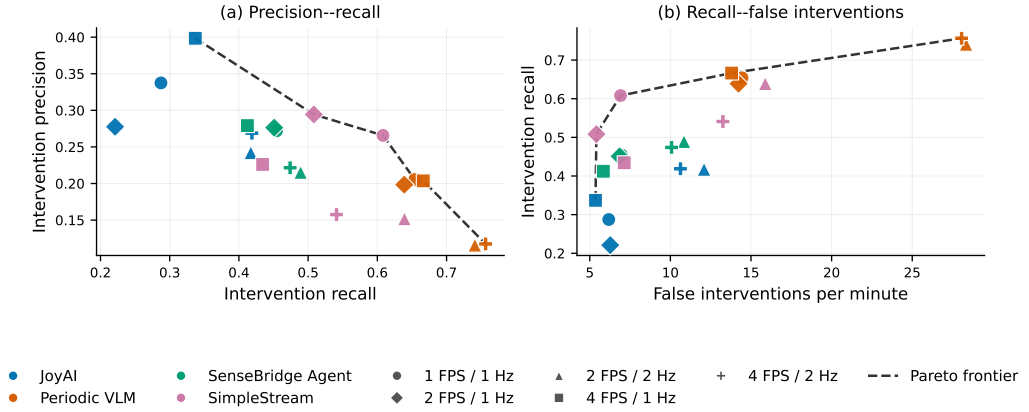


Figure 18: Precision–recall and recall–false-intervention trade-offs under different input frame rates and decision frequencies. Each point denotes one measured configuration of JoyAI, Periodic VLM, SimpleStream, or SenseBridge Agent with Qwen3-VL-8B-Instruct. Color identifies the baseline, marker shape identifies the joint FPS/Hz setting, and the dashed curves indicate the global Pareto frontiers. Higher precision and recall are preferable; lower false-intervention rate is preferable.

Table 7: Detailed performance in *Shopping & Service*.

Category	Model	Intervention Regulation P	R	F1	Grounding and Response Quality FC	Conci.	Understand.	Value.	Action.	Overall	Info. Burden False/min	Chars/min
Video-language models	JoyAI	0.0662	0.0531	0.0589	12.1622	4.92	4.81	2.31	2.36	3.60	9.1459	961
	LiveStar-8B	0.2635	0.1157	0.1608	6.7703	4.11	4.19	1.28	1.27	2.71	3.7995	1175
	MMDuet2	0.0854	0.1078	0.0953	12.4431	3.65	4.28	1.90	1.92	2.94	13.0805	10620
	VideoLLM-online	0.3523	0.0737	0.1219	8.5443	4.89	4.43	1.62	1.68	3.16	2.1926	120
	SimpleStream	0.2747	0.5182	0.359	29.4365	4.08	4.72	2.60	2.65	3.51	6.4577	7907
Periodic VLM	GPT-5.6-Luna	0.2063	0.6656	0.315	35.0422	4.43	4.65	2.87	2.94	3.72	12.8102	10947
	Qwen3-VL-4B	0.2019	0.6513	0.3082	24.6408	3.97	4.63	2.40	2.46	3.37	14.4021	17795
	Qwen3-VL-8B	0.1982	0.6395	0.3026	24.1283	4.37	4.72	2.37	2.42	3.47	14.5823	10204
SenseBridge Agent	GPT-5.6-Luna	0.2218	0.6284	0.3278	38.3421	4.27	4.47	2.75	2.92	3.60	10.092	8784
	Qwen3-VL-8B	0.1475	0.4764	0.2252	21.5821	4.41	4.49	2.26	2.25	3.35	11.9542	4040

Table 8: Detailed performance in *Public Facilities*.

Category	Model	Intervention Regulation			Grounding and Response Quality						Info. Burden	
		P	R	F1	FC	Conci.	Understand.	Value.	Action.	Overall	False/min	Chars/min
Video-language models	JoyAI	0.3643	0.3063	0.3328	12.7439	4.90	4.75	2.11	2.15	3.48	5.5413	1007
	LiveStar-8B	0.2393	0.0857	0.1262	7.4118	4.08	4.20	1.25	1.27	2.70	3.2939	957
	MMDuet2	0.2125	0.3171	0.2545	11.9519	3.73	4.32	2.13	2.17	3.09	9.2751	8731
	VideoLLM-online	0.3818	0.0806	0.1331	22.5877	4.95	4.75	2.16	2.21	3.52	2.0453	139
	SimpleStream	0.2805	0.4783	0.3536	31.6752	4.19	4.79	2.72	2.78	3.62	5.2559	6554
Periodic VLM	GPT-5.6-Luna	0.2186	0.7193	0.3353	37.6668	4.54	4.72	2.92	2.97	3.79	10.9875	9880
	Qwen3-VL-4B	0.2033	0.6688	0.3118	25.207	4.21	4.70	2.38	2.44	3.43	14.4121	16511
	Qwen3-VL-8B	0.1978	0.6509	0.3034	25.1551	4.50	4.77	2.44	2.50	3.55	15.3396	8652
SenseBridge Agent	GPT-5.6-Luna	0.2326	0.6912	0.3481	40.4377	4.45	4.54	2.98	3.14	3.78	9.1562	8616
	Qwen3-VL-8B	0.3213	0.5060	0.3931	26.1939	4.47	4.57	2.32	2.31	3.42	5.4699	4812

Table 9: Detailed performance in *Transportation*.

