EGOMM: TOWARDS MULTIMODAL EGOCENTRIC INTELLIGENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Wearables capture synchronized video, audio, and inertial signals, yet existing egocentric understanding systems remain largely centered on vision. When cameras are occluded, poorly lit, or directed away from an event, complementary evidence can remain in sound and motion. Learning from this evidence requires models that integrate wearable signals and evaluations that distinguish their contributions. We present EgoMM, connecting benchmarking, multimodal learning, and long-horizon memory around this goal. We establish EgoMMU, to our knowledge the first benchmark to assess egocentric language models across visual, acoustic, and inertial tasks with explicit modality requirements. Its five task families form thirteen task-modality combinations, supported by construction-time checks and controlled ablations. We then develop EgoMind, which connects all three modalities to a vision-language backbone through staged sensor pretraining and modality-matched instruction tuning. At each of four model sizes, a single checkpoint accepts any non-empty subset of the streams. Pretraining improves all five task families at every size, alongside competitive performance across external egocentric benchmarks. Finally, we build EgoBuddy to organize multimodal observations into hierarchical text memory for long-horizon egocentric understanding. Video-audio evaluations examine retrieval and question answering under progressively smaller visual frame budgets. Together, these components advance egocentric understanding toward a richer account of what is "seen" and "heard", and how the wearers "move".



Figure 1: **Egocentric Understanding Beyond Just Vision.** Wearable sensors capture complementary aspects of everyday experience: "Vision" reveals objects and surroundings, "Audio" captures sound events and vocal cues, and "IMU" measures motion and head orientation. In this work, we introduce EgoMM, which brings together benchmarking, multimodal learning, and long-horizon memory to advance egocentric understanding through complementary wearable signals for real-world use cases.

1 INTRODUCTION

Egocentric perception brings intelligent systems into the flow of everyday life, where understanding depends on what a person sees, hears, and experiences through movement (Grauman et al., 2022). Wearable devices increasingly capture these complementary signals together, connecting first-person video with audio and inertial measurements (Grauman et al., 2024; Yang et al., 2025). This creates an opportunity to build assistants with a richer account of everyday experience.

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.

That experience often exceeds what a camera can reveal. In a dim room, objects become difficult to distinguish while conversation remains audible. Glasses resting on a desk may retain an unchanging view as activities continue nearby. During active use, motion signals capture turns, pauses, and changes in posture that scene appearance alone may not resolve (Ma et al., 2024). Following everyday life therefore calls for complementary sensing across changing conditions.

Progress in egocentric understanding has nevertheless been driven primarily by video, supported by large datasets (Grauman et al., 2022; Damen et al., 2022) and advances in video-language learning (Lin et al., 2022; Zhao et al., 2023). Audio-visual models broaden this perspective (Xu et al., 2025b; Qwen Team, 2025a), but important questions remain about how different signals contribute. A visible instrument can suggest a sound, and a familiar activity can suggest movement, even when the corresponding event was neither heard nor sensed. Aggregate scores can therefore conceal limitations in acoustic and motion understanding. Table 14 compares how existing egocentric resources cover these signals and their associated learning and evaluation tasks.

EgoMM brings evaluation, learning, and memory together to connect wearable sensing with a more complete understanding of everyday experience. At its foundation, EgoMMU makes the contribution of each modality an explicit dimension of evaluation. As shown in Figure 6, our benchmark integrates five task families that span semantic understanding, reasoning, and temporal localization through diverse task-modality combinations. Each item specifies its intended evidence, with automated checks and human verification supporting construction and controlled ablations examining modality dependence. Recognizing a scene, interpreting its soundscape, and understanding wearer motion place distinct demands on a model, as illustrated in Figure 7. The benchmark provides a common basis for studying these individual capabilities and their combination.

Turning this evaluation perspective into learned capability is the role of EgoMind, which brings vision, audio, and IMU signals into a shared language representation. We support this learning with nearly a thousand hours of recordings and 1.75 million instruction-tuning items. Staged sensor pretraining accommodates the modalities' different starting points: audio builds on pretrained acoustic representations, while the inertial pathway learns motion-language correspondence before alignment with text generation. Modality-matched instruction tuning connects these representations to diverse tasks while grounding supervision in the supplied inputs. A single checkpoint accepts any non-empty subset of the three streams, supporting different sensor combinations and controlled comparisons without retraining. Across four model sizes, pretraining consistently improves all five task families, while external evaluations demonstrate competitive performance beyond EgoMMU.

Understanding an experience also requires preserving observations whose relevance may emerge only later. EgoBuddy addresses this temporal dimension by organizing multimodal descriptions into a hierarchical text memory for subsequent retrieval and answering. It connects the detail of individual clips with broader temporal context, allowing questions to draw on observations distributed across long recordings. Visual input is treated as a budgeted resource, bringing the quality of remembered information into direct relation with the amount of visual evidence supplied. Video-audio experiments examine retrieval and question answering across decreasing frame budgets, studying how complementary perception supports memory when visual observations become sparse.

To summarize, the key contributions of this work are fourfold:

- To the best of our knowledge, EgoMM makes the **first** attempt to unite modality-specific benchmarking, tri-modal language learning, and long-horizon memory for egocentric scene understanding.
- We establish EgoMMU with a diverse spectrum of task-modality combinations across five task families, supported by item verification and controlled analyses of modality dependence.
- We develop EgoMind to integrate vision, audio, and IMU through staged sensor pretraining and modality-matched instruction tuning, supporting flexible input combinations and evaluations.
- Finally, we build EgoBuddy for long-horizon memory and conduct extensive evaluations of benchmark performance, sensor contributions, pretraining, and visual frame budgets for practical uses.

2 Related Work

Egocentric Datasets & Benchmarks. Large-scale datasets (Grauman et al., 2022; 2024; Yang et al., 2025; Damen et al., 2022) and surveys (Plizzari et al., 2024) support research on everyday activities,

skilled interactions, and long-term experience. Benchmarks cover reasoning, perspective taking, and assistive question answering (Mangalam et al., 2023; Cheng et al., 2023; Xiao et al., 2025), alongside nighttime vision (Zhang et al., 2026) and emergency medical assistance (Weerasinghe et al., 2026). MMG-Ego4D (Gong et al., 2023) studies action recognition and modality generalization with video, audio, and IMU, while EgoSound (Zhu et al., 2026) evaluates sound perception and audio-visual reasoning. Table 14 compares representative resources. EgoLife (Yang et al., 2025) annotates audio requirements, providing a precedent for identifying required evidence. **EgoMMU** extends modality-specific evaluation across five language-understanding task families, with construction checks and controlled ablations examining dependence on visual, acoustic, and motion evidence.

Multimodal Egocentric Scene Understanding. Video-language architectures (Qwen Team, 2025b; Zhu et al., 2025) and egocentric models (Ye et al., 2025) support instruction-based visual understanding, while omni-modal systems process audio and video jointly (Qwen Team, 2025a; Gemini Team, Google, 2026). Egocentric audio research examines sound-producing actions (Huh et al., 2023), audio-visual grounding, and hallucination (Seth et al., 2026). Sensor-language methods align wearable measurements with text (Moon et al., 2022; Zhang et al., 2025d), motion-language models describe behavior (Chen et al., 2024d), and shared embedding frameworks connect heterogeneous modalities (Girdhar et al., 2023). Language generation further requires representations that express relevant signal properties. **EgoMind** combines learned sensory representations and interpretable motion descriptions through staged pretraining and modality-matched instruction tuning, supporting diverse tasks and input combinations within one checkpoint.

Egocentric Assistants & Efficient Capture. Wearable memory assistance dates to SenseCam (Hodges et al., 2006). EgoLife (Yang et al., 2025) connects week-scale recordings with long-horizon questions, while LightMem-Ego (Chen et al., 2026) organizes visual and acoustic observations into hierarchical memory. Streaming benchmarks (Yan et al., 2025) and online video-language models (Chen et al., 2024c) examine responses as observations arrive. Efficiency methods reduce visual processing through spatiotemporal compression (Shen et al., 2025), while EgoTrigger (Paruchuri et al., 2025) uses audio to guide image capture. **EgoBuddy** studies long-horizon understanding through memory populated by multimodal clip descriptions. Our video-audio experiments compare captioners and uniform frame budgets under shared consolidation, retrieval, and answering procedures, examining how recorded evidence supports subsequent understanding.

3 EGOMM: MULTIMODAL EGOCENTRIC UNDERSTANDING

3.1 PROBLEM FORMULATION

We represent a 30 s egocentric clip as $x = (\mathbf{v}, \mathbf{a}, \mathbf{u})$, comprising video frames, an audio waveform, and IMU measurements when available. In the dual-sensor setting, $\mathbf{u} = (\mathbf{u}^L, \mathbf{u}^R)$ contains left and right acceleration and angular-velocity streams. Given an instruction q, the model generates text from a non-empty subset M of the available modalities M^{av}, that is,

$$\hat{y} = f_\phi\big(\mathbf{X}_M(x), q\big), \qquad \varnothing \neq M \subseteq M^{\text{av}} \subseteq \{V, A, I\}. \tag{1}$$

Here, $\mathbf{X}_M(x)$ denotes the corresponding input tokens and ϕ the trainable parameters. The reference y_M must be supported by the supplied evidence. A benchmark item's designated modality identifies the evidence needed to answer it, independently of which additional streams the evaluated model receives. Section D.1 details the input conventions.

3.2 EGOMMU: EVALUATING MULTIMODAL UNDERSTANDING

Tasks and Evidence. Our benchmark contains 3,762 items from 1,196 clips, organized into five task families: multiple choice, narration, fact checking, open-ended question answering, and temporal grounding. Four families each contain vision, audio, and IMU variants, while narration integrates the available modalities into one description, yielding 13 evaluation columns. This organization examines how task performance changes with the evidence required: recognizing visual details, interpreting acoustic events, and characterizing wearer motion. Figures 6 and 7 illustrate representative use cases.

Evidence-Grounded Construction. Benchmark clips come from recording groups held out from training, preventing neighboring windows from crossing the split. We author items from clip media,

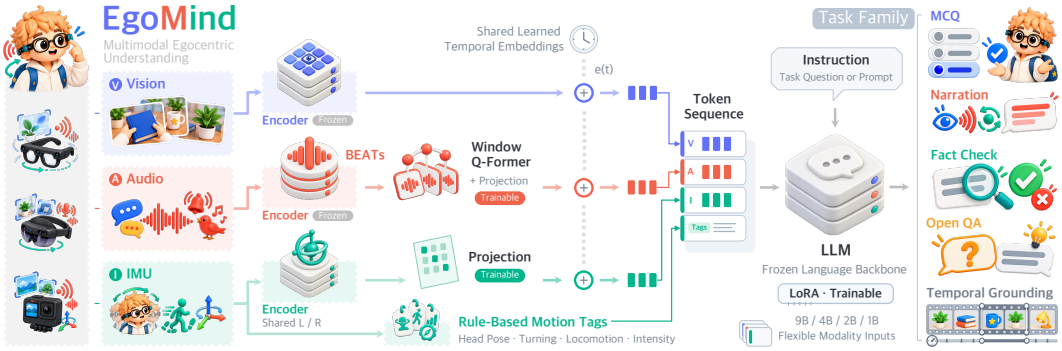


Figure 2: **Architecture Overview.** Available **Vision**, **Audio**, **IMU** streams are encoded into separate token blocks with shared temporal embeddings. IMU signals additionally provide rule-based motion tags. The tokens, tags, and task instruction condition a LoRA-adapted language backbone for five task families. The audio Q-Former and projection, IMU pathway, and LoRA adapters are trainable; the vision encoder, BEATs, and base language backbone remain frozen. See Appendix D.2 for details.

source annotations, and signal-derived descriptions. For IMU questions and claims, predefined rules convert measured features into structured motion descriptions that constrain prompt-based authoring. Automated checks and human verification assess format, reference correctness, ambiguity, and consistency with the designated evidence. Section B.3 describes the construction process.

Modality Dependence and Scoring. Text-only and field-conditioned probes examine shortcuts in wording and item structure, while temporal baselines characterize overlap attainable without conditioning on the clip. Complementary inference-time ablations withhold individual modalities from the same checkpoint, keeping the items and remaining inputs fixed. Together, these analyses assess whether performance reflects the intended evidence (Section B.4). Multiple choice uses answer accuracy, temporal grounding uses interval overlap, and open-ended responses and narration use task-specific grading. Fact checking combines verdict correctness with justification quality, assigning no credit to an incorrect verdict. Section C.2 provides the metrics and scoring conventions.

3.3 EgoMind: Learning from Wearable Signals

Unified Sensor Interface. **EgoMind** connects wearable signals to a Qwen3.5 vision-language backbone (Qwen Team, 2026), as illustrated in Figure 2. The visual pathway retains the backbone's vision encoder. Audio passes through frozen BEATs (Chen et al., 2022) and a trainable window Q-Former with an output projection; IMU signals pass through a learned encoder and projection. Each pathway produces tokens at the backbone's embedding width. With all modalities present, we concatenate their representations and the IMU tag text, that is,

$$\mathbf{X}_{\{V,A,I\}}(x) = \left[\mathbf{z}_V; \mathbf{z}_A; \mathbf{z}_I; \mathbf{z}_{\text{tag}}\right]. \tag{2}$$

The instruction q follows these blocks. Shared learned temporal embeddings encode clip-relative time across sensor tokens. Token blocks for absent modalities are omitted, and IMU tag text is included only when IMU is supplied. The vision encoder, BEATs, and base language-model weights remain frozen, with language adaptation performed through LoRA (Hu et al., 2022). The same design supports four backbone sizes. Due to page limits, see Section D.2 for additional details.

Motion Representation. The IMU pathway combines temporal signal representations with explicit motion attributes. A convolutional stem and Transformer, shared across the left / right sensors, encode motion dynamics into learned tokens. In parallel, predefined rules map signal features to $\text{Tag}(\mathbf{u})$, which is embedded as $\mathbf{z}_{\text{tag}}$ and describes properties such as movement intensity and head orientation. The two forms expose complementary information: continuous signal patterns and interpretable summaries of selected motion properties. Section A.4 details the feature-to-tag mapping.

Staged Sensor Pretraining. The pathways enter training with different initial representations. Phase A connects the pretrained audio encoder to language generation by optimizing the Q-Former and projection on acoustic captions (Eq. 17). Supervision covers sound events and vocal characteristics, excluding transcribed speech and conversational content. Phase I first learns an IMU representation from scratch through symmetric contrastive alignment with caption embeddings (Eq. 20), establishing motion-language correspondence before introducing generation. This encoder is pretrained once and shared as the initialization across model sizes. A subsequent caption-alignment step trains the

Table 1: **Capture Platforms and IMU Motion Tags.** We compare Vision, Audio, IMU configurations across three capture platforms, together with coordinate conventions and signal-derived motion tags. Appendix A describes the fine-grained data sources, preprocessing, and motion-tag extraction.

	Project Aria	HoloLens 2	GoPro
V Modality: Vision			
Camera Type	RGB, Rolling Shutter	RGB, PV, Autofocus	Action Cam, Fixed-Focus
Resolution	1408 × 1408	1920 × 1080	Up to 4K–5K
Field Of View	110° (Fisheye)	∼65°	∼120° (wide)
A Modality: Audio			
Microphones	7-Mic Spatial Array	5-Channel Array	3 Mics (Stereo)
Sample Rate	48 kHz	-	48 kHz
I Modality: IMU			
Inertial Units	2 (Bilateral, L+R)	1 (Head)	1 (Head)
Sample Rate	1000 / 800 Hz (R / L)	Not Reported	∼200 Hz (GPMF)
Coordinate			
Motion Tags	• *Pitch* — Gravity g_y, g_z	Looking Up · Looking Ahead · Looking Down	
	• *Roll* — Gravity g_x, g_y	Tilt Left · Upright · Tilt Right	
	• *Head Motion* — Turn + Nod Rate	Still Head · Slow Scan · Active Scan · Quick Turn	
	• *Locomotion* — Accel + Gyro	Stationary · Walk · Run · Rhythmic · Light · Vehicle	
	• *Intensity* — Accel RMS	Still · Light · Moderate · Vigorous	

encoder and a backbone-specific projection through the frozen language model (Eq. 21). Tag text is excluded from this step, so caption generation is conditioned on learned signal tokens. Section D.3 details the objectives and trainable components.

Modality-Matched Instruction Tuning. We assemble nearly a thousand hours of egocentric recordings into 1.75 million instruction-tuning items. Recordings span Project Aria, HoloLens 2, and GoPro-class devices (Table 1); synchronized preprocessing and modality-specific annotation organize their available signals into training targets (Section A). Figure 5 illustrates label generation.

Joint instruction tuning updates the trainable sensor components and LoRA adapters using modality-matched targets. The sampling distribution π retains every modality required by a target and samples additional available streams, complemented by examples constructed for specific input combinations. Supervision remains grounded in the supplied modalities, including when multiple streams are combined. We minimize the mean negative log-likelihood over answer tokens and the end-of-sequence token, indexed by $\mathcal{S}$, that is,

$$\mathcal{L}(\phi) = -\mathbb{E}_{(x,q)}\, \mathbb{E}_{M\sim\pi}\left[\frac{1}{|\mathcal{S}|}\sum\nolimits_{k\in\mathcal{S}} \log p_\phi\big(y_{M,k} \mid y_{M,<k}, \mathbf{X}_M(x), q\big)\right]. \tag{3}$$

Here, y_M denotes a reference supported by M; the same target may supervise multiple subsets that retain its required evidence. The resulting checkpoint accepts any non-empty modality subset at inference. Due to page limits, for additional details, kindly refer to Section D.4.

3.4 EgoBuddy: Long-Horizon Multimodal Memory

Observations become useful beyond their capture time when they can be retained and retrieved in context. EgoBuddy uses EgoMind-9B to describe successive clips and organizes these records into hourly episodes and day summaries, following hierarchical-memory designs (Yang et al., 2025; Chen et al., 2026). Hierarchical retrieval selects relevant days and hours, while questions about the most recent occurrence of an event use keyword search; selected clip records then support answering. Gemini-3.5-Flash performs consolidation, retrieval, and answering over the text memory. This separates the content preserved by perception from its subsequent retrieval and use.

We examine this memory under a per-clip visual budget, uniformly sampling B frames while retaining the full audio input. The reported experiments use video and audio, comparing EgoMind-generated records with EgoGPT records under shared downstream procedures. They evaluate retrieval over precomputed memory with coarse day/hour temporal restrictions, rather than strictly causal online

Table 2: **EgoMMU Benchmark Results.** Performance of state-of-the-art multimodal models across task-modality combinations. Evaluation inputs and metrics are specified in Appendix C. All scores are higher the better. Due to page limits, see Appendix E.1 for complete results of 50+ models.

Model	Size	Type	MCQ			Narr	Fact Check			Open QA			Temporal		
			V	A	I		V	A	I	V	A	I	V	A	I
Group: Open-Source Omni Models															
Qwen3-Omni	30B	V A	63.9	59.8	37.6	39.3	54.2	54.5	53.6	56.4	42.8	47.7	33.3	11.6	13.3
Gemma 4	12B	V A	60.6	41.3	35.2	36.9	52.7	43.3	50.9	53.2	33.0	38.3	14.2	8.6	8.3
Gemma 4	4B	V A	53.1	33.2	32.3	32.9	43.6	31.1	40.3	33.9	32.5	30.9	11.5	7.1	6.3
Gemma 4	2B	V A	43.9	29.7	29.7	31.0	36.7	26.8	25.3	27.0	25.6	29.9	11.0	5.2	4.2
Intern-Omni	9B	V A	59.1	53.7	36.1	40.9	51.6	45.5	39.7	44.7	36.1	43.7	10.6	7.9	5.7
MiniCPM-o	8B	V A	54.9	45.8	36.1	40.0	55.0	44.3	32.7	50.1	35.9	34.3	15.4	7.9	4.5
Qwen2.5-Omni	7B	V A	56.9	46.6	39.4	41.7	43.7	46.7	33.9	44.1	36.7	40.9	14.1	10.8	11.5
Qwen2.5-Omni	3B	V A	51.9	39.7	31.0	34.3	36.5	40.1	34.2	29.8	30.3	31.6	9.8	10.0	10.8
VideoLLaMA2.1	7B	V A	54.7	44.1	35.9	36.8	50.2	37.2	43.1	36.7	35.1	43.5	16.2	8.7	10.8
Baichuan-Omni-1.5	7B	V A	52.1	43.7	34.7	33.1	49.9	42.6	33.1	39.1	34.3	40.7	9.3	5.4	6.1
EgoGPT	7B	V A	57.3	52.8	35.1	42.7	46.8	52.1	39.6	45.3	40.7	38.0	19.5	12.1	8.7
Phi-4-Multimodal	6B	V A	50.3	36.7	34.1	25.3	37.7	38.6	40.1	24.7	35.8	36.6	9.1	9.8	10.0
EgoMind	9B	V A I	70.3	67.3	66.2	52.1	61.6	62.0	66.7	61.9	52.6	62.2	34.5	16.7	20.3
EgoMind	4B	V A I	61.5	55.1	47.2	44.7	54.7	53.8	57.1	56.2	44.8	52.8	29.5	14.0	18.5
EgoMind	2B	V A I	56.7	48.9	43.8	39.3	47.3	50.7	49.2	42.3	38.0	43.5	25.2	10.5	14.7
EgoMind	1B	V A I	50.5	39.4	39.0	36.1	41.9	46.1	44.7	37.2	35.7	39.4	18.1	10.3	12.5
Group: Open-Source Vision-Only Models															
InternVL3	78B	V	55.6	30.2	38.4	37.2	52.0	35.9	45.8	46.8	30.4	43.5	17.4	11.3	9.8
InternVL3	8B	V	51.4	23.6	40.7	35.7	39.7	21.0	47.2	43.6	33.6	42.0	13.8	5.9	5.2
Qwen3.5	35B	V	63.7	24.3	38.7	34.7	54.7	32.4	46.1	50.7	30.8	43.8	28.1	7.0	5.3
Qwen3.5	9B	V	56.2	25.9	37.2	30.1	49.1	29.4	46.7	45.6	29.8	44.6	24.7	5.1	4.9
Qwen3-VL	32B	V	54.7	24.4	38.5	30.5	51.5	36.8	51.3	32.4	32.0	43.7	18.7	6.8	4.9
Qwen3-VL	8B	V	52.5	22.7	38.3	29.3	50.8	36.4	45.4	30.2	30.4	44.0	15.0	6.1	4.6
Qwen3-Omni	30B	V	63.1	29.9	42.0	34.1	52.7	36.1	53.1	57.1	29.7	44.3	29.7	8.5	11.4
Qwen2.5-VL	72B	V	54.7	21.6	42.7	35.5	51.5	26.0	45.2	38.8	34.8	35.8	17.9	8.6	8.8
Qwen2.5-VL	7B	V	51.2	21.0	40.3	33.7	50.2	19.0	42.1	36.8	30.8	30.0	16.8	7.0	4.7
Qwen2.5-Omni	7B	V	55.4	19.5	38.1	39.4	44.1	17.7	36.8	42.3	26.7	40.3	15.3	7.7	11.2
Qwen2.5-Omni	3B	V	52.3	20.6	33.7	32.9	36.0	16.8	34.2	30.0	25.1	32.7	10.3	6.1	11.2
Group: Proprietary Models															
Gemini 3.8 Flash	–	V A	79.1	66.1	35.8	51.8	71.6	70.7	55.2	70.8	60.1	50.3	31.7	15.0	19.8
Gemini 3.1 Pro	–	V A	78.2	65.0	45.4	46.2	70.3	70.4	49.3	68.0	52.0	46.3	39.7	16.9	15.3
Gemini 3.1 Flash Lite	–	V A	71.2	51.0	34.8	44.7	66.2	62.8	44.7	62.1	44.3	34.7	22.4	14.7	15.2
GPT 5.4	–	V	72.3	28.7	44.1	51.8	70.5	32.3	52.8	72.2	32.1	50.1	28.9	9.5	14.9
GPT 5.4 Mini	–	V	64.8	20.1	31.6	49.6	62.9	33.7	54.6	54.3	30.7	41.3	20.4	8.2	12.0
Opus 5	–	V	75.1	30.7	40.2	52.8	72.6	32.1	64.7	70.1	42.7	45.8	38.7	7.6	12.8
Sonnet 5	–	V	63.7	29.7	26.1	42.7	67.7	26.1	61.6	56.6	33.6	33.2	26.9	4.3	10.7
Haiku 4.5	–	V	45.5	20.4	23.0	27.0	44.7	28.8	47.8	43.2	26.4	35.8	18.0	3.3	5.4

operation (Section D.5). Varying B connects long-horizon performance to the amount of visual evidence supplied; the frame-budget settings are detailed in Section D.6.

4 Experiments

We conduct extensive benchmark comparisons, controlled input ablations, sensor-pretraining analyses, and long-horizon memory experiments. Comparisons cover open vision-language models, omni-modal models, and proprietary systems, alongside four sizes of EgoMind. Due to page limits, implementation details, modality handling, and evaluation protocols are detailed in Appendix D.

We report EgoMMU and EgoAVU-Bench (Seth et al., 2026) results together with external evaluations on EgoSchema, EgoThink, EgoBlind, and EgoPolice (Mangalam et al., 2023; Cheng et al., 2023; Xiao et al., 2025; Gonzalez Saez-Diez et al., 2026). EgoLife-QA (Yang et al., 2025) evaluates long-horizon understanding. Each benchmark retains its scoring protocol. Due to page limits, answer parsing and grading details are specified in Appendix C. The main tables present representative configurations. We include the complete results of more than 60 models and their variants in Appendix E.1.

4.1 Benchmark Performance

Capabilities Across Modalities. Table 2 shows consistent improvements across all task-modality columns as EgoMind scales from 1B to 9B. IMU MCQ accuracy increases from 39.0% to 66.2%, placing the 9B model ahead of Gemini 3.1 Pro (45.4%) on this task. Their visual MCQ accuracies are 70.3% and 78.2%, respectively. These modality-dependent rankings show why visual recognition alone does not characterize wearer-motion understanding. Cross-model comparisons also reflect differences in available inputs and training. Temporal grounding retains substantial headroom, with constant-span references providing context for interpreting localization scores (Appendix B.4).

Audio-Visual Description Quality. Table 3 evaluates description quality, sound-source attribution, and hallucination on EgoAVU-Bench (Seth et al., 2026). Relative to Qwen3.5-9B, EgoMind-9B

Table 3: **EgoAVU-Bench Results.** Evaluations of audio-visual segment narration (AVSN), sound-source attribution (SSA), and audio-visual hallucination (AVH). † denotes fine-tuning on EgoAVU-Instruct. All scores are higher the better. Due to page limits, see Appendix E.1 for complete results.

Model	Size	Type	AVSN-R	AVSN-S	SSA	AVH				Overall (LLM Judge)
						Action	Sound	Object	Average	
Open-Source Omni Models										
Qwen3-Omni	30B	V A	14.81	1.57	1.40	3.14	2.55	2.03	2.57	2.14
Intern-Omni	9B	V A	14.70	1.82	1.47	2.37	2.41	2.27	2.35	2.07
MiniCPM-o	8B	V A	12.19	2.06	1.43	2.41	2.47	2.21	2.36	2.12
Baichuan-Omni	7B	V A	14.21	1.79	1.49	2.60	2.44	2.31	2.45	2.13
EgoAVU†	7B	V A	28.70	2.63	3.20	–	–	–	–	–
Qwen2.5-Omni	7B	V A	13.39	1.99	1.50	2.80	2.42	2.33	2.52	2.21
VideoLLaMA2	7B	V A	13.89	1.71	1.51	3.23	2.58	2.41	2.74	2.29
EgoGPT	7B	V A	14.30	1.52	1.37	2.68	2.39	2.37	2.48	2.07
Phi4-Multimodal	6B	V A	16.90	1.69	1.42	1.90	2.10	2.13	2.04	1.85
EgoMind†	9B	V A I	26.77	2.41	3.27	3.61	2.98	2.87	3.15	3.03
EgoMind	9B	V A I	17.53	1.73	1.61	3.27	2.71	2.58	2.85	2.38
Open-Source Vision-Only Models										
Qwen3.5	35B	V	16.21	1.71	1.59	3.11	2.50	2.21	2.61	2.22
Qwen3.5	9B	V	16.47	1.69	1.48	3.32	2.62	1.87	2.60	2.20
Proprietary Models										
Gemini 3.8 Flash	–	V A	15.98	1.83	2.27	3.37	2.88	2.61	2.95	2.59
Gemini 3.1 Pro	–	V A	15.37	1.71	2.31	3.66	3.01	2.57	3.08	2.65
GPT 5.4	–	V	16.44	1.61	2.20	3.51	2.61	2.31	2.81	2.45
Opus 5	–	V	15.37	2.03	2.07	3.33	2.77	2.53	2.88	2.55
Sonnet 5	–	V	16.10	2.17	1.81	3.21	2.41	2.41	2.68	2.40
Haiku 4.5	–	V	11.53	1.87	1.42	2.77	2.41	2.31	2.50	2.16

Table 4: **Results on Public Egocentric Benchmarks.** All scores are higher the better. Due to page limits, see Appendix C for detailed evaluation protocols and Appendix E.1 for complete results.

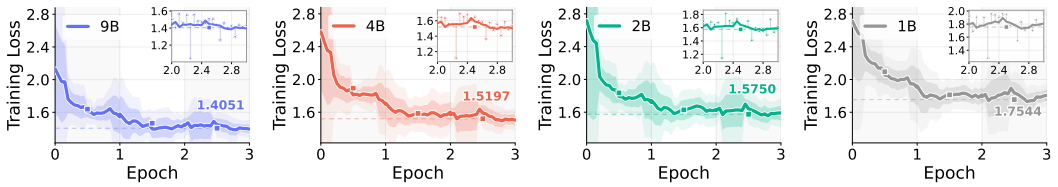
Model	Size	Type	EgoSchema	EgoBlind						EgoThink	EgoPolice		
				inform	safety	navig	resour	tool	commu		1s	10s	60s
Group: Open-Source Omni Models													
Qwen3-Omni	30B	V A	61.0	55.6	61.0	28.6	40.4	40.0	40.5	65.6	56.5	58.5	55.2
Gemma 4	12B	V A	68.4	50.5	46.6	26.5	36.4	30.0	45.9	64.3	53.1	56.0	48.4
Qwen2.5-Omni	7B	V A	62.4	47.2	42.3	30.7	32.3	31.2	35.1	62.6	52.6	53.2	46.1
EgoGPT	7B	V A	75.4	46.4	59.0	22.4	36.4	22.9	43.2	61.4	53.8	54.3	50.4
Phi-4-Multimodal	6B	V A	59.2	44.8	40.0	23.0	31.3	28.6	38.6	58.7	51.4	46.6	39.3
EgoMind	9B	V A I	72.1	54.3	63.1	28.6	42.7	43.1	47.7	73.6	66.7	62.0	54.3
Group: Open-Source Vision-Only Models													
Qwen3.5	9B	V	67.1	52.3	61.6	27.6	40.4	45.7	45.9	73.4	59.1	53.7	43.0
Qwen3-VL	32B	V	64.0	51.0	58.6	30.7	42.1	37.6	44.3	66.9	56.0	53.3	40.0
InternVL3	8B	V	68.2	52.7	60.3	32.8	36.4	32.9	48.6	62.3	53.8	54.4	40.6
Group: Proprietary Models													
Gemini 3.8 Flash	–	V A	79.0	64.9	59.7	45.9	58.6	58.6	59.5	77.1	81.6	84.2	81.8
Gemini 3.1 Pro	–	V A	76.5	60.0	61.1	47.2	48.5	62.0	51.4	75.1	82.3	83.0	79.7
GPT 5.4	–	V	78.2	62.7	62.5	45.4	47.5	60.0	53.2	74.6	75.6	78.9	76.3
Opus 5	–	V	76.7	58.7	66.1	45.4	47.0	55.3	52.8	64.7	72.8	76.3	70.1
Sonnet 5	–	V	74.0	44.5	64.9	37.3	46.3	40.0	44.3	68.4	63.4	65.7	60.8
Haiku 4.5	–	V	72.1	38.4	54.4	36.7	45.5	42.4	37.0	64.3	55.4	59.1	52.0

improves the object-hallucination rating from 1.87 to 2.58, while the action-hallucination rating remains close (3.32 versus 3.27); higher ratings indicate fewer hallucinations. The overall score increases from 2.20 to 2.38. These category-dependent gains motivate separate assessment of objects, actions, and acoustic properties, with complete results and discussion in Appendix E.1.

Transfer Across Egocentric Tasks. Table 4 examines capabilities beyond EgoMMU. Compared with its Qwen3.5-9B backbone (Qwen Team, 2026), EgoMind-9B improves EgoSchema accuracy from 67.1% to 72.1% and retains similar EgoThink performance (73.4 versus 73.6). EgoPolice gains range from 7.6 to 11.3 percentage points across its three observation lengths. Transfer is nevertheless nonuniform: EgoBlind information-seeking accuracy improves from 52.3% to 54.3%, while the tool category decreases from 45.7% to 43.1%. These mixed outcomes motivate evaluation across task categories and temporal settings, with complete results and further analysis reported in Appendix E.1.

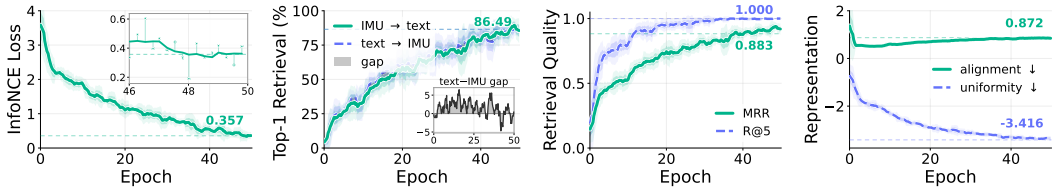
4.2 Modality Dependence

Selective Sensitivity to Audio. Controlled comparisons reveal whether changes follow the evidence required by a task. Removing audio from Qwen3-Omni-30B (Qwen Team, 2025a) reduces audio MCQ accuracy from 59.8% to 29.9%, while visual MCQ remains close (63.9% versus 63.1%). Across six paired omni-modal variants, audio MCQ, fact checking, and open-ended scores all decrease. The concentration of losses in acoustic tasks supports their modality dependence. Due to page limits, the full comparisons and analyses are placed in Appendix E.2.



(a) EgoMind-9B (b) EgoMind-4B (c) EgoMind-2B (d) EgoMind-1B

Figure 3: **Audio Alignment.** Training loss across model sizes during audio caption alignment in Phase A with a frozen language backbone. See Appendix E.3.1 for detailed convergence analysis.



(a) Loss Convergence (b) Retrieval Accuracy (c) Retrieval Quality (d) Representations

Figure 4: **IMU Contrastive Learning.** Loss, retrieval performance, and embedding geometry during contrastive pretraining in Phase I. See Appendix E.3.2 for metric definitions and further analysis.

Table 5: **Long-Horizon Question Answering on EgoLife-QA.** Accuracy across five question types. See Appendix C.1 for evaluation settings and Appendix D.5 for memory construction and retrieval.

Model	Type	#Frames	EgoLifeQA					Average
			EntityLog	EventRecall	HabitIns	RelationMap	TaskMaster	
Group: Retrieval-Based Models								
VideoRAG	–	64	19.2	15.9	19.7	28.0	14.3	20.0
LightRAG	–	64	40.8	40.4	36.0	32.0	50.7	39.2
Group: Proprietary Models								
Gemini 3.1 Flash Lite	V A	–	44.0	41.3	42.6	47.2	38.1	43.2
Gemini 2.5 Pro	V A	–	45.6	48.4	51.7	41.6	52.4	46.8
Group: Open-Source Omni Models								
VideoChat-Flash-7B	V	1024	36.8	39.7	34.4	31.2	41.3	36.4
InternVideo2.5-8B	V	512	34.4	37.3	42.6	30.4	31.7	34.8
VideoLLaMA3-7B	V	128	32.8	35.7	37.7	27.2	33.3	32.8
LongVA-7B	V	128	33.6	38.1	36.1	39.2	28.6	35.8
Qwen3.5-9B	V	64	28.8	31.7	45.9	33.6	36.5	33.8
LLaVA-OneVision	V	30 (FPS=1)	36.8	34.9	31.1	22.4	28.6	30.8
EgoRetriever	V	30 (FPS=1)	42.5	45.1	37.7	35.8	46.2	41.5
EgoGPT (EgoIT-99K)	V A	30 (FPS=1)	35.2	36.5	27.9	29.6	36.5	33.1
EgoGPT (EgoIT-99K+D1)	V A	30 (FPS=1)	39.2	36.5	31.1	33.6	39.7	36.0
EgoSelf	V A	30 (FPS=1)	38.4	42.9	44.3	36.0	41.3	40.6
EgoBuddy	V A	1 (FPS=1/30)	43.0	47.1	33.4	38.9	43.9	41.3
EgoBuddy	V A	3 (FPS=1/10)	44.3	47.8	36.7	38.7	49.2	43.3
EgoBuddy	V A	6 (FPS=1/5)	47.5	52.0	43.1	37.1	52.7	46.5
EgoBuddy	V A	15 (FPS=1/2)	48.6	52.3	46.7	40.1	53.6	48.3
EgoBuddy	V A	30 (FPS=1)	48.9	53.7	48.9	42.3	55.6	50.0
Oracle†	–	–	66.1	65.9	37.3	41.5	66.7	56.4

Multimodal Scores and Acoustic Evidence. EgoAVU-Bench exhibits a complementary pattern: vision-only Qwen3.5-9B exceeds audio-enabled Qwen3-Omni-30B on SSA (1.48 *vs.* 1.40). Visible context can support plausible acoustic descriptions, so scores on mixed-media tasks alone do not isolate audio perception. Cross-model differences also reflect architecture and training; paired input ablations provide a more controlled assessment of how predictions depend on the available evidence.

Complementarity Across Three Streams. Table 23 extends the analysis to EgoMind. Removing IMU, including its tag text, lowers 9B IMU MCQ accuracy from 66.2% to 34.7%. Across the three larger sizes, vision and audio scores change by at most 2.5 points under this removal. Narration decreases under every modality ablation at every size, supporting the contribution of complementary observations to the shared description task. Residual scores after withholding also indicate that context and learned priors remain relevant. Complete comparisons are provided in Appendix E.2.

4.3 Sensor Pretraining

Benefits After Instruction Tuning. Table 24 compares random and pretrained sensor initialization under matched instruction tuning. Pretraining improves all five task families at every size. At 9B, MCQ increases from 40.1 to 67.9, and temporal grounding from 11.4 to 23.8. These gains show the value of sensor alignment before joint task learning. Both configurations retain deterministic IMU

Table 6: **Retrieval HIT Rates.** Retrieval quality under matched visual frame budgets. HIT rates assess four temporal resolutions; medΔ measures median distance from the reference moment to the nearest retrieved clip. See Appendix C.2 for detailed metric definitions and the no-skill baseline.

Model	Type	#Frames	Frame Rate	HIT Rate				medΔ	Acc | HIT	Acc | MISS
				DAY	HOUR	EVENT	CLIP			
EgoGPT	V A	30	1	78.1%	33.0%	22.6%	4.4%	55.0 min	66.7%	37.6%
EgoBuddy	V A	30	1	83.7%	58.9%	52.6%	18.7%	5.6 min	73.8%	35.0%
EgoGPT	V A	15	1 / 2	77.9%	32.1%	20.3%	4.1%	61.2 min	67.1%	38.0%
EgoBuddy	V A	15	1 / 2	83.2%	57.3%	49.8%	16.2%	8.9 min	68.3%	34.7%
EgoGPT	V A	6	1 / 5	77.3%	31.7%	17.6%	3.9%	70.5 min	66.1%	38.3%
EgoBuddy	V A	6	1 / 5	82.6%	55.6%	47.0%	15.7%	11.3 min	69.7%	33.3%
EgoGPT	V A	3	1 / 10	77.1%	30.6%	17.1%	3.7%	76.3 min	68.6%	35.4%
EgoBuddy	V A	3	1 / 10	82.3%	54.7%	44.3%	15.1%	12.7 min	70.1%	36.3%
EgoGPT	V A	1	1 / 30	76.3%	29.8%	16.3%	3.4%	77.1 min	70.0%	39.8%
EgoBuddy	V A	1	1 / 30	79.8%	52.3%	42.1%	14.7%	15.9 min	65.3%	38.9%
No-Skill Floor†	–	–	–	75.9%	7.1%	3.5%	0.2%	–	–	–

tags, so this comparison measures pretraining benefits in their presence. The 9B factorial study in Figure 11 further examines audio and IMU initialization through training loss.

Learning Dynamics. Audio alignment achieves 83–92% of its total loss reduction in the first epoch, while larger backbones maintain lower losses thereafter (Figure 3). IMU contrastive pretraining reaches 86.5% top-1 retrieval accuracy among 64 within-batch candidates (Figure 4). Later training improves retrieval as matched embeddings move farther apart and their distribution becomes more dispersed. This distinguishes relative ranking quality from absolute pairwise proximity. Subsequent IMU alignment achieves 89–92% of its loss reduction in the first epoch (Figure 10), connecting the contrastively initialized representation to generation. Section E.3 analyzes these trajectories in detail.

Signal Dependence and Held-Out Performance. Training loss alone does not establish that a pathway uses the recorded signal. Input-substitution probes show positive IMU specificity at every size: replacing a clip's signal with another clip's measurements lowers agreement with the unchanged reference by 7.49 BLEU-4 points at 9B (Figure 9). Continued loss reduction need not improve held-out captioning, particularly at smaller sizes (Figure 8). Together, downstream ablations and signal-substitution probes provide complementary evidence beyond optimization curves.

4.4 Long-Horizon Understanding

Remembered Evidence and Answer Quality. We evaluate EgoBuddy on 475 EgoLife-QA questions for one participant, with localization measured on 270 questions with fixed reference moments. Tables 5 and 6 report answering and retrieval separately. At 30 frames per clip, replacing EgoGPT records with EgoMind-generated records increases HOUR HIT from 33.0% to 58.9% and CLIP HIT from 4.4% to 18.7%, while median temporal distance decreases from 55.0 to 5.6 minutes. Under shared downstream procedures, these differences show how perception records affect subsequent evidence retrieval. DAY, HOUR, EVENT, and CLIP results provide complementary temporal views.

Understanding Under Reduced Visual Input. Reducing the budget from 30 to 6 frames lowers QA accuracy from 50.0% to 46.5%, while the median retrieval distance increases from 5.6 to 11.3 minutes. At one frame per clip, QA accuracy remains 41.3%. Sparse visual input thus supports substantial answering capability while reducing the temporal precision of retrieval. These experiments retain full-clip audio and use uniform frame sampling over precomputed text memory with coarse temporal restrictions (Sections D.5 and D.6). They characterize the relationship between visual evidence, retrieval, and answering under the evaluated memory protocol.

5 Conclusion

We presented EgoMM, uniting benchmarking, multimodal learning, and long-horizon memory for egocentric understanding. Our experiments demonstrate the benefits of staged sensor pretraining, distinguish the contributions of complementary wearable signals, and show competitive performance across external benchmarks. Controlled modality ablations help link these gains to the evidence provided. Video-audio memory experiments further examine retrieval and answering under reduced visual frame budgets. These findings motivate more precise temporal grounding and adaptive capture toward wearable assistants that understand and remember everyday experience.

AI USE STATEMENT

We used AI tools to assist with language editing and the creation of illustrative figures. AI models also supported annotation generation, benchmark item authoring, and automated grading, as documented in the appendices. Generated benchmark items underwent automated checks and human verification. The authors take full responsibility for the manuscript and its reported findings.

REPRODUCIBILITY STATEMENT

We describe data processing and benchmark construction in Appendices A and B, evaluation protocols in Appendix C, and model architecture and training procedures in Appendix D. Appendix F provides prompts for captioning, benchmark authoring, and automated grading. Additional results and ablations appear in Appendix E, while Appendix G identifies the datasets, model checkpoints, and API services used.

ETHICS STATEMENT

This work builds on existing datasets and models, with their sources and licensing information documented in Appendix G. Our annotation pipeline constrains generated descriptions to observable evidence and replaces personal names and source-specific identity labels with clip-grounded descriptors. Because egocentric recordings may contain sensitive information about wearers and bystanders, downstream use should preserve source access restrictions and protect personal privacy. Applications in assistive or other safety-sensitive settings require further validation under their intended deployment conditions.

REFERENCES

Anthropic. Claude 4 model card. *Technical report*, 2025.

Natan Bagrov, Eugene Khvedchenia, Borys Tymchenko, Shay Aharon, Lior Kadoch, Tomer Keren, Ofri Masad, Yonatan Geifman, Ran Zilberstein, Tuomas Rintamaki, Matthieu Le, and Andrew Tao. Efficient video sampling: Pruning temporally redundant tokens for faster VLM inference. *arXiv preprint arXiv:2510.14624*, 2025. URL https://arxiv.org/abs/2510.14624.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025. URL https://arxiv.org/abs/2502.13923.

Keshigeyan Chandrasegaran, Agrim Gupta, Lea M. Hadzic, Taran Kota, Jimming He, Cristóbal Eyzaguirre, Zane Durante, Manling Li, Jiajun Wu, and Li Fei-Fei. HourVideo: 1-hour video-language understanding. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL https://arxiv.org/abs/2411.04998.

Vivek Chavan et al. IndEgo: A dataset of industrial scenarios and collaborative work for egocentric assistants. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2025. URL https://arxiv.org/abs/2511.19684.

Baiyu Chen, Wilson Wongso, Zechen Li, Yonchanok Khaokaew, Hao Xue, and Flora Salim. COMODO: Cross-modal video-to-IMU distillation for efficient egocentric human activity recognition. *arXiv preprint arXiv:2503.07259*, 2025. URL https://arxiv.org/abs/2503.07259.

Changan Chen, Kumar Ashutosh, Rohit Girdhar, David Harwath, and Kristen Grauman. SoundingActions: Learning how actions sound from narrated egocentric videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27242–27252, 2024a. URL https://arxiv.org/abs/2404.05206.

Changan Chen, Puyuan Peng, Ami Baid, Zihui Xue, Wei-Ning Hsu, David Harwath, and Kristen Grauman. Action2Sound: Ambient-aware generation of action sounds from egocentric videos. In *European Conference on Computer Vision*, 2024b. URL https://arxiv.org/abs/2406.09272.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024c. URL https://arxiv.org/abs/2406.11816.

Ling-Hao Chen, Shunlin Lu, Ailing Zeng, Hao Zhang, Benyou Wang, Ruimao Zhang, and Lei Zhang. MotionLLM: Understanding human behaviors from human motions and videos. *arXiv preprint arXiv:2405.20340*, 2024d. URL https://arxiv.org/abs/2405.20340.

Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, and Furu Wei. BEATs: Audio pre-training with acoustic tokenizers. *arXiv preprint arXiv:2212.09058*, 2022. URL https://arxiv.org/abs/2212.09058.

Yijun Chen, Boyi Xiao, Yixian Zhao, Haoting Xia, Buqiang Xu, Jizhan Fang, Yanya Li, Yaqi Zheng, Xuehai Wang, Zirui Xue, Liuxin Zhang, Hui Li, and Ningyu Zhang. LightMem-Ego: Your AI memory for everyday life. *arXiv preprint arXiv:2607.11487*, 2026. URL https://arxiv.org/abs/2607.11487.

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. *arXiv preprint arXiv:2311.15596*, 2023. URL https://arxiv.org/abs/2311.15596.

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*, 2022. URL https://arxiv.org/abs/2006.13256.

Fei Deng, Yinghui He, Chuntong Chu, Ge Wang, Han Ding, Jinsong Han, and Fei Wang. MobiDiary: Autoregressive action captioning with wearable devices and wireless signals. *arXiv preprint arXiv:2601.08204*, 2026. URL https://arxiv.org/abs/2601.08204.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. CLAP: Learning audio concepts from natural language supervision. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2023. URL https://arxiv.org/abs/2206.04769.

Rosario Forte, Giuseppe Lando, and Antonino Furnari. EgoStream: A diagnostic benchmark for streaming episodic memory in egocentric vision. *arXiv preprint arXiv:2605.31557*, 2026. URL https://arxiv.org/abs/2605.31557.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, et al. Video-MME: The first-ever comprehensive evaluation benchmark of multi-modal LLMs in video analysis. *arXiv preprint arXiv:2405.21075*, 2024. URL https://arxiv.org/abs/2405.21075.

Gemini Team, Google. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next-generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. URL https://arxiv.org/abs/2507.06261.

Gemini Team, Google. Gemini 3.1. *Technical report*, 2026.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. ImageBind: One embedding space to bind them all. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15180–15190, 2023. URL https://arxiv.org/abs/2305.05665.

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. In *European Conference on Computer Vision*, 2026. URL https://arxiv.org/abs/2607.11523.

Xinyu Gong, Sreyas Mohan, Naina Dhingra, Jean-Charles Bazin, Yilei Li, Zhangyang Wang, and Rakesh Ranjan. MMG-Ego4D: Multimodal generalization in egocentric action recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6481–6491, 2023. URL https://arxiv.org/abs/2305.07214.

Max Gonzalez Saez-Diez, Jihoon Chung, Adam D. Wolsky, Gregory Lanzalotto, Dean Knox, Jonathan Mummolo, Brandon M. Stewart, and Olga Russakovsky. EgoPolice: A benchmark for egocentric video understanding in high-stakes police body-worn camera footage. *arXiv preprint arXiv:2607.06468*, 2026. URL https://arxiv.org/abs/2607.06468.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4D: Around the world in 3,000 hours of egocentric video. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18973–18990, 2022. URL https://arxiv.org/abs/2110.07058.

Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-Exo4D: Understanding skilled human activity from first- and third-person perspectives. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19383–19400, 2024. URL https://arxiv.org/abs/2311.18259.

Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. AudioCLIP: Extending CLIP to image, text and audio. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 976–980, 2022. URL https://arxiv.org/abs/2106.13043.

Jiaming Han, Kaixiong Gong, Yiyuan Zhang, Jiaqi Wang, Kaipeng Zhang, Dahua Lin, Yu Qiao, Peng Gao, and Xiangyu Yue. OneLLM: One framework to align all modalities with language. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26574–26585, 2024. URL https://arxiv.org/abs/2312.03700.

Harish Haresamudram, Chi Ian Tang, Sungho Suh, Paul Lukowicz, and Thomas Plötz. Past, present, and future of sensor-based human activity recognition using wearables: A surveying tutorial on a still challenging task. *arXiv preprint arXiv:2411.14452*, 2024. URL https://arxiv.org/abs/2411.14452.

Steve Hodges, Lyndsay Williams, Emma Berry, Shahram Izadi, James Srinivasan, Alex Butler, Gavin Smyth, Narinder Kapur, and Kenneth R. Wood. SenseCam: A retrospective memory aid. In *International Conference on Ubiquitous Computing*, pp. 177–193, 2006.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2106.09685.

Xiaohu Huang, Hao Zhou, and Kai Han. PruneVid: Visual token pruning for efficient video large language models. *arXiv preprint arXiv:2412.16117*, 2024a. URL https://arxiv.org/abs/2412.16117.

Yifei Huang, Jilan Xu, Baoqi Pei, Yuping He, Guo Chen, Lijin Yang, Xinyuan Chen, Yaohui Wang, Zheng Nie, Jinyao Liu, Guoshun Fan, et al. Vinci: A real-time embodied smart assistant based on egocentric vision-language model. *arXiv preprint arXiv:2412.21080*, 2024b. URL https://arxiv.org/abs/2412.21080.

Jaesung Huh, Jacob Chalk, Evangelos Kazakos, Dima Damen, and Andrew Zisserman. EPIC-Sounds: A large-scale dataset of actions that sound. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1–5, 2023. URL https://arxiv.org/abs/2302.00646.

Sheikh Asif Imran, Mohammad Nur Hossain Khan, Subrata Biswas, and Bashima Islam. LLaSA: A sensor-aware LLM for natural language reasoning of human activity from IMU data. *arXiv preprint arXiv:2406.14498*, 2024. URL https://arxiv.org/abs/2406.14498.

Sijie Ji, Xinzhe Zheng, and Chenshu Wu. HARGPT: Are LLMs zero-shot human activity recognizers? *arXiv preprint arXiv:2403.02727*, 2024. URL https://arxiv.org/abs/2403.02727.

Giuseppe Lando, Rosario Forte, Giovanni Maria Farinella, and Antonino Furnari. How far can off-the-shelf multimodal large language models go in online episodic memory question answering? *arXiv preprint arXiv:2506.16450*, 2025. URL https://arxiv.org/abs/2506.16450.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-OneVision: Easy visual task transfer. *Transactions on Machine Learning Research*, 2025a. URL https://arxiv.org/abs/2408.03326.

Jiahao Nick Li, Zhuohao Zhang, and Jiaju Ma. OmniQuery: Contextually augmenting captured multimodal memory to enable personal question answering. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2025b. URL https://arxiv.org/abs/2409.08250.

Xiang Li, Heqian Qiu, Lanxiao Wang, Hanwen Zhang, Chenghao Qi, Linfeng Han, Huiyu Xiong, and Hongliang Li. Challenges and trends in egocentric vision: A survey. *Machine Intelligence Research*, 23(1):1–33, 2026. URL https://arxiv.org/abs/2503.15275.

Yadong Li, Jun Liu, Tao Zhang, Tao Zhang, Song Chen, Tianpeng Li, Zehuan Li, Lijun Liu, Lingfeng Ming, Guosheng Dong, Da Pan, et al. Baichuan-Omni-1.5 technical report. *arXiv preprint arXiv:2501.15368*, 2025c. URL https://arxiv.org/abs/2501.15368.

Yin Li, Miao Liu, and James M. Rehg. In the eye of the beholder: Gaze and actions in first person video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):6731–6747, 2023. URL https://arxiv.org/abs/2006.00626.

Zechen Li, Shohreh Deldari, Linyao Chen, Hao Xue, and Flora D. Salim. SensorLLM: Aligning large language models with motion sensors for human activity recognition. In *Conference on Empirical Methods in Natural Language Processing*, pp. 354–379, 2025d. URL https://arxiv.org/abs/2410.10624.

Zhe Li, Weihao Yuan, Yisheng He, Lingteng Qiu, Shenhao Zhu, Xiaodong Gu, Weichao Shen, Yuan Dong, Zilong Dong, and Laurence T. Yang. LaMP: Language-motion pretraining for motion generation, retrieval, and captioning. *arXiv preprint arXiv:2410.07093*, 2024. URL https://arxiv.org/abs/2410.07093.

Kevin Qinghong Lin, Alex Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Zhongcong Xu, Difei Gao, Rongcheng Tu, Wenzhe Zhao, Weijie Kong, Chengfei Cai, Hongfa Wang, Dima Damen, Bernard Ghanem, Wei Liu, and Mike Zheng Shou. Egocentric video-language pretraining. In *Advances in Neural Information Processing Systems*, 2022. URL https://arxiv.org/abs/2206.01670.

Jiasen Lu, Christopher Clark, Sangho Lee, Zichen Zhang, Savya Khosla, Ryan Marten, Derek Hoiem, and Aniruddha Kembhavi. Unified-IO 2: Scaling autoregressive multimodal models with vision, language, audio, and action. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26429–26445, 2024. URL https://arxiv.org/abs/2312.17172.

Zhaoyang Lv, Nicholas Charron, Pierre Moulon, Alexander Gamino, Cheng Peng, Chris Sweeney, Edward Miller, et al. Aria everyday activities dataset. *arXiv preprint arXiv:2402.13349*, 2024. URL https://arxiv.org/abs/2402.13349.

Lingni Ma, Yuting Ye, Fangzhou Hong, Vladimir Guzov, Yifeng Jiang, Rowan Postyeni, Luis Pesqueira, Alexander Gamino, Vijay Baiyya, Hyo Jin Kim, et al. Nymeria: A massive collection of multimodal egocentric daily motion in the wild. In *European Conference on Computer Vision*, pp. 445–465, 2024. URL https://arxiv.org/abs/2406.09905.

Neelu Madan, Andreas Møgelmose, Rajat Modi, Yogesh S. Rawat, and Thomas B. Moeslund. Foundation models for video understanding: A survey. *arXiv preprint arXiv:2405.03770*, 2024. URL https://arxiv.org/abs/2405.03770.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. EgoSchema: A diagnostic benchmark for very long-form video language understanding. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2308.09126.

MiniCPM Team. MiniCPM-o 2.6: A GPT-4o level MLLM for vision, speech, and multimodal live streaming on your phone. *Technical report*, 2025.

Seungwhan Moon, Andrea Madotto, Zhaojiang Lin, Alireza Dirafzoon, Aparajita Saraf, Amy Bearman, and Babak Damavandi. IMU2CLIP: Multimodal contrastive learning for IMU motion sensors from egocentric videos and text. *arXiv preprint arXiv:2210.14395*, 2022. URL https://arxiv.org/abs/2210.14395.

Girish Narayanswamy, Xin Liu, Kumar Ayush, Yuzhe Yang, Xuhai Xu, Shun Liao, Jake Garrison, Shyam A. Tailor, Jacob E. Sunshine, Yun Liu, Tim Althoff, et al. Scaling wearable foundation models. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.13638.

OpenAI. GPT-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. URL https://arxiv.org/abs/2410.21276.

Jongwoo Park, Kanchana Ranasinghe, Kumara Kahatapitiya, Wonjeong Ryu, Donghyun Kim, and Michael S. Ryoo. Too many frames, not all useful: Efficient strategies for long-form video QA. *arXiv preprint arXiv:2406.09396*, 2024. URL https://arxiv.org/abs/2406.09396.

Akshay Paruchuri, Sinan Hersek, Lavisha Aggarwal, Qiao Yang, Xin Liu, Achin Kulshrestha, Andrea Colaco, Henry Fuchs, and Ishan Chatterjee. EgoTrigger: Toward audio-driven image capture for human memory enhancement in all-day energy-efficient smart glasses. *IEEE Transactions on Visualization and Computer Graphics*, 2025. URL https://arxiv.org/abs/2508.01915.

Toby Perrett, Ahmad Darkhalil, Saptarshi Sinha, Omar Emara, Sam Pollard, Kranti Parida, Kaiting Liu, Prajwal Gatti, Siddhant Bansal, Kevin Flanagan, et al. HD-EPIC: A highly-detailed egocentric video dataset. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. URL https://arxiv.org/abs/2502.04144.

Chiara Plizzari, Gabriele Goletto, Antonino Furnari, Siddhant Bansal, Francesco Ragusa, Giovanni Maria Farinella, Dima Damen, and Tatiana Tommasi. An outlook into the future of egocentric vision. *International Journal of Computer Vision*, 132(11), 2024. URL https://arxiv.org/abs/2308.07123.

Shraman Pramanick, Yale Song, Sayan Nag, Kevin Qinghong Lin, Hardik Shah, Mike Zheng Shou, Rama Chellappa, and Pengchuan Zhang. EgoVLPv2: Egocentric video-language pre-training with fusion in the backbone. In *IEEE/CVF International Conference on Computer Vision*, 2023. URL https://arxiv.org/abs/2307.05463.

Kevin Pu, Ting Zhang, Naveen Sendhilnathan, Sebastian Freitag, Raj Sodhi, and Tanya R. Jonker. ProMemAssist: Exploring timely proactive assistance through working memory modeling in multi-modal wearable devices. In *Proceedings of the ACM Symposium on User Interface Software and Technology*, pp. 56:1–56:19, 2025. URL https://arxiv.org/abs/2507.21378.

Qwen Team. Qwen3-Omni technical report. *arXiv preprint arXiv:2509.17765*, 2025a. URL https://arxiv.org/abs/2509.17765.

Qwen Team. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025b. URL https://arxiv.org/abs/2511.21631.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5. Technical blog post.

Gorjan Radevski, Dusan Grujicic, Matthew Blaschko, Marie-Francine Moens, and Tinne Tuytelaars. Multimodal distillation for egocentric action recognition. In *IEEE/CVF International Conference on Computer Vision*, pp. 5190–5201, 2023. URL https://arxiv.org/abs/2307.07483.

Fadime Sener, Dibyadip Chatterjee, Daniel Shelepov, Kun He, Dipika Singhania, Robert Wang, and Angela Yao. Assembly101: A large-scale multi-view video dataset for understanding procedural activities. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21064–21074, 2022. URL https://arxiv.org/abs/2203.14712.

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. *arXiv preprint arXiv:2602.06139*, 2026. URL https://arxiv.org/abs/2602.06139.

Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, Zhuang Liu, Hu Xu, Hyunwoo J. Kim, Bilge Soran, Raghuraman Krishnamoorthi, Mohamed Elhoseiny, and Vikas Chandra. LongVU: Spatiotemporal adaptive compression for long video-language understanding. In *International Conference on Machine Learning*, 2025. URL https://arxiv.org/abs/2410.17434.

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

Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed. Learning audio-visual speech representation by masked multimodal cluster prediction. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2201.02184.

Alessandro Suglia, Claudio Greco, Katie Baker, Jose L. Part, Ioannis Papaioannou, Arash Eshghi, Ioannis Konstas, and Oliver Lemon. AlanaVLM: A multimodal embodied AI foundation model for egocentric video understanding. *arXiv preprint arXiv:2406.13807*, 2024. URL https://arxiv.org/abs/2406.13807.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. SALMONN: Towards generic hearing abilities for large language models. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.13289.

Yunlong Tang, Jing Bi, Siting Xu, Luchuan Song, Susan Liang, Teng Wang, Daoan Zhang, Jie An, Jingyang Lin, Rongyi Zhu, Ali Vosoughi, Chao Huang, Zeliang Zhang, Pinxin Liu, Mingqian Feng, Feng Zheng, Jianguo Zhang, Ping Luo, Jiebo Luo, and Chenliang Xu. Video understanding with large language models: A survey. *IEEE Transactions on Circuits and Systems for Video Technology*, 36(2):1355–1376, 2025. URL https://arxiv.org/abs/2312.17432.

Yixian Tian. How well can your video model remember? measuring memory-budget trade-offs in long video understanding. *arXiv preprint arXiv:2606.20726*, 2026. URL https://arxiv.org/abs/2606.20726.

Quang-Linh Tran, Hoang-Bao Le, Tuong-Nghiem Diep, Binh Nguyen, Gareth J. F. Jones, and Cathal Gurrin. OpenLifelogQA: An open-ended multi-modal lifelog question-answering dataset. In *International Symposium on Information and Communication Technology*, 2025. URL https://arxiv.org/abs/2508.03583.

Haoqin Tu, Jianwen Chen, Zijun Wang, Siwei Han, et al. VisualClaw: A real-time, personalized agent for the physical world. *arXiv preprint arXiv:2606.16295*, 2026. URL https://arxiv.org/abs/2606.16295.

Jian Wang, Rishabh Dabral, Diogo Luvizon, Zhe Cao, Lingjie Liu, Thabo Beeler, and Christian Theobalt. Ego4o: Egocentric human motion capture and understanding from multi-modal input. *arXiv preprint arXiv:2504.08449*, 2025. URL https://arxiv.org/abs/2504.08449.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pp. 9929–9939, 2020. URL https://arxiv.org/abs/2005.10242.

Xin Wang, Taein Kwon, Mahdi Rad, Bowen Pan, Ishani Chakraborty, Sean Andrist, Dan Bohus, Ashley Feniello, Bugra Tekin, Felipe Vieira Frujeri, Neel Joshi, and Marc Pollefeys. HoloAssist: an egocentric human interaction dataset for interactive AI assistants in the real world. In *IEEE/CVF International Conference on Computer Vision*, 2023.

Yanghai Wang, Jiahao Wang, Jiafu Tang, Yuanxing Zhang, Zhe Cao, Hanyan Bian, Zijie Zhang, Weiliang Luo, Zhiyu Pan, Zixuan Dong, Jiaheng Liu, and Zhaoxiang Zhang. AVSCap: Orchestrating audio-visual synergy for omni-modal video captioning. *arXiv preprint arXiv:2607.12820*, 2026a. URL https://arxiv.org/abs/2607.12820.

Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Chenting Wang, Guo Chen, Baoqi Pei, Ziang Yan, Rongkun Zheng, Jilan Xu, Zun Wang, Yansong Shi, Tianxiang Jiang, Songze Li, Hongjie Zhang, Yifei Huang, Yu Qiao, Yali Wang, and Limin Wang. InternVideo2: Scaling foundation models for multimodal video understanding. In *European Conference on Computer Vision*, 2024. URL https://arxiv.org/abs/2403.15377.

Yun Wang, Junbin Xiao, Han Lyu, Yifan Wang, Jing Zuo, Zhanjie Zhang, Hong Huang, Dapeng Wu, and Angela Yao. Keep it in mind: User centric continual spatial intelligence reasoning in egocentric video streams. In *International Conference on Machine Learning*, 2026b. URL https://arxiv.org/abs/2606.15200.

Keshara Weerasinghe, Xueren Ge, Tessa Heick, Lahiru Nuwan Wijayasingha, Anthony Cortez, Abhishek Satpathy, John Stankovic, and Homa Alemzadeh. EgoEMS: A high-fidelity multimodal egocentric dataset for cognitive assistance in emergency medical services. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026. URL https://arxiv.org/abs/2511.09894.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. LongVideoBench: A benchmark for long-context interleaved video-language understanding. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL https://arxiv.org/abs/2407.15754.

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 Datasets and Benchmarks Track*, 2025. URL https://arxiv.org/abs/2503.08221.

Boshen Xu, Ziheng Wang, Yang Du, Zhinan Song, Sipeng Zheng, and Qin Jin. Do egocentric video-language models truly understand hand-object interactions? In *International Conference on Learning Representations*, 2025a. URL https://arxiv.org/abs/2405.17719.

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*, 2025b. URL https://arxiv.org/abs/2503.20215.

Jiaqi Yan, Ruilong Ren, Jingren Liu, Shuning Xu, Ling Wang, Yiheng Wang, Xinlin Zhong, Yun Wang, Long Zhang, Xiangyu Chen, Changzhi Sun, et al. TeleEgo: Benchmarking egocentric AI assistants in the wild. *arXiv preprint arXiv:2510.23981*, 2025. URL https://arxiv.org/abs/2510.23981.

Antoine Yang, Arsha Nagrani, Paul Hongsuck Seo, Antoine Miech, Jordi Pont-Tuset, Ivan Laptev, Josef Sivic, and Cordelia Schmid. Vid2Seq: Large-scale pretraining of a visual language model for dense video captioning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. URL https://arxiv.org/abs/2302.14115.

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. *arXiv preprint arXiv:2503.03803*, 2025. URL https://arxiv.org/abs/2503.03803.

Hanrong Ye, Haotian Zhang, Erik A. Daxberger, Lin Chen, Zongyu Lin, Yanghao Li, Bowen Zhang, Haoxuan You, Dan Xu, Zhe Gan, Jiasen Lu, and Yinfei Yang. MM-Ego: Towards building egocentric multimodal LLMs for video QA. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.07177.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *National Science Review*, 2024. URL https://arxiv.org/abs/2306.13549.

Liping Yuan, Jiawei Wang, Haomiao Sun, Yuchen Zhang, and Yuan Lin. Tarsier2: Advancing large vision-language models from detailed video description to comprehensive video understanding. *arXiv preprint arXiv:2501.07888*, 2025. URL https://arxiv.org/abs/2501.07888.

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, Peng Jin, et al. VideoLLaMA 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025a. URL https://arxiv.org/abs/2501.13106.

Deheng Zhang, Yuqian Fu, Runyi Yang, Yang Miao, Tianwen Qian, Xu Zheng, Guolei Sun, Ajad Chhatkuli, Xuanjing Huang, Yu-Gang Jiang, Luc Van Gool, and Danda Pani Paudel. EgoNight: Towards egocentric vision understanding at night with a challenging benchmark. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2510.06218.

Hang Zhang, Xin Li, and Lidong Bing. Video-LLaMA: An instruction-tuned audio-visual language model for video understanding. In *Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2023. URL https://arxiv.org/abs/2306.02858.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. LLaVA-Video: Video instruction tuning with synthetic data. *Transactions on Machine Learning Research*, 2025b. URL https://arxiv.org/abs/2410.02713.

Yulin Zhang, Cheng Shi, Yang Wang, and Sibei Yang. Eyes wide open: Ego proactive video-LLM for streaming video. In *Advances in Neural Information Processing Systems*, 2025c. URL https://arxiv.org/abs/2510.14560.

Yuwei Zhang, Kumar Ayush, Siyuan Qiao, A. Ali Heydari, Girish Narayanswamy, Maxwell A. Xu, Ahmed A. Metwally, Shawn Xu, Jake Garrison, Xuhai Xu, Tim Althoff, et al. SensorLM: Learning the language of wearable sensors. *arXiv preprint arXiv:2506.09108*, 2025d. URL https://arxiv.org/abs/2506.09108.

Yue Zhao, Ishan Misra, Philipp Krähenbühl, and Rohit Girdhar. Learning video representations from large language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. URL https://arxiv.org/abs/2212.04501.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. MLVU: Benchmarking multi-task long video understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13691–13701, 2025a. URL https://arxiv.org/abs/2406.04264.

Wenqi Zhou, Kai Cao, Hao Zheng, Yunze Liu, Xinyi Zheng, Miao Liu, Per Ola Kristensson, Walterio Mayol-Cuevas, Fan Zhang, Weizhe Lin, and Junxiao Shen. X-LeBench: A benchmark for extremely long egocentric video understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 15206–15222, 2025b. URL https://arxiv.org/abs/2501.06835.

Xiaolin Zhou, Liu Liu, Tingyang Xiao, et al. HoloAgent-0: A unified embodied agent framework with 3d spatial memory. *arXiv preprint arXiv:2606.23565*, 2026. URL https://arxiv.org/abs/2606.23565.

Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, HongFa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, Wancai Zhang, Zhifeng Li, Wei Liu, and Li Yuan. LanguageBind: Extending video-language pretraining to N-modality by language-based semantic alignment. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.01852.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25589–25598, 2026. URL https://arxiv.org/abs/2602.14122.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025. URL https://arxiv.org/abs/2504.10479.

Heqing Zou, Tianze Luo, Guiyang Xie, Victor Zhang, Fengmao Lv, Guangcong Wang, Junyang Chen, Zhuochen Wang, Hansheng Zhang, and Huaijian Zhang. From seconds to hours: Reviewing multimodal large language models on comprehensive long video understanding. *arXiv preprint arXiv:2409.18938*, 2024. URL https://arxiv.org/abs/2409.18938.

APPENDIX

A Data Curation 18
A.1 Data Sources 18
A.2 Dataset Composition 19
A.3 Processing Pipeline 21
A.4 IMU Signal Characteristics 22
A.5 Label Generation 24
B The EgoMMU Benchmark 37
B.1 Benchmark Overview 37
B.2 Benchmark Statistics 38
B.3 Task Item Construction 39
B.4 Construct Validity 42
C Evaluation Settings & Protocols 43
C.1 Benchmark Statistics 43
C.2 Metrics & Judging 45
C.3 Grading Configurations & Prompts 47
D Model Architecture & Training 49
D.1 Problem Formulation 49
D.2 Model Architecture 49
D.3 Training Setup 50
D.4 Modality Handling 53
D.5 Streaming Memory 54
D.6 Frame-Budgeted Capture 56
D.7 Symbols & Notations 56
E Additional Experiments & Analyses 58
E.1 Complete Benchmark Results 58
E.2 Modality Dependence and Ablations 59
E.3 Encoder Pretraining 62
E.4 Qualitative Examples 66
F Prompts 84
F.1 Single-Modality Captioning 84
F.2 Multi-Modality Captioning 87
F.3 Benchmark Authoring 89
F.4 Grading 93
G Public Resources Used 97
G.1 Datasets 97
G.2 Models & Checkpoints 97
G.3 Proprietary Model Services 98
H Limitations & Broader Impact 99
H.1 Limitations 99
H.2 Broader Impact 99

A Data Curation

This section describes the construction of the EgoMM dataset and benchmark, including its data sources, composition, processing pipeline, IMU signal characteristics, and annotation procedure. We explain how egocentric recordings from different sources are organized into synchronized clips and annotated to support multimodal egocentric scene understanding.

A.1 Data Sources

Large-scale egocentric datasets (Grauman et al., 2022; 2024; Yang et al., 2025; Damen et al., 2022; Li et al., 2023; Sener et al., 2022) have supported substantial progress in first-person perception, as summarized in recent surveys (Plizzari et al., 2024; Li et al., 2026). These datasets cover diverse activities, participants, and environments, with some providing synchronized third-person views and detailed annotations of actions and interactions. Their recordings and annotations support both contrastive video-text pretraining (Lin et al., 2022; Pramanick et al., 2023; Zhao et al., 2023) and instruction tuning for egocentric understanding (Suglia et al., 2024; Ye et al., 2025; Xu et al., 2025a; Yang et al., 2023). We construct EgoMM from these existing resources, selecting recordings based on modality availability, activity coverage, and annotation quality. Human annotations provide supervision for events and activities, while audio and IMU signals contribute complementary information about sounds and wearer motion. Our data processing pipeline retains relevant source annotations and organizes the available modalities into a unified clip-level format.

Video-Centric Datasets. Existing egocentric datasets typically organize their annotations, benchmarks, and processing tools around video. Although many also record audio and IMU signals, the availability and format of these modalities vary across datasets. Audio may be unavailable or excluded from processed releases, while IMU measurements are often stored separately with device-specific sampling rates and coordinate systems. Additional processing is therefore required to use these modalities jointly for training and evaluation. We extract the available audio and IMU signals, align them with the corresponding video intervals, and standardize their representations. Each resulting sample associates visual observations with the sounds and motion recorded during the same interval, providing aligned inputs for multimodal learning.

IMU Signals and Language. Previous studies have explored mapping IMU signals to text for activity recognition and description (Moon et al., 2022; Imran et al., 2024; Ji et al., 2024; Zhang et al., 2025d; Li et al., 2025d; Narayanswamy et al., 2025; Haresamudram et al., 2024), generating descriptions from body motion (Chen et al., 2024d; Li et al., 2024; Wang et al., 2025), and aligning multiple modalities in a shared embedding space (Girdhar et al., 2023; Zhu et al., 2024; Han et al., 2024; Lu et al., 2024). These approaches demonstrate the potential of language supervision for learning semantic representations from motion signals. We extend this setting by combining IMU measurements with video and audio from the same recording interval. The dataset provides modality-specific descriptions and joint annotations that integrate the available information. This design supports learning from individual modalities and their combinations, allowing models to interpret wearer motion together with the corresponding visual and auditory context.

Primary Data Sources. The primary sources of EgoMM contain recordings captured using head-mounted Project Aria devices. Within each recording, video, audio, and measurements from two onboard IMUs are collected from the same wearer with synchronized timestamps. EgoLife (Yang et al., 2025) provides unscripted recordings of daily life, covering routine activities and interpersonal interactions over extended periods. Ego-Exo4D (Grauman et al., 2024) contributes recordings of skilled activities with detailed atomic-action annotations and expert commentary. These two datasets offer complementary coverage of everyday behavior and fine-grained task execution. Their synchronized sensor data and existing annotations provide the foundation for constructing multimodal clips and generating supervision for egocentric understanding.

Additional Data Sources. We incorporate additional datasets to increase diversity in devices, participants, activities, and modality coverage. HoloAssist (Wang et al., 2023) and Aria Everyday Activities (Lv et al., 2024) provide recordings from additional devices and participants. Nymeria (Ma et al., 2024) contributes motion-rich recordings, while HD-EPIC (Perrett et al., 2025) and IndEgo (Chavan et al., 2025) extend coverage to kitchen and industrial manipulation using Aria devices. Ego4D (Grauman et al., 2022) adds workplace and trade activities, and EgoAVU-Instruct (Seth et al., 2026)

Table 7: **Egocentric Capture Device Summary.** The specifications of typical devices used for multimodal egocentric scene understanding tasks, including different configurations and setups in terms of sensing using the Vision, Audio, and IMU modalities, respectively.

Specification	 Project Aria	 HoloLens 2	 Go Pro	 Insta360	 Phone
Form Factor	Glasses	Headset	Action Camera	360° Camera	Handset
Placement	Head	Head	Helmet or Chest	Helmet or Chest	Handheld or Chest
Sensor Time Base	One Clock, ns Resolution	Per Sensor, Aligned Offline	Independent, ±1% Drift	Not Published	Host Clock, No Guarantee
Raw Sensor Access	First-Party SDK	Research Mode API	Embedded Metadata	Vendor Tooling	OS API, Usually Unreleased
V Modality: Vision					
Scene Camera	RGB, 110°, 2880^2 at 30 fps	Photo/Video	Wide-Angle	Dual-Lens 360°	Wide, Ultra-Wide
Other Cameras	2× Mono SLAM, 150°, 640×480	4× Grayscale 640×480; Depth 512^2	None	None	Depth on Some Models
Eye Tracking	2× Camera, to 90 fps	Built In	None	None	None
A Modality: Audio					
Microphones	7× Spatial Array	Array	Built In	Built In	2–3
Sample Rate	48 kHz	Not Published	Not Published	Not Published	Not Published
I Modality: IMU					
Units	Two, Left and Right	One	One	One	One
Placement	Head, Bilateral	Head	Follows the Mount	Follows the Mount	Follows the Mount
Sample Rate	1000 / 800 Hz	Not Published	200 Hz; Gyro 400 Hz Pre-HERO6	Not Published	OS API, Not Guaranteed
Other Sensors	Magnetometer 10 Hz, Barometer 50 Hz, GPS	Magnetometer	Magnetometer (360° Only), GPS 10–18 Hz	Not Published	Magnetometer, Barometer, GPS

provides additional audio-visual supervision. We use only the modalities available in each source. For example, Nymeria (Ma et al., 2024) does not provide audio, and its IMU captions are generated from human action annotations filtered to retain motion-related information.

Recording Devices. The recording device determines the available modalities, synchronization properties, and sensor placement. Table 7 compares five device categories used for egocentric data collection. A shared hardware clock facilitates temporal alignment across modalities, whereas independently timestamped signals require additional alignment during processing. Sensor placement also determines the interpretation of IMU measurements: head-mounted sensors capture head motion, while sensors attached to other locations measure movement at those positions. These differences motivate our use of Aria recordings as the primary data sources and inform the processing of recordings from other devices. The table provides a broader comparison of device configurations, while the datasets incorporated into EgoMM are specified above.

A.2 Dataset Composition

Dataset Scale. EgoMM contains around 116,000 clips collected from the sources described in Section A.1. The training split consists of 1.75 million annotated items, corresponding to approximately 972.9 hours of recordings. Each clip spans 30 seconds, providing a consistent temporal window for organizing the available video, audio, and IMU signals. This duration captures local actions and their surrounding context while keeping the multimodal input manageable for training and inference.

Table 8: **Activity Coverage of the EgoMM Instruction Tuning Dataset.** Sources of multimodal egocentric data: EL EgoLife (Yang et al., 2025), EX Ego-Exo4D (Grauman et al., 2024), AE AEA (Lv et al., 2024), NY Nymeria (Ma et al., 2024), HE HD-EPIC (Perrett et al., 2025), IE IndEgo (Chavan et al., 2025), HA HoloAssist (Wang et al., 2023), E4 Ego4D (Grauman et al., 2022), EB EgoBlind (Xiao et al., 2025), EP EgoPolice (Gonzalez Saez-Diez et al., 2026). *Splits*: train denotes data used for model training, val denotes data used for model validation. *Statistics*: **Samples** the number of instruction-tuning items, **Clips** is the number of distinct recordings.

	Activity	Data Sources	Split	Statistics		Per-Modality Cues		
				Samples	Clips	V	A	I
Cooking & Eating								
	Cooking	EX E4 HE	train / val	127.1K	13.7K	Utensils	Sizzle/Timer	Looking Down
	Food Preparation	E4 HE EL	train / val	88.8K	5.5K	Ingredients	Chopping	Looking Down
	Making Coffee	HA HE EL	train / val	25.1K	1.3K	Machine	Machine	Stationary
	Eating	E4 EL NY	train / val	54.2K	4.1K	Food	Chewing	Stationary
	Total			**295.2K**	**24.6K**	≈ 202.8 h		
Daily Living								
	Cleaning	E4 NY HE	train / val	103.3K	6.1K	Objects	Ambient	Walking
	Laundry	E4 NY	train / val	30.6K	1.8K	Garments	Washer Hum	Looking Down
	Home Tour	NY E4 EL	train / val	30.0K	2.1K	Rooms	Footsteps	Walk & Scan
	Finding Item	NY E4 EL	train / val	31.3K	2.1K	Visual Scan	Rummaging	Head Turns
	Commuting	E4 EL NY	train / val	144.6K	9.7K	Route	Footsteps	Walking
	Total			**339.9K**	**21.8K**	≈ 179.4 h		
Making & Repair								
	Assembly & Disassembly	HA IE E4	train / val	298.9K	14.2K	Parts	Clicks	Looking Down
	Electronics Setup	HA IE E4	train / val	108.7K	5.2K	Device	Beeps	Looking Down
	Woodworking	E4 IE	train / val	51.0K	2.5K	Workpiece	Sawing	Looking Down
	Bike Repair	E4 EX	train / val	22.5K	2.6K	Frame	Tool Sound	Looking Down
	Appliance Repair	HA E4 IE	train / val	5.4K	459	Appliance	Tool Sound	Looking Down
	Auto Repair	E4	train	3.5K	189	Engine	Tools	Looking Down
	Construction	E4	train / val	12.3K	598	Site	Machinery	Walking
	Total			**502.2K**	**25.7K**	≈ 213.1 h		
Workplace & Industry								
	Office Work	E4 EL NY	train / val	82.7K	6.1K	Screen	Typing	Stationary
	Logistics	E4 EL IE	train / val	42.3K	2.6K	Warehouse	Scanner Beeps	Walking
	Industrial Assembly	IE E4	train / val	2.7K	201	Station	Machinery	Looking Down
	Total			**127.7K**	**8.9K**	≈ 73.4 h		
Sports & Fitness								
	Basketball	EX E4	train	10.0K	1.3K	Court	Bounce	Sprint/Jump
	Soccer	E4 EX	train / val	10.8K	1.2K	Field	Kicks	Sprint/Turn
	Climbing	EX E4	train / val	8.7K	1.3K	Holds	Gear Clinks	Head Up
	Workout	E4 NY EL	train / val	53.3K	3.4K	Gym	Breathing	Vigorous
	Total			**82.8K**	**7.3K**	≈ 60.7 h		
Social, Music & Play								
	Music	E4 EX EL	train / val	29.1K	3.4K	Instrument	Melody	Looking Down
	Dance	NY EX EL	train / val	23.9K	2.4K	Studio	Music	Rhythmic
	Games	E4 NY EL	train / val	115.2K	6.7K	Players	Chatter	Stationary
	Relaxing	E4 EL NY	train / val	114.4K	7.1K	Seated Scene	Ambient	Stationary
	Conversation	EL E4 AE	train / val	94.8K	7.8K	Faces	Speech/Tone	Stationary
	Total			**377.5K**	**27.3K**	≈ 224.6 h		
Health & Safety								
	First Aid	EX	train / val	1.6K	188	Procedure	Counting	Kneeling
	Health Test	EX E4	train / val	10.5K	2.1K	Test Kit	Spoken Steps	Looking Down
	Visual Assistance	EB	val	1.3K	1.2K	Surroundings	Spoken Query	Held Steady
	Incident Response	EP	val	12.0K	10.1K	Subjects	Raised Voices	Rapid Motion
	Total			**25.3K**	**13.6K**	≈ 18.9 h		
	Overall Statistics			**1.75M**	**116.7K**	≈ 972.9 h		

We use the same clip boundaries throughout data processing, training, and benchmark evaluation to maintain alignment between the sensor inputs and their annotations.

Supervision Format. Each clip supports multiple forms of supervision, including a separate caption for each available modality, a fused caption integrating the available modalities, and task-specific annotations for multiple-choice questions, fact checking, open-ended question answering, temporal grounding, and narration. Single-modality captions describe information supported by their respective inputs, while fused captions combine complementary observations across modalities. This organization supports both modality-specific and multimodal training: individual inputs can be paired with their corresponding captions, and combined inputs with annotations that require the supplied modalities. It also provides a consistent annotation structure for examining the contributions of individual modalities across different tasks.

Activity and Modality Coverage. Table 8 summarizes 32 activities grouped into seven domains, listing their data sources, use in training or evaluation, and representative visual, audio, and motion cues. The number of annotated items differs by more than an order of magnitude across domains, reflecting the availability of source recordings and our emphasis on sensor coverage rather than a balanced activity distribution. Modality availability also varies across sources. Every clip contains video, most include audio, and IMU measurements are included where they are recorded and released by the source dataset. We retain this variation and provide annotations according to the modalities available in each clip.

A.3 Processing Pipeline

Our processing pipeline consists of recording retrieval, temporal segmentation, sensor alignment and extraction, source annotation collection, modality-specific captioning, multimodal annotation generation, and validation. Each stage organizes the available inputs into a consistent clip-level format. This subsection describes the preprocessing of video, audio, and IMU signals, including temporal alignment, coordinate normalization, and sampling-rate adjustment. The captioning and annotation procedures are detailed in Section A.5.

A.3.1 Windowing & Synchronization

We extract each clip from a shared temporal interval of the source recording to preserve alignment across the available modalities. Cutting compressed video through stream copying can constrain the start time to a nearby keyframe. When keyframes are sparsely distributed, the resulting video interval may differ from the requested interval, introducing an offset relative to audio and IMU signals extracted using the specified timestamps. We therefore decode and re-encode the video during segmentation to obtain frame-accurate boundaries. Audio and IMU signals are extracted over the corresponding interval using the source timestamps.

A.3.2 IMU Normalization

Recording devices use different coordinate systems and sampling rates, requiring source-specific preprocessing before their IMU signals can be combined. We express the measurements in the Aria device coordinate system before feature extraction, using the device information summarized in Table 7. This convention provides a common reference for interpreting acceleration, angular velocity, and gravity-based head orientation. In particular, consistent axis definitions allow head-pitch features to be interpreted across recordings from different sources.

Recordings already expressed in the Aria coordinate system require no transformation. For Ego4D (Grauman et al., 2022), we apply a 90° rotation to align the source axes with the Aria convention, using the same rotation for acceleration and angular velocity. For HoloAssist (Wang et al., 2023), the current preprocessing applies a source-specific sign correction to one axis of both signals. We check the resulting head-pitch estimates as part of the validation described below. HD-EPIC (Perrett et al., 2025) provides per-frame device poses rather than raw IMU measurements. For this source, we estimate angular velocity from relative rotations between consecutive frames and their time intervals, and derive acceleration using a rigid-body model. These estimates retain the pose sampling rate and are identified as pose-derived signals wherever they are used.

We also standardize sampling rates where downsampling is required. Native rates vary by more than an order of magnitude across devices. Signals recorded above the target rate are downsampled with anti-aliasing filtering to suppress frequency components above the new Nyquist limit. This avoids the aliasing that can result from directly retaining every n-th sample. Signals recorded below the target rate retain their original sampling rate and are not upsampled. This treatment preserves the available temporal resolution without presenting interpolated values as additional measurements.

A.4 IMU Signal Characteristics

This subsection describes the IMU features used for annotation and model input, their mapping to discrete motion labels, and their variation across activities. We extract complementary features from acceleration and angular velocity to characterize motion amplitude, head orientation, and temporal dynamics. Representative windows are visualized using consistent axis limits to facilitate comparisons across activities.

A.4.1 Windowed Feature Extraction

We compute features over 2 s windows with 50% overlap. Signals recorded above 200 Hz are downsampled to this rate, while lower-rate signals retain their original sampling rate, following Section A.3. To estimate head orientation and linear motion separately, we apply a zero-phase second-order Butterworth low-pass filter with a cutoff frequency of 0.5 Hz:

$$\mathbf{g} = \mathrm{LPF}_{0.5}(\mathbf{a}), \qquad \mathbf{a}_\ell = \mathbf{a} - \mathbf{g}, \qquad m[n] = \|\mathbf{a}_\ell[n]\|, \tag{4}$$

where $\mathbf{a}$ is the measured acceleration, $\mathbf{g}$ estimates the gravity component, and m is the magnitude of the residual acceleration. This separation approximates gravity using the slowly varying component of the signal. Zero-phase filtering is used during offline preprocessing to avoid introducing a temporal shift between the estimated gravity and the original measurements.

A.4.2 Amplitude, Orientation & Rotation

We extract features describing motion amplitude, changes in acceleration magnitude, and head orientation:

$$a_{\mathrm{rms}} = \sqrt{\overline{m^2}}, \qquad j_{\mathrm{rms}} = \sqrt{\overline{\dot{m}^2}}, \qquad \mathrm{pitch} = \mathrm{atan2}(\bar{g}_z, -\bar{g}_y), \qquad \mathrm{roll} = \mathrm{atan2}(\bar{g}_x, -\bar{g}_y), \tag{5}$$

where the overline denotes averaging within a window and $\dot{m}$ is the temporal derivative of m. The first two quantities characterize motion intensity and abruptness, respectively. Pitch and roll are estimated from the window-mean gravity direction under the coordinate convention defined above. These estimates avoid drift from angular-velocity integration, although they remain sensitive to errors in the gravity estimate. From the gyroscope, we compute mean absolute angular velocities $\overline{|\omega_y|}$ and $\overline{|\omega_x|}$ to characterize turning and nodding, together with the overall magnitude $\omega_{\mathrm{rms}} = \sqrt{\overline{\|\boldsymbol{\omega}\|^2}}$. Integrated yaw is analyzed separately below.

A.4.3 Periodicity and Steps

We characterize repetitive motion using the dominant frequency and normalized autocovariance of the acceleration magnitude:

$$f_{\mathrm{dom}} = \underset{0.5 \le \nu \le 5}{\arg\max} |\hat{m}(\nu)|, \qquad r = \max_{0.2 \le \tau \le 1.5} \frac{\rho(\tau)}{\rho(0)}, \tag{6}$$

where $\hat{m}(\nu)$ is the spectrum of m at frequency ν in Hz, and $\rho(\tau)$ is its autocovariance at lag τ in seconds. The frequency and lag ranges are selected to capture common patterns of repetitive human motion. Higher values of r indicate stronger periodicity within the selected lag range, while lower values indicate weaker repetition. We estimate step counts from peaks in m that exceed an amplitude threshold and are separated by at least 0.3 s. The threshold is increased for low-amplitude windows to reduce detections caused by fluctuations in nearly stationary signals.

Dual-IMU Feature Aggregation. For recordings with two IMUs, we aggregate their features according to the quantity being measured. Rotation-invariant magnitudes are averaged across the two

Table 9: **Inertial Tag Vocabulary.** We consider five key motion aspects, each is a closed set of values with the condition that defines it. a_{rms} is the root-mean-square linear acceleration in m/s^2, f_{dom} the dominant motion frequency in Hz, r the motion regularity, $\dot{s}$ the step rate in steps per second, ω_{rms} the root-mean-square angular rate in rad/s, and pitch and roll the head angles in degrees.

Aspect	Value	Defining Condition
Locomotion	`Stationary`	$a_{\text{rms}} < 0.5$
	`Walking`	$1.5 \leq f_{\text{dom}} \leq 2.5, \dot{s} \geq 1.0, r > 0.4$
	`Running`	$2.5 < f_{\text{dom}} \leq 4.0, \dot{s} > 2.5, r > 0.3$
	`Light Motion`	$a_{\text{rms}} \geq 0.5$
	`Rhythmic Motion`	$a_{\text{rms}} > 3.0, f_{\text{dom}} > 0.5$
	`Vehicle`	$0.8 < a_{\text{rms}} < 1.5, f_{\text{dom}} < 