EGOSPECT: EGOCENTRIC SPATIAL PERCEPTION AND EVENT-CASCADE TESTBENCH

Anonymous authors
Paper under double-blind review

ABSTRACT

EgoSPECT is a first-person audio-visual benchmark for evaluating event-driven embodied intelligence across perception, priority assessment, and behavioral decision-making. Unlike benchmarks that evaluate these capabilities in isolation, EgoSPECT uses explicitly grounded question chains to measure the full event cascade while localizing failures to individual stages. We collect approximately 60 hours of egocentric recordings across six indoor and outdoor scenarios and evaluate three multimodal large language models under audiovisual, audio-only, video-only, and text-only conditions. Across the evaluated tasks, perception is the primary bottleneck: models frequently miss weak or masked events, confuse event categories, and over-report ambient sounds as hazards. In the complex multi-event scenario, end-to-end accuracy of 27.1%–62.8% rises to 69.0%–79.2% under conditional evaluation, while the three tested models reach 100% on 96 closed-form downstream questions when the ground-truth event pair is supplied as a premise. Eight-way spatial direction estimation remains near chance, but controlled DOA intervention substantially improves performance, with Oracle DOA revealing further headroom. These results show that robust embodied perception requires not only recognizing events, but also distinguishing relevant events from ambient sound and jointly grounding event semantics with spatial location under noisy and concurrent conditions.

1 INTRODUCTION

Problem and Motivation: As wearable devices, smart glasses, and embodied assistants enter home care, mobility assistance, and humanoid robotics, these systems must continuously perceive surrounding events from a first-person perspective. A useful assistant does not merely "hear a sound"; it must complete a full cognitive chain: *spatial perception* $\rightarrow$ *event recognition* $\rightarrow$ *priority assessment* $\rightarrow$ *behavioral decision*, answering four progressively deeper questions: Where is the sound? What happened? How urgent is it? What should be done now? This chain is a core ingredient of situated awareness for embodied AI: an agent must bind event semantics to spatial location, distinguish relevant events from ambient sound, and act under uncertainty.

In safety-critical scenarios, knowing only "where the loudest sound is" is insufficient. What is needed is the joint association between *what* happened and *where* it happened. Under realistic acoustic conditions—background noise, room reverberation, concurrent sources, and weak target events—the loudest sound is not necessarily the most important, and conventional direction-of-arrival (DOA) estimators point toward the highest-energy source rather than the semantically relevant event.

Limitations of Existing Benchmarks: Existing work has three gaps. **Gap 1: single-capability evaluation.** Most benchmarks test an isolated ability—sound localization, audio event recognition, or instruction following—rather than the complete perception-to-decision chain. **Gap 2: lack of first-person real wearable scenarios.** Existing spatial-audio benchmarks rely on controlled microphone-array recordings or synthesized audio, which differ from the reverberation, noise, and occlusion encountered by a wearable device in the wild. **Gap 3: no cascaded benchmark that localizes errors stage by stage.** No benchmark links perception, prioritization, and decision while identifying the stage at which an error occurs. Prior first-person audio-visual resources confirm this picture: Ego4D treats audio as secondary and focuses on wearer activity rather than external safety-critical events; EgoSound, EgoAVU, and MRSAudio evaluate isolated sound understanding

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.

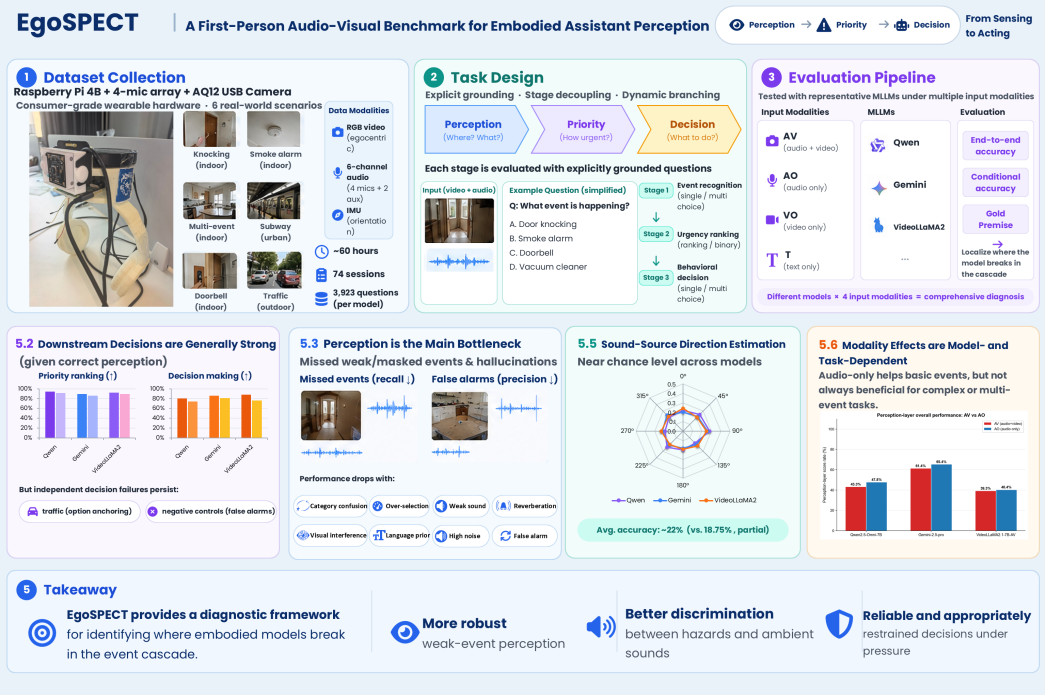

Figure 1: EgoSPECT overview: dataset collection, task design, evaluation pipeline, and representative results across six first-person scenarios.

or event localization without linking grounding, localization, priority, and action; SAVVY-Bench is dominated by human speech with limited concurrent sources and no negative controls; synthetic benchmarks such as SoundSpaces and BeDAViN provide multi-source noise but are simulated rather than captured by a worn device; RealImpact and SoundCam are audio-centric impact-sound datasets; and WAGIBench infers wearer intent rather than responding to external events. Table 4 summarizes these differences.

EgoSPECT: We present **EgoSPECT**, a first-person audio-visual event-cascade testbench comprising approximately 60 hours of recordings across 74 sessions and six indoor and outdoor scenarios. Explicitly grounded question design yields thousands of traceable evaluation items that decouple perception, prioritization, and decision-making (Figure 1). We deliberately use a low-cost DIY head-mounted recorder (about 2000 RMB, four-channel microphone array, synchronized 720p first-person camera, known geometry and timestamps) rather than commercial AI glasses, which often do not expose raw multi-channel audio synchronized with first-person video or impose closed interfaces and unknown preprocessing. This enables scalable collection in real indoor and outdoor environments while retaining raw acoustic cues; validation on commercial glasses is a natural future extension.

Beyond assistive wearables, EgoSPECT targets a core capability for humanoid and embodied AGI: multimodal spatial reasoning from first-person experience. A humanoid must not only detect events but bind what happened to where it happened, prioritize among concurrent events, and select appropriate behavior. By evaluating this cascade under noise, reverberation, weak events, and negative controls, EgoSPECT provides a diagnostic benchmark for the situated-awareness abilities that future embodied agents will need.

Contributions. EgoSPECT provides (i) a first-person audio-visual benchmark covering the complete event cascade (perception, priority, decision); (ii) a grounded evaluation protocol that writes upstream answers into downstream questions, enabling stage-level attribution of errors; (iii) prior-, premise-, and layout-controlled probes (question-only, blind-majority, gold-premise, option-shuffle) that validate the benchmark against language priors and yield a cross-model failure taxonomy of omission, over-selection, category confusion, and decision anchoring; (iv) systematic evidence

that perception—especially spatial and event grounding—is the primary bottleneck, with downstream reasoning becoming substantially more reliable once the correct premise is supplied; and (v) controlled DOA intervention experiments showing that explicit spatial cues substantially improve MLLM direction accuracy while Oracle DOA reveals considerable headroom. Our goal is not to replace learned spatial understanding with an external DOA module, but to test whether an embodied model can jointly bind an event's semantics with its spatial location under realistic multi-event conditions.

2 Related Work

Egocentric spatial audio and localization. Conventional direction-of-arrival methods (GCC-PHAT, SRP-PHAT, beamforming) estimate source direction from microphone arrays under controlled conditions, and SAVVY-Bench (Chen et al., 2025) extends spatial reasoning to dynamic Aria-glass (Lv et al., 2024) recordings. Simulation platforms (Chen et al., 2020; 2023) and their noisy multi-source extension (Shi et al., 2025) render spatial audio in virtual 3D scenes rather than capturing it on a moving wearer, while real-world spatial audio datasets rely on fixed third-person arrays (Shimada et al., 2023) or cover narrow sound classes such as impact sounds and room-acoustic signatures (Clarke et al., 2023; Wang et al., 2023a). Egocentric conversation datasets (Northcutt et al., 2022; Donley et al., 2021) record binaural or multi-channel audio from glasses but involve only seated multi-party speech. These approaches localize the loudest source without binding a direction to a semantic event and treat localization as the endpoint. EgoSPECT instead asks where a specific event is and whether it warrants action. **Egocentric audio-visual benchmarks.** Ego4D (Grauman et al., 2025) advances egocentric video understanding but treats audio as secondary, as do other large-scale first-person datasets centered on the wearer's own actions (Damen et al., 2018; 2022; Wang et al., 2023b; Yang et al., 2025), while EgoSound (Zhu et al., 2026), EgoAVU (Seth et al., 2026), and MRSAudio (Guo et al., 2025) target sound understanding or sound-event localization. Third-person audio-visual QA (Li et al., 2022; Yang et al., 2022) and audio understanding benchmarks (Sakshi et al., 2025; Wang et al., 2025) evaluate sound understanding without first-person spatial grounding. Such benchmarks assess an isolated capability without linking grounding, localization, and downstream action. EgoSPECT combines first-person audio-video input with a full perception-to-decision chain. **Wearable agents and goal inference.** WAGIBench (Veerabadran et al., 2025) studies wearable behavioral prediction but infers the wearer's internal intent rather than responding to external events. A wave of recent egocentric benchmarks evaluates MLLMs on first-person QA, situated awareness, or smart-glasses assistance (Cheng et al., 2024a; Ye et al., 2025; Xiao et al., 2025; Yang et al., 2025; Chang et al., 2025; Lei et al., 2026; Jiang et al., 2026; Li et al., 2026), but these efforts are vision-only and none requires localizing and prioritizing external sound events. EgoSPECT instead asks, given what is happening and where, what the agent should do. A detailed dimension-by-dimension comparison with these and further related benchmarks is provided in § A.

3 The EgoSPECT Testbench

3.1 Task Formulation: The Perception–Priority–Decision Chain

EgoSPECT models an embodied assistant's environmental understanding as a three-stage cascade: **perception** $\rightarrow$ **priority** $\rightarrow$ **decision**. **Perception** jointly establishes *what* happened (event identity) and *where* it happened relative to the wearer (spatial location). **Priority** determines which perceived event matters more, and **decision** selects what the agent should do. Each stage takes the preceding stage's output as its input: priority ranks the perceived events, and decision acts on the highest-priority one.

3.2 Dataset and Data Collection

Using a consumer-grade wearable device (four-channel microphone array plus first-person camera), we collect approximately 60 hours across 74 sessions, covering six indoor and outdoor scenarios: three basic indoor events (knock, doorbell, smoke alarm), a reverberant indoor multi-event concurrent scenario, a high-noise subway platform scenario, and an open outdoor traffic scenario. The

total hardware cost is approximately 2000 RMB, using off-the-shelf components. Detailed hardware specifications are provided in § B.

3.3 Scenario and Question Design

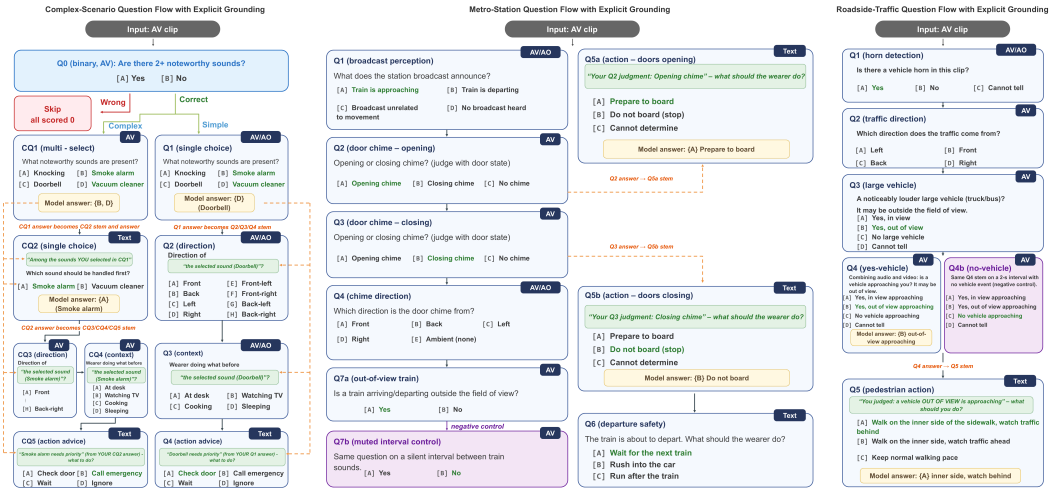

Figure 2: Question pipelines for the three scenario groups. **Left:** the complex multi-event scenario. **Middle:** the metro-station scenario. **Right:** the roadside-traffic scenario.

We use the complex multi-event scenario as an illustrative example; its pipeline (Figure 2) proceeds as Q0 (screening) → CQ1 (event enumeration) → CQ2 (urgency ranking) → behavioral decision. The metro and traffic scenarios follow analogous grounded pipelines (Figure 2, middle and right), while the three basic scenarios contain a single event, so no inter-event ranking is needed and priority is merged into the decision item. Full question templates are provided in § C.

3.4 Explicit Grounding and Stage Decoupling

To make the three stages independently measurable, we write upstream outcomes into downstream stems through two mechanisms that reduce error propagation across stages. **Premise externalization.** Downstream stems state the upstream conclusion as a given premise, so each item evaluates only its own stage; for instance, a decision item asks what to do once the event is already given as a knock. **Dynamic branching.** In the multi-event questionnaire, ranking options and later stems are built at runtime from the events the model itself selected, via placeholders such as {`q2_sound`}, with the correct answer generated on the corresponding branch. Downstream scoring thus always follows the model's own upstream state rather than failing by association. We additionally retain a full end-to-end score: if Q0 screening is incorrect, all subsequent items are skipped and scored zero, measuring the cascade's end-to-end capability.

3.5 Evaluation Protocol

We report five measures: **strict exact match**, **partial-direction score**, **conditional accuracy** (only downstream items actually reached), **gold-premise accuracy** (the correct event pair supplied as the question premise), and a **human baseline** from 10 participants on the same task. Figure 2 illustrates the task structure; dataset statistics are summarized in § D, and detailed scoring rules and edge cases in § F.

4 Experimental Setup

Evaluated Models. We evaluate three representative multimodal large language models (**Qwen2.5-Omni-7B** (Xu et al., 2025), **Gemini-2.5-Pro** (Gemini Team, 2025), and **VideoLLaMA2.1-7B-AV** (Cheng et al., 2024b)). Qwen and VideoLLaMA are deployed locally (bfloat16); Gemini is

invoked through its official API. All three models are evaluated on the same question sets and clips under the same scoring rules; the only differences are model-specific inference wrappers and audio preprocessing. We feed the same four-channel raw microphone audio to every model; how each consumes this input differs internally—Qwen's official inference pipeline downmixes to mono, while Gemini receives the raw channels natively. This reflects each model's own interface configuration rather than an input distinction introduced by us.

Input Conditions. Each question is answered under controlled input conditions: audiovisual (AV), audio-only (AO), video-only (V), and text-only (T), where applicable, to quantify how each input channel affects accuracy.

Metrics. We use the following measures:

- **Total score rate.** Across any scope (scenario, capability level, or modality), the score is the ratio of earned points to total points (earned/total), rather than the mean of per-question accuracies. A single question is considered correct only if it earns a full score of 1.0; any partial score below 1.0 is counted as an error.
- **Direction hit rate.** For eight-way direction questions, if the ground truth is a diagonal direction, answering either of its two component cardinal directions receives 0.5; all other answers are scored by exact match. The random baselines are 12.5% for strict exact match and 18.75% for the partial-direction score.
- **Modality effect.** AV, AO, and T results are reported separately to quantify how each modality affects perception accuracy.
- **Negative-control false alarm rate.** For items such as the traffic "no vehicle approaching" control, we report the proportion of cases in which the model incorrectly reports an event.

Prompt and Decoding. Models use default decoding settings. The complete prompt templates are provided in § C.

5 Results and Analysis

5.1 Overall Benchmark Performance

Figure 3 shows the overall score rate across the six scenarios. Gemini-2.5-Pro attains the highest and most consistent scores across all scenarios, the complex multi-event scenario is consistently the weakest for all three models, and performance varies substantially across the noisy traffic and subway settings. Exact per-scenario values are annotated in Figure 3.

5.2 Performance Along the Event Cascade

Along the perception–priority–decision chain, the three models show a consistent layered pattern (Figure 4).

Perception is the common lowest layer (Qwen 48.4%, Gemini 65.1%, VideoLLaMA 42.2%): spatial direction is near random for all three models, while the other perception sub-dimensions vary widely across CQ1, subway, and traffic (Figure 4).

Priority is the strongest layer (93.6%–100%). **Decision** is intermediate (55.0%–99.0%), punctuated by an extreme traffic failure in which Qwen answered all 298 items incorrectly (0%), traced to option-order anchoring (§ 5.5).

5.3 Why Perception Is the Primary Bottleneck

Three pieces of evidence form a complete attribution chain (Table 1):

- **End-to-end:** complex-scenario scores are 27.1%–62.8%.
- **Conditional:** when only items actually reached are counted, scores rise to 69.0%–79.2%.

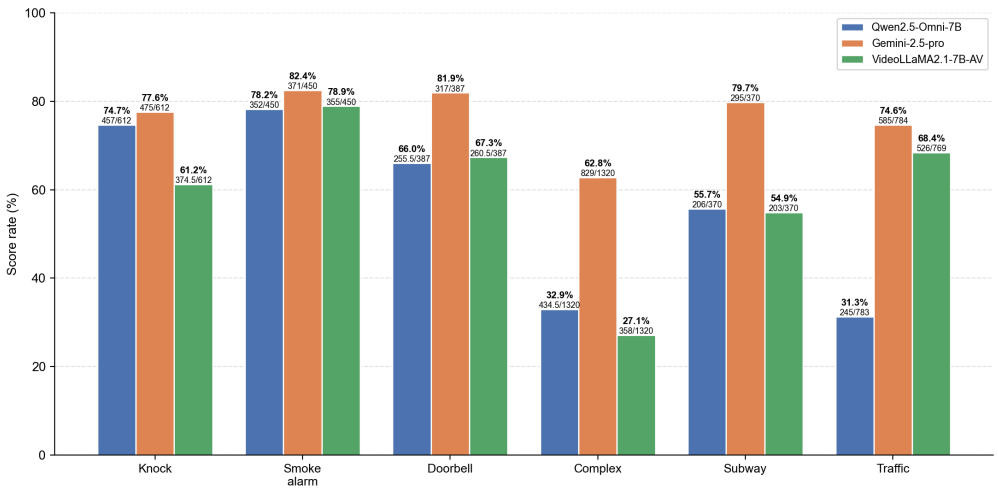

Figure 3: Overall score rate of the three models across six scenarios. Complex bars include the supplementary Q0/CQ1 perception retests (360 pairs); end-to-end-only complex scores are reported in Table 1. Each bar is annotated with the score rate (top) and earned/total (bottom).

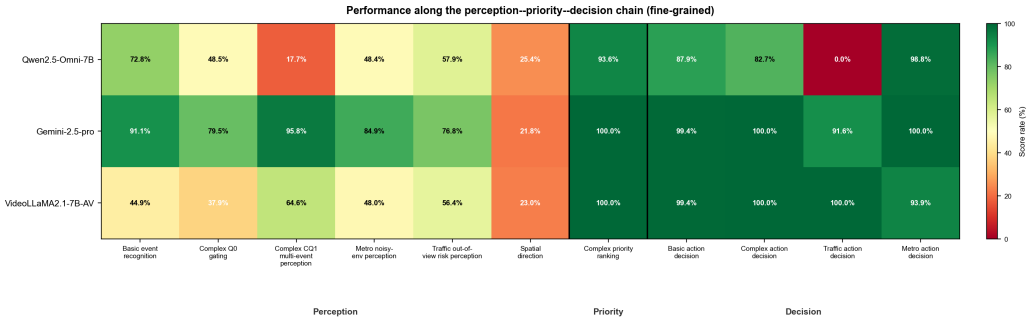

Figure 4: Score rate of the three models across 11 fine-grained dimensions along the perception–priority–decision chain (6 perception, 1 priority, 4 decision).

- **Gold premise:** when the correct event pair is supplied directly as a text premise, all three models answer all 96 closed-form ranking and action questions correctly (100%).

This three-tier ladder shows that once the correct event pair is given, ranking and decision-making are essentially solved *for the current closed-form, explicitly grounded tasks*; this does not establish that open-ended planning, long-horizon interaction, or real-world action selection is solved. The $\sim$40-point gap between end-to-end and conditional scores is largely due to Q0 gating: a wrong screening answer skips all follow-up items and scores zero. Within the evaluated tasks, perception (especially front-end event screening and grounding) accounts for the largest observed loss in end-to-end performance.

Table 1: End-to-end vs. conditional vs. gold-premise accuracy on the complex multi-event scenario. Conditional accuracy restricts clips to those where Q0 was answered correctly; gold-premise supplies the ground-truth event pair as text (no audio/video; all three models).

Model	End-to-end	Conditional	Gold premise
Qwen2.5-Omni-7B	27.1%	71.4%	100%
Gemini-2.5-pro	62.8%	79.2%	100%
VideoLLaMA2.1-7B-AV	27.1%	69.0%	100%

5.4 Systematic Perception Failures

5.4.1 Weak and Ambiguous Events

Weak events missed under high noise. At the recorded subway station, the door-opening chime is objectively much weaker than the door-closing chime, and detection rates diverge sharply (§ G): Qwen detects only 3.3% of opening chimes against 100% of closing chimes; Gemini 40.0% vs. 100%; VideoLLaMA 43.3% vs. 55.6%. When the weak opening chime is inaudible, Qwen often answers "the train doors are closed," falling back on visual cues rather than audio—under high noise, weak acoustic events are not reliably detected.

Asymmetric category confusion. All three models recognize the smoke alarm 100% of the time and never misclassify it downward, yet transient knocking and doorbell sounds are frequently upgraded to the louder, more salient smoke alarm—VideoLLaMA misclassifies 116/136 knocks as smoke alarms (§ G). Together with the question-only results (§ 5.5), this points to a "smoke alarm is most urgent" prior as a partial driver of the bias. Since we do not independently manipulate loudness, duration, reverberation, or masking, we treat acoustic salience as a correlate rather than an established causal factor.

Model-specific structural failures. The three models fail in structurally different ways. Qwen mainly over-selects: in 67 of its 81 CQ1 errors it selects all three options, even though select-all earns zero credit. VideoLLaMA shows the opposite tendency: across the 48 AV complex items it never selects knocking (0/48), typically answering "smoke alarm + doorbell" or "smoke alarm only," and it recognizes only 5.1% of knocking events even in the basic scenario—a systematic cross-scenario omission of knocking rather than conservative under-selection. Gemini is substantially more stable, with only 15 errors and exactly the correct pair on 90.7% of CQ1 items.

5.4.2 Human Baseline and Negative Controls

We recruit 10 human participants on five sub-dimensions (about 700 questions; Table 2). Humans score 100% on basic event recognition and 100% on complex CQ1, confirming these items are not inherently unfair. On the weak subway opening chime, humans score 71.4%, indicating some intrinsic difficulty, but still outperform all three models (3.3%–43.3%). Details of test tool are in 8.

The traffic Q4b serves as a negative control: it presents vehicle-free intervals where the correct answer is "no approaching vehicle." Humans correctly stay silent 83.8% of the time, whereas all three models false-alarm frequently—Qwen 62/62 (0% correct), Gemini 61/62 (1.6%), VideoLLaMA 52/62 (16.1%). Models produce reasoning such as "although no vehicle is visible, traffic sound is audible, so a vehicle may be approaching out of view"—mistaking continuous ambient road noise for an approaching vehicle. Together, these results show that model failures are not driven by task difficulty but by model-specific deficiencies—especially a cross-model inability to withhold a response when nothing noteworthy occurs.

Table 2: Human baseline versus three MLLMs (AV condition, score rate %).

Dimension	Human	Qwen	Gemini	VideoLLaMA
Basic event recognition	100	46.5–57.4	96.3–97.3	0.0–23.3
Complex CQ1	100	17.7	95.8	64.6
Subway opening chime	71.4	3.3	40.0	43.3
Negative control (no car)	83.8	0.0	1.6	16.1

5.5 Controlled Failure Analysis

Prior and Shortcut Controls: To rule out language shortcuts, we run three controls. **Question-only event recognition.** Without audio or video, both open-weight models fall *below* the 25% random baseline on knocking and doorbell, proving the question text contains no shortcut; in contrast, smoke-alarm questions are answered at 86%–100% *without any audio*, revealing a strong smoke-alarm prior. **Q0 yes-bias.** Under the question-only condition, both open-weight models answer "yes" to *all* 84 simple negative clips (0% accuracy); since option order is randomized and the

"yes" answers split roughly evenly across the two option letters, this is a semantic failure to remain silent rather than a position bias toward a fixed letter. **Blind-majority prior on CQ1.** All 48 complex clips contain the smoke alarm (paired with knocking in 33 and the doorbell in 15), so a blind strategy that always answers "smoke alarm + knocking" without listening earns 84.4% partial credit (68.8% exact match). Only Gemini-2.5-Pro (95.8%) exceeds this blind baseline, providing stronger evidence of genuine multi-event perception; VideoLLaMA2.1-7B-AV (64.6%) falls below it, while Qwen2.5-Omni-7B (17.7%) reflects systematic over-selection rather than prior-driven guessing. We use this prior as the reference for CQ1 and report the imbalance as a limitation.

Grounded Decision Anchoring: When the correct event premise is supplied, downstream ranking and decision-making are substantially more reliable: complex CQ2 reaches 93.6%–100%, and CQ5 behavioral decisions are 100% for Gemini and VideoLLaMA, 82.7% for Qwen. This near-ceiling performance is possible because explicit grounding writes the upstream conclusion directly into the question stem, so the downstream task no longer depends on the model's own perception. **Position anchoring.** The traffic Q5 in the main run used a fixed option order, and Qwen selected the same position on all 298 questions (0% accuracy). A shuffled-option control—identical stems, still text-only—recovers Qwen to 76.6% (229/299), confirming that the original failure was option-order anchoring rather than decision incompetence. VideoLLaMA is largely unaffected (79.2% $\rightarrow$ 79.6%).

Spatial Perception and DOA Intervention: On eight-direction questions, all three models score only 21.9%–23.5% partial-direction points (random expectation 18.75%) and 8.0%–12.1% strict exact match (random 12.5%), with all eight directions near chance and no systematic front–back or left–right reversal (per-direction breakdown in § G). To decompose whether the failure originates in signal-level sensing or MLLM reasoning, we run a four-layer intervention on indoor doorbell direction estimation (Table 3, 132 clips per condition):

Table 3: DOA–MLLM intervention on indoor doorbell eight-direction localization (132 clips per condition). Pure DOA is signal-level only; all other rows feed the direction estimate to the MLLM as a text hint. AV and AO differ minimally; direction is a purely auditory cue.

Method	Qwen	Gemini	VideoLLaMA
Pure DOA (signal-level, no MLLM)	53.0%	53.0%	53.0%
MLLM alone, AV	14.4%	7.6%	9.8%
MLLM alone, AO	12.9%	12.9%	12.9%
DOA $\rightarrow$ MLLM, AV	51.5%	50.0%	46.2%
DOA $\rightarrow$ MLLM, AO	53.0%	44.7%	47.0%
Oracle DOA $\rightarrow$ MLLM, AV	87.1%	87.1%	81.8%
Oracle DOA $\rightarrow$ MLLM, AO	94.7%	92.4%	85.6%

Pure DOA alone achieves only 53% in the reverberant indoor room (vs. 96% outdoors), confirming that the signal-level estimator itself degrades under reverberation. MLLMs alone cannot localize from raw streams (9.8%–14.4% AV, 12.9% AO), near the 12.5% chance level. Feeding the estimated DOA as a text hint raises accuracy to 46.2%–53.0%, essentially matching pure DOA, and replacing it with the Oracle (ground-truth) direction reaches 81.8%–94.7%. The bottleneck is thus acquiring a reliable spatial cue from raw input: once the correct direction is supplied, the MLLM uses it well.

These results do *not* suggest replacing learned spatial understanding with an external DOA module. Instead, they reveal a gap between raw sensory input and semantic–spatial grounding: an external DOA cue substantially improves performance, and Oracle DOA exposes considerable additional headroom. However, a standalone DOA estimator does not identify which semantic event produced a sound, nor does it resolve competing events under strong noise. For an embodied assistant, spatial perception therefore cannot be reduced to estimating the direction of the strongest sound; the relevant capability is to associate an event's acoustic identity and semantic meaning with its corresponding spatial location.

5.6 Modality Effects

Vision is not uniformly helpful. On basic event recognition, audio-only (AO) consistently outperforms audiovisual (AV) by 10–17.5 points, suggesting that vision often introduces interfering priors;

yet vision helps in some multi-event cases—on the traffic, it allows VideoLLaMA to confirm a second concurrent source (AV 92.6% vs. AO 63.0% on Q1). The direction of the modality effect is thus model- and task-dependent rather than uniform (Figure 5). Beyond aggregate scores, AV and AO frequently produce different answers for the same clip, indicating that cross-modal grounding is not stable even when the underlying recording is identical.

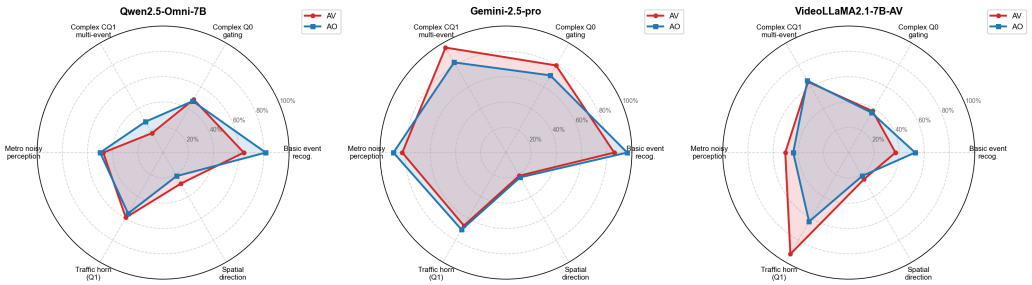

Figure 5: Perception-layer performance across six dimensions: AV (red) vs AO (blue), for the three evaluated models.

The resulting failure modes are model-specific, as summarized in § G.

6 Limitations

Scenario and Task Coverage: EgoSPECT currently comprises six scenarios, about 60 hours, and 74 sessions, covering indoor home safety (knocking, doorbell, smoke alarm), a reverberant indoor multi-event setting, a noisy subway platform, and outdoor traffic; they do not cover every real-world environment. The task format is primarily multiple-choice, and open-ended generation or long-horizon planning tasks are not yet included.

Spatial Localization and Audio Interface: Multi-source semantic–spatial disentanglement—localizing the direction of each concurrent sound separately—remains an open challenge.

Dataset Bias and Validation Scope: All 48 complex-scenario clips contain the smoke alarm, and the doorbell-plus-smoke-alarm combination is less frequent than the knock-plus-smoke-alarm combination, inflating the blind majority prior on CQ1 (84.4%). A more balanced event-composition distribution is planned for the next release.

7 Conclusion

We present **EgoSPECT**, an egocentric audio-visual benchmark that organizes evaluation along the full perception–priority–decision chain rather than isolating sound localization. Collected on a consumer-grade headset across roughly 60 hours, 74 sessions, and six indoor and outdoor scenarios, EgoSPECT evaluates three MLLMs under audiovisual, audio-only, video-only, and text-only conditions. The main difficulty of current MLLMs lies in raw sensory perception and grounding—jointly understanding spatial direction and event semantics, and abstaining when nothing noteworthy occurs. When the correct event is supplied as a text premise, downstream ranking and decision-making approach ceiling, showing that the perception front-end dominates end-to-end loss. Failures are systematic and bidirectional: weak-event misses, category confusion, over-selection, and false alarms on negative controls, with residual decision failures such as option-order anchoring persisting even after grounding. DOA interventions confirm that explicit spatial cues substantially improve directional accuracy, and the required capability is not locating the loudest sound but associating each semantic event with its spatial location; adding vision is not uniformly beneficial and remains model- and task-dependent. These findings have direct implications for *humanoid and embodied AGI*, which must bind event semantics to spatial location before acting, making semantic-spatial event understanding under concurrent and noisy conditions a prerequisite for future embodied assistants. Extending EgoSPECT to commercial AI glasses and to low-reverberation open-space direction control is an immediate next step.

AI USE STATEMENT

In this work, we used generative AI tools solely for language polishing of the manuscript (e.g., grammar correction and phrasing improvements). All research ideas, experimental designs, code, and the original manuscript draft were produced entirely by the authors. We have reviewed all AI-assisted text and take full responsibility for the final content of this work.

ETHICS STATEMENT

All publicly available datasets used in this work either contain no private information or have been anonymized with the informed consent of the participants. Data collected by the authors that has not yet been released is used under controlled, research-only conditions. We are actively anonymizing this data and will release it publicly once all privacy concerns have been properly addressed.

REPRODUCIBILITY STATEMENT

We are committed to ensuring the reproducibility of our work. An anonymized version of our code is available at `https://anonymous.4open.science/r/EgoSPECT-D346/`. The dataset will be publicly released upon publication, following the completion of the privacy processing described in the Ethics Statement.

REFERENCES

Eun Chang, Zhuangqun Huang, Yiwei Liao, Sagar Bhavsar, Amogh Param, Tammy Stark, Adel Ahmadyan, Xiao Yang, Jiaqi Wang, Ahsan Abdullah, et al. Wearvqa: A visual question answering benchmark for wearables in egocentric authentic real-world scenarios. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Changan Chen, Unnat Jain, Carl Schissler, Sebastia Vicenc Amengual Gari, Ziad Al-Halah, Vamsi Krishna Ithapu, Philip Robinson, and Kristen Grauman. Soundspaces: Audio-visual navigation in 3d environments. In *European Conference on Computer Vision (ECCV)*, 2020.

Changan Chen, Carl Schissler, Sanchit Garg, Philip Kobernik, Alexander Clegg, Paul Calamia, Dhruv Batra, Philip Robinson, and Kristen Grauman. Soundspaces 2.0: A simulation platform for visual-acoustic learning. *Transactions on Machine Learning Research*, 2023.

Mingfei Chen, Zijun Cui, Xiulong Liu, Jinlin Xiang, Caleb Zheng, Jingyuan Li, and Eli Shlizerman. Savvy: Spatial awareness via audio-visual llms through seeing and hearing. *arXiv preprint arXiv:2506.05414*, 2025.

Sijie Cheng, Zhicheng Guo, Jingwen Wu, Kechen Fang, Peng Li, Huaping Liu, and Yang Liu. Egothink: Evaluating first-person perspective thinking capability of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14291–14302, 2024a.

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and Lidong Bing. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024b.

Samuel Clarke, Ruohan Gao, Mason Wang, Mark Rau, Julia Xu, Jui-Hsien Wang, Doug L. James, and Jiajun Wu. Realimpact: A dataset of impact sound fields for real objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1516–1525, 2023.

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

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision*, 130:33–55, 2022.

Jacob Donley, Vladimir Tourbabin, Jung-Suk Lee, Mark Broyles, Hao Jiang, Jie Shen, Maja Pantic, Vamsi Krishna Ithapu, and Ravish Mehra. Easycom: An augmented reality dataset to support algorithms for easy communication in noisy environments. *arXiv preprint arXiv:2107.04174*, 2021.

Gemini Team. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Vincent Cartillier, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Devansh Kukreja, et al. Ego4d: Around the world in 3,600 hours of egocentric video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(11):9468–9509, 2025.

Wenxiang Guo, Changhao Pan, Zhiyuan Zhu, Xintong Hu, Yu Zhang, Li Tang, Rui Yang, Han Wang, Zongbao Zhang, Yuhan Wang, Yixuan Chen, Hankun Xu, Ke Xu, Pengfei Fan, Zhetai Chen, Yanhao Yu, Qianye Huang, Fei Wu, and Zhou Zhao. Mrsaudio: A large-scale multimodal recorded spatial audio dataset with refined annotations. *arXiv preprint arXiv:2510.10396*, 2025.

Zhuohang Jiang, Xu Yuan, Haohao Qu, Shanru Lin, Kanglong Liu, Wenqi Fan, and Qing Li. Superglasses: Benchmarking vision language models as intelligent agents for ai smart glasses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Yijia Lei, Jinzhao Li, Yichi Zhang, Jiacheng Hua, Yin Li, and Miao Liu. Egosat: A comprehensive benchmark of egocentric streaming interaction understanding. *arXiv preprint arXiv:2606.24422*, 2026.

Chuhan Li, Rilyn Han, Joy Hsu, Yongyuan Liang, Rajiv Dhawan, Jiajun Wu, Ming-Hsuan Yang, and Xin Eric Wang. Saw-bench: Learning situated awareness in the real world. *arXiv preprint arXiv:2602.16682*, 2026.

Guangyao Li, Yake Wei, Yapeng Tian, Chenliang Xu, Ji-Rong Wen, and Di Hu. Learning to answer questions in dynamic audio-visual scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Zhaoyang Lv, Nicholas Charron, Pierre Moulon, Alexander Gamino, Cheng Peng, Chris Sweeney, Edward Miller, Huixuan Tang, Jeff Meissner, Jing Dong, Kiran Somasundaram, Luis Pesqueira, Mark Schwesinger, Omkar M. Parkhi, Qiao Gu, Renzo De Nardi, Shangyi Cheng, Steve Saarinen, Vijay Baiyya, Yuyang Zou, et al. Aria everyday activities dataset. *arXiv preprint arXiv:2402.13349*, 2024.

Curtis G. Northcutt, Shengxin Zha, Steven Lovegrove, and Richard Newcombe. Egocom: A multi-person multi-modal egocentric communications dataset. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6783–6793, 2022.

S. Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. Mmau: A massive multi-task audio understanding and reasoning benchmark. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025.

Ashish Seth, Xinhao Mei, Changsheng Zhao, Varun Nagaraja, Ernie Chang, Gregory P. Meyer, Gael Le Lan, Yunyang Xiong, Vikas Chandra, Yangyang Shi, Dinesh Manocha, and Zhipeng Cai. Egoavu: Egocentric audio-visual understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Zhanbo Shi, Lin Zhang, Linfei Li, and Ying Shen. Towards audio-visual navigation in noisy environments: A large-scale benchmark dataset and an architecture considering multiple sound-sources. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 14673–14680, 2025.

Kazuki Shimada, Archontis Politis, Parthasaarathy Sudarsanam, Daniel Krause, Kengo Uchida, Sharath Adavanne, Aapo Hakala, Yuichiro Koyama, Naoya Takahashi, Shusuke Takahashi, Tuomas Virtanen, and Yuki Mitsufuji. Starss23: An audio-visual dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2023.

Vijay Veerabadran, Fanyi Xiao, Nitin Kamra, Pedro Matias, Joy Chen, Caley Drooff, Brett D. Roads, Riley Williams, Ethan Henderson, Xuanyi Zhao, Kevin Carlberg, Joseph Tighe, and Karl Ridgeway. Benchmarking egocentric multimodal goal inference for assistive wearable agents. *arXiv preprint arXiv:2510.22443*, 2025.

Bin Wang, Xunlong Zou, Geyu Lin, Shuo Sun, Zhuohan Liu, Wenyu Zhang, Zhengyuan Liu, AiTi Aw, and Nancy Chen. Audiobench: A universal benchmark for audio large language models. In *Proceedings of NAACL-HLT*, pp. 4297–4316, 2025.

Mason Wang, Samuel Clarke, Jui-Hsien Wang, Ruohan Gao, and Jiajun Wu. Soundcam: A dataset for finding humans using room acoustics. *arXiv preprint arXiv:2311.03517*, 2023a.

Xin Wang, Taein Kwon, Mahdi Rad, Bowen Pan, Ishani Chakraborty, Sean Andrist, Dan Bohus, Ashley Feniello, Bugra Tekin, Felipe Vieira Frujeri, et al. Holoassist: An egocentric human interaction dataset for interactive ai assistants in the real world. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20270–20281, 2023b.

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. *arXiv preprint arXiv:2503.08221*, 2025.

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

Jingkang Yang, Shuai Liu, Hongming Guo, Yuhao Dong, Xiamengwei Zhang, Sicheng Zhang, Pengyun Wang, Zitang Zhou, Binzhu Xie, Ziyue Wang, et al. Egolife: Towards egocentric life assistant. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Pinci Yang, Xin Wang, Xuguang Duan, Hong Chen, Runze Hou, Cong Jin, and Wenwu Zhu. Avqa: A dataset for audio-visual question answering on videos. In *Proceedings of the 30th ACM International Conference on Multimedia*, 2022.

Hanrong Ye, Haotian Zhang, Erik Daxberger, Lin Chen, Zongyu Lin, Yanghao Li, Bowen Zhang, Haoxuan You, Dan Xu, Zhe Gan, et al. Mm-ego: Towards building egocentric multimodal llms. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025.

Bingwen Zhu, Yuqian Fu, Qiaole Dong, Guolei Sun, Tianwen Qian, Yuzheng Wu, Danda Pani Paudel, Yanwei Fu, and Xiangyang Xue. Egosound: Benchmarking sound understanding in egocentric videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Appendices

Content

A Extended Comparison with Related Work **14**

B Hardware BOM and Assembly **14**

C Prompt Templates **15**

D Dataset Card **15**

E Ethics and Data Release **17**

F Additional Evaluation Details **17**

G Additional Results **18**

H Qualitative Case Studies **18**

I Ongoing Extensions **20**

A Extended Comparison with Related Work

Table 4 extends §2 with a dimension-by-dimension comparison against additional egocentric, spatial-audio, and audio-visual benchmarks.

Table 4: Comparison with related egocentric audio-visual benchmarks. "partial" denotes a capability present only in limited form.

	Ego centric	Spat. dir.	Event sem.	Multi-ev. priority	Action decision	Neg. controls	Stage-wise diagnosis
EgoSound (Zhu et al., 2026)	✓	✓	✓	—	—	—	—
EgoAVU (Seth et al., 2026)	✓	partial	✓	—	—	partial	—
SAVVY-Bench (Chen et al., 2025)	✓	✓	partial	—	—	—	—
WAGIBench (Veerabadran et al., 2025)	✓	—	—	—	—	—	—
SoundSpaces (Chen et al., 2020)	partial	✓	—	—	partial	—	—
SoundSpaces 2.0 (Chen et al., 2023)	partial	✓	—	—	partial	—	—
BeDAViN (Shi et al., 2025)	partial	✓	partial	partial	partial	—	—
STARSS23 (Shimada et al., 2023)	—	✓	✓	partial	—	—	—
RealImpact (Clarke et al., 2023)	—	partial	partial	—	—	—	—
SoundCam (Wang et al., 2023a)	—	✓	partial	—	—	—	—
EgoCom (Northcutt et al., 2022)	✓	partial	partial	—	—	—	—
EasyCom (Donley et al., 2021)	✓	partial	partial	—	—	—	—
Ego4D (Grauman et al., 2025)	✓	partial	✓	—	—	—	—
EPIC-KITCHENS (Damen et al., 2018; 2022)	✓	—	✓	—	—	—	—
HoloAssist (Wang et al., 2023b)	✓	—	✓	—	partial	—	—
EgoLife (Yang et al., 2025)	✓	—	partial	—	partial	—	—
EgoThink (Cheng et al., 2024a)	✓	—	partial	—	partial	—	—
MM-Ego (Ye et al., 2025)	✓	—	partial	—	—	—	—
EgoBlind (Xiao et al., 2025)	✓	—	partial	—	partial	—	—
WearVQA (Chang et al., 2025)	✓	—	—	—	—	—	—
EgoSAT (Lei et al., 2026)	✓	—	partial	—	partial	—	—
SuperGLASSES (Jiang et al., 2026)	✓	—	—	—	partial	—	—
SAW-Bench (Li et al., 2026)	✓	partial	partial	—	partial	—	—
MUSIC-AVQA (Li et al., 2022)	—	partial	✓	—	—	—	—
AVQA (Yang et al., 2022)	—	partial	✓	—	—	—	—
MMAU (Sakshi et al., 2025)	—	—	✓	—	partial	—	—
AudioBench (Wang et al., 2025)	—	—	✓	—	—	—	—
EgoSPECT (Ours)	✓	✓	✓	✓	✓	✓	✓

B Hardware BOM and Assembly

The acquisition headset is assembled entirely from off-the-shelf components at a total cost of about 2,000 CNY. The main controller is a Raspberry Pi 4B. The audio front end is an XVF3800 four-microphone array that outputs six-channel, 16 kHz, 16-bit PCM audio: channels 1–2 carry the device's processed reference/beamformed streams, while channels 3–6 carry the four raw microphone channels, which are the four channels supplied to every model. An AQ12 USB camera module (12 MP autofocus with a 4 mm, 80° distortion-free lens) captures first-person video, recorded at 1280×720 at approximately 15 fps. A BNO055 inertial measurement unit records head pose, but the IMU data are not used as model input or in scoring. Audio and video are aligned through unified hardware timestamps, with every video frame timestamped on the same time axis as the audio. The bill of materials is given in Table 5.

Table 5: Bill of materials for the acquisition headset.

Component	Model	Role
Main controller	Raspberry Pi 4B	Recording, synchronization, storage
Microphone array	XVF3800 4-mic board	Six-channel output; ch3–ch6 = four raw mic channels
Camera	AQ12 USB module	1280×720 first-person video, ≈15 fps
IMU	BNO055	Head-pose logging (not used in scoring)
Power	Power bank	Mobile power supply
Frame	Head-mount	Wearable mounting

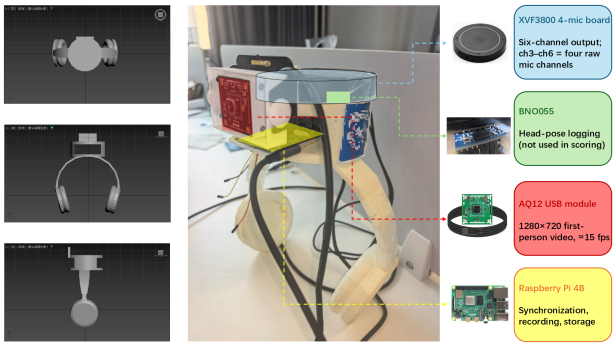

Figure 6: The acquisition headset, assembled from off-the-shelf components. (a) Worn on the head; (b) front view; (c) side view; (d) top view of the XVF3800 four-microphone array.

C Prompt Templates

All questions share a single request format that concatenates the question stem, an option list (options separated by two spaces), and a fixed answering instruction. Downstream items in a chain (Q4, CQ2, CQ5) receive *no* audio or video: the model answers only from the perception premise written into the stem (premise externalization; see §3).

Single-select instruction: First output the option letter (format: Answer: X), then give brief reasoning in 2–3 sentences. If you cannot decide, you must still guess the most likely option and output it.

Multi-select instruction: First output the option letters (format: Answer: X, Y, or Answer: X and Y), then give brief reasoning in 2–3 sentences. You must select two or more options. If you cannot decide, you must still guess the most likely options and output them.

Request format.

```
{stem}
A. {option 1}  B. {option 2}  C. {option 3} ...
{instruction}
```

Stage-specific stems. Table 6 gives the original Chinese stem and its English rendering for each chain stage; placeholders are filled at runtime with the upstream perception output.

Perceived-event to action mapping.

- Knock or doorbell → "go to the door and check."
- Smoke alarm → "evacuate immediately and check for smoke/fire."
- No abnormal sound → "no action needed; continue the current activity."

D Dataset Card

Item schema. Each question is stored as a JSON record carrying the session and clip identifiers, question id, modality, stem, option list, the gold option letter, and (where applicable) the ground-truth direction. A representative record is shown below.

```
{
  "session": "20260819_150114",
  "clip": "evt05",
  "qid": "Q1_perception",
  "modality": "av",
  "stem": "In the latter half of this first-person "
          "recording, what sound suddenly appears?",
```

Table 6: Original Chinese stems and English renderings by chain stage.

Stage	English rendering
Q0 gating	Are there two or more noteworthy sounds in this segment (e.g. smoke alarm, doorbell, knocking)? There may not be; ordinary background sounds such as games or TV can be ignored.
Q1 perception	In the latter half of this first-person recording, what sound suddenly appears in the environment?
Q2 direction	From which direction did that sound initially originate?
Q3 context	Just before that sudden sound occurred, what was the wearer doing?
CQ2 priority	Among the sounds you selected in Q1, which needs to be handled first?
Q4 decision	Your perception conclusion was: {perceived}. The wearer is working at a computer; as their assistant, what should you now advise them to do?
CQ5 decision	You judged that {perceived} is the highest priority; based on this, what do you most advise the wearer to do?
Traffic Q4	Combining the video and sound, is a vehicle approaching you, and from where? It may not be in the field of view.
Traffic Q5	You judged {...}; as a pedestrian on the sidewalk, what should you do now?

Table 7: EgoSPECT scenarios and question counts per model. *Event×type units* counts event-level units after multiplying by the number of question types per event; for example, the 68 knocking events yield 4 question types per event under one input condition (AV), giving $68\times4\approx272$ units before additional modality conditions. *Q–C pairs* counts clip–modality evaluations. AV: audiovisual; A: audio-only; V: video-only; T: text stem only.

Scenario	Category	Sess.	Event×type units	Q–C pairs	Conditions
Knocking	Basic	8	272	612	AV/A/V
Smoke alarm	Basic	8	200	450	AV/A/V
Doorbell	Basic	8	172	387	AV/A/V
Complex multi-event	Priority	20	708	1,320[†]	AV/A
Subway platform	Public	17	261	370	AV/A/T
Outdoor traffic	Outdoor	13	757	784[‡]	AV/A/T
Total (six core)		**74**	**2,370**	**3,923**	
Outdoor open-space control[⋆]	Control	in progress	—	—	dir.

[†]The complex scenario contains 1,320 question–condition evaluations: 960 end-to-end full-chain tests (a wrong Q0 skips follow-up questions and counts as 0), plus 360 independent Q0/CQ1 perception retests (AV+AO, perception only, no gating).
[‡]Each model is given 784 traffic questions; under extreme conditions a few questions did not receive complete answers and are excluded from the final score.
[⋆]Data collection in progress; matched direction questions in a low-reverberation open space.


```
  "options": {"A": "No abnormal sound",
              "B": "Knocking",
              "C": "Doorbell",
              "D": "Smoke alarm"},
  "gold": "C",
  "direction": "right"
}
```

Worked basic chain. For a single doorbell clip the model is gated by Q0 (gold *no*), then asked Q1 perception (gold *doorbell*), Q2 direction (one of eight), Q3 context (gold *working at a computer*), and the externalized Q4 decision (gold *go to the door and check*). The complex chain follows the gated flow in the task diagram of §3.3.

Clip segmentation. Each session is cut at event boundaries; the target event lies in the latter half of a clip and is preceded by ambient context, while negative-control clips contain ambient sound only. The 48 multi-event clips each contain a smoke alarm paired with either a knocking (33) or a doorbell (15) sound.

Annotation tool. Recordings are annotated with a mobile web tool (Fig. 7): the main screen provides a timeline with a MARK button and event labels, the Records view lists timestamped events with their direction, and the backend log aligns audio–video timestamps and muxes the streams via ffmpeg.

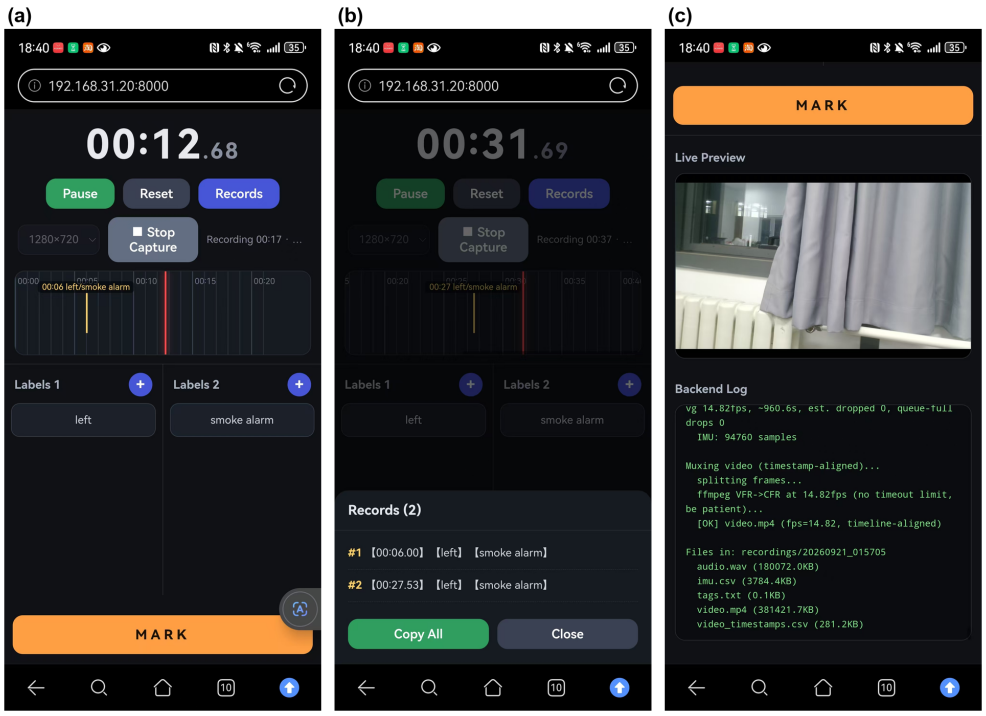

Figure 7: The annotation tool. (a) Main labeling screen with timeline and MARK; (b) Records list of timestamped events; (c) backend log for timestamp alignment and stream muxing.

E Ethics and Data Release

All recordings are collected in public spaces or in consented private settings. Every identifiable face is blurred before public release, and all recorded conversations were made with the involved participants' consent. Publicly released data contain no sensitive personal information or have undergone the corresponding privacy treatment; data not yet public will be released only after privacy processing is complete. The benchmark relies on passive observation of everyday scenes and does not intervene on participants, and therefore poses minimal risk.

F Additional Evaluation Details

Scoring rules. *Strict exact match:* an item scoring 1 is counted correct, otherwise incorrect. *Partial-direction score:* for eight-direction items, if the ground truth is a diagonal direction, an answer correctly identifying either of its two adjacent cardinal directions receives 0.5; all other answers

are scored by exact match. *Conditional accuracy:* only downstream items actually reached are counted. *Gold-premise accuracy:* the correct event pair is supplied directly as the question premise. **Premise externalization (worked example).** The decision item reads, "Suppose you have already determined that the sound is a knock; what should you advise the wearer to do?" This item evaluates decision-making alone, regardless of whether the model perceived the event correctly.

Answer-conditioned (dynamic) stems. Ranking options are built at runtime from the sounds the model selected in the multi-select item; subsequent direction, detail, and decision items reference a placeholder (e.g., {`q2_sound`}) for the chosen sound, with the correct answer generated on that branch. Even when upstream recognition is wrong, the downstream level is scored fairly under the situation the model believes it is in, rather than being marked wrong by association.

Human evaluation interface. The human baseline is collected with a dedicated tool (Fig. 8): a landing page for selecting the stimulus folder, a setup page for the question bank and participant ID, a stimulus playback screen, a post-playback perception item, and a results summary that logs every answer. Ten participants answered approximately 700 questions in total (§5.5).

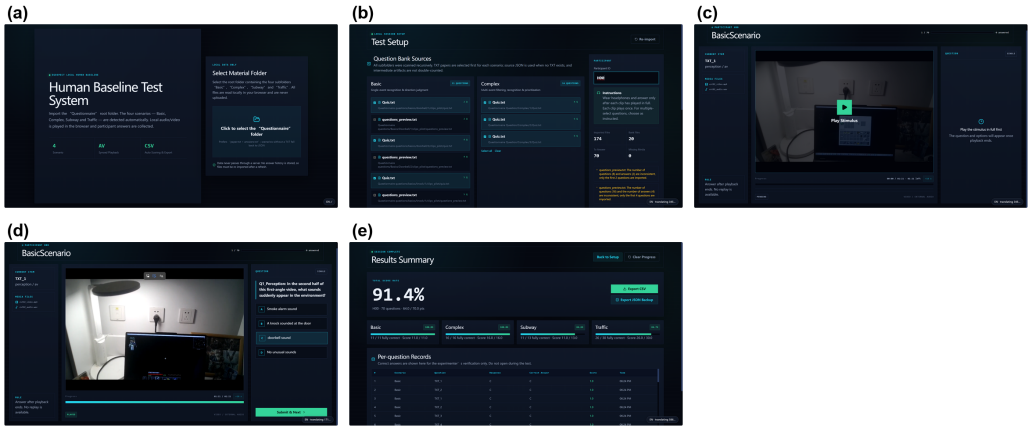

Figure 8: The human-evaluation interface. (a) Landing/folder selection; (b) test setup; (c) stimulus playback; (d) post-playback perception item; (e) results summary.

G Additional Results

Table 8: Detection of door-opening vs. door-closing chimes in the noisy subway scenario. At the recorded station the opening chime is substantially weaker than the closing chime.

Model	Opening chime (weak)	Closing chime (loud)
Qwen2.5-Omni-7B	1/30 (3.3%)	30/30 (100%)
Gemini-2.5-pro	12/30 (40.0%)	27/27 (100%)
VideoLLaMA2.1-7B-AV	13/30 (43.3%)	15/27 (55.6%)

H Qualitative Case Studies

To illustrate the quantitative results, we present representative successes and failures. Each case gives the session/clip, the item, the model's verbatim answer (translated from Chinese), the gold answer, and an error tag.

Case 1 (grounded downstream success). When the ground-truth event pair is supplied as a text premise, both open-weight models answer all 96 closed-form ranking and action items correctly, e.g. "a smoke alarm outranks a doorbell; advise the wearer to evacuate." This isolates downstream competence from perception and shows that, on the current closed-form grounded tasks, performance is near ceiling.

Table 9: Model-specific failure modes. Each cell names the dominant error type and the supporting statistic from §5.2– §5.6.

	Perception errors	Negative-control failures	Decision errors
Qwen2.5-Omni-7B	Doorbell 47.7% (18/86 misheard as smoke alarm); opening chime 3.3%; CQ1 *over-selection* (67/81 errors pick all three); audio-only hallucinates absent animals	Q0 simple clips 136/168 false alarms; question-only Q0 on simple clips 0/84 (yes-bias); Q4b 62/62 (reports vehicle approaching when none present)	Traffic Q5 *position anchoring*: 0/298 under fixed option order, recovering to 76.6% with shuffled options
Gemini-2.5-pro	Opening chime 40%; knock confused with smoke alarm 17/136; otherwise most stable (only 15 CQ1 errors; the only model above the blind CQ1 prior)	Q4b 61/62; Q0 simple clips 59/168	Strongest overall; only 25/299 traffic Q5 errors
VideoLLaMA2.1-7B-AV	Knock upgraded to smoke alarm 116/136; doorbell 44.2% (45/86 as smoke); CQ1 *category confusion*: never selects knocking (0/48), mislabels it as doorbell; second-source identification 31.3% (below the 68.8% blind guess)	Q0 simple clips 164/168 (worst); Q4b 52/62	Traffic Q5 284/284 correct (79.6% under shuffled options)

Table 10: Basic-scenario Q1 event-recognition accuracy and main confusions (true event per row). Transient knocking and doorbell sounds are frequently upgraded to the louder, more salient smoke alarm.

Model	True event	Accuracy	Main confusion
Qwen2.5-Omni-7B	Knocking	105/136 (77.2%)	30 as smoke alarm
	Doorbell	41/86 (47.7%)	26 as knocking; 18 as smoke alarm
	Smoke alarm	100/100 (100%)	—
Gemini-2.5-pro	Knocking	118/136 (86.8%)	17 as smoke alarm
	Doorbell	82/86 (95.3%)	4 as smoke alarm
	Smoke alarm	100/100 (100%)	—
VideoLLaMA2.1-7B-AV	Knocking	7/136 (5.1%)	116 as smoke alarm
	Doorbell	38/86 (44.2%)	45 as smoke alarm
	Smoke alarm	100/100 (100%)	—

Case 2 (gating false alarm; Qwen, 20260819_150114/evt01). On a single-event clip the model answered Q0 "yes": "I can hear a doorbell and knocking, indicating multiple noteworthy sounds." Gold: *no*. The stem explicitly names the salient events and states that ordinary game/TV audio can be ignored, yet the model still reports multiple hazards, failing to suppress an alarm tendency.

Case 3 (category confusion; Qwen, 20260819_150114/evt05). Asked what sound appears, the model answered *smoke alarm*: "a smoke-alarm-like alarm suddenly appears, possibly triggered by smoke or fire." Gold: *doorbell*. Transient knocking and doorbell sounds are repeatedly upgraded to the louder smoke alarm; VideoLLaMA2 shows the same pattern at scale, misclassifying 116/136 knocks as smoke alarms (Table 10).

Case 4 (over-selection; Qwen CQ1). On the multi-select perception item, 67 of Qwen's 81 errors select all three events regardless of the clip, inflating the reported event set. Tag: *over-selection*.

Case 5 (omission; VideoLLaMA2 CQ1). VideoLLaMA2 never selects knocking on the multi-select item (0/48), absorbing it into other categories. Tag: *omission / category confusion*.

Case 6 (negative-control false alarm; Qwen traffic, 20260831_103825). On clips where no vehicle is approaching either in or out of view, the model answered Q4b "a vehicle is approaching from outside the field of view": "there is no sign of an approaching vehicle in the video, but I hear a car passing, suggesting one may approach from outside the view." Qwen answered this way on 62/62 such clips. Tag: *inability to remain silent on a negative control*.

Case 7 (option-position anchoring; Qwen traffic Q5). Under the fixed option order Qwen scored 0/298 on the decision item, effectively anchoring to one option position; with shuffled options in the

text-only condition it recovered to 76.6%. The reasoning itself is plausible, but the answer collapses to a position, evidencing stability failures under dense, noisy conditions. Tag: *position anchoring / instability*.

Case 8 (weak-sound miss in noise; subway). Qwen detected the objectively weaker opening chime at 3.3% versus 100% for the louder closing chime (Table 8). In the high-noise station, weaker events fall below the model's effective sensitivity. Tag: *weak-sound sensitivity under high noise*.

I ONGOING EXTENSIONS

- Low-reverberation open-space direction control (data collection in progress).
- Self-voice / own-speech perception (testing in progress).