Category	Model	Intervention Regulation			Grounding and Response Quality						Info. Burden	
		P	R	F1	FC	Conci.	Understand.	Value.	Action.	Overall	False/min	Chars/min
Video-language models	JoyAI	0.1043	0.0839	0.093	6.8421	4.88	4.69	2.22	2.23	3.51	7.786	917
	LiveStar-8B	0.2259	0.1153	0.1527	4.4295	4.31	4.23	1.24	1.24	2.75	4.6279	1180
	MMDuet2	0.1032	0.1509	0.1226	10.6412	4.04	4.54	2.02	2.03	3.16	13.8639	4756
	VideoLLM-online	0.396	0.1038	0.1645	20.6897	4.92	4.73	2.22	2.30	3.54	2.304	181
	SimpleStream	0.2833	0.4948	0.3603	30.774	4.08	4.73	2.56	2.64	3.50	6.5744	8095
Periodic VLM	GPT-5.6-Luna	0.2189	0.7128	0.335	41.2626	4.54	4.72	3.04	3.17	3.87	12.2749	10511
	Qwen3-VL-4B	0.2067	0.673	0.3163	23.8128	4.15	4.67	2.35	2.39	3.39	16.1878	17426
	Qwen3-VL-8B	0.1043	0.3396	0.1596	22.2798	4.43	4.68	2.28	2.32	3.43	23.9937	9455
SenseBridge Agent	GPT-5.6-Luna	0.2403	0.6897	0.3564	45.6871	4.41	4.53	2.93	3.11	3.74	8.8189	8914
	Qwen3-VL-8B	0.1835	0.4812	0.2657	21.0476	4.47	4.53	2.14	2.11	3.31	10.5866	4609

Table 10: Additional experiment of selected baselines in 1 FPS/1 Hz, 2 FPS/2 Hz, 4 FPS/1 Hz, 4 FPS/2 Hz configurations.

Model	FPS	Hz	Intervention Regulation			Grounding and Response Quality						Information Burden	
			P	R	F1	FC	Conci.	Understand.	Value.	Action.	Overall	False/min	Chars/min
JoyAI	1	1	0.337	0.287	0.310	16.07	4.83	4.68	1.93	1.95	3.35	6.18	1310
JoyAI	2	2	0.242	0.417	0.306	16.35	4.85	4.69	1.99	2.01	3.39	12.09	2438
JoyAI	4	1	0.398	0.337	0.365	18.33	4.87	4.68	2.12	2.15	3.45	5.36	1176
JoyAI	4	2	0.269	0.419	0.327	16.09	4.88	4.71	1.97	2.02	3.40	10.62	1998
SimpleStream	1	1	0.266	0.608	0.370	28.42	4.22	4.71	2.52	2.58	3.51	6.91	8672
SimpleStream	2	2	0.152	0.639	0.245	26.30	4.21	4.69	2.40	2.45	3.44	15.88	12769
SimpleStream	4	1	0.226	0.434	0.297	26.91	4.25	4.70	2.47	2.53	3.49	7.14	5819
SimpleStream	4	2	0.158	0.541	0.244	26.05	4.25	4.72	2.45	2.50	3.48	13.26	10200
Periodic VLM	1	1	0.205	0.654	0.313	21.47	4.63	4.78	2.32	2.36	3.52	14.44	7470
Periodic VLM	2	2	0.116	0.741	0.200	20.37	4.64	4.77	2.24	2.28	3.48	28.34	14814
Periodic VLM	4	1	0.204	0.666	0.312	24.19	4.60	4.75	2.39	2.44	3.54	13.80	8171
Periodic VLM	4	2	0.117	0.756	0.203	19.93	4.64	4.74	2.19	2.23	3.45	28.06	14374
SenseBridge Agent	1	1	0.257	0.443	0.326	23.85	4.70	4.75	2.41	2.46	3.58	6.93	3932
SenseBridge Agent	2	2	0.161	0.497	0.243	21.04	4.68	4.75	2.26	2.31	3.50	13.40	6917
SenseBridge Agent	4	1	0.243	0.439	0.313	26.19	4.67	4.76	2.44	2.49	3.60	7.74	3986
SenseBridge Agent	4	2	0.157	0.492	0.238	20.98	4.72	4.79	2.31	2.36	3.54	13.30	6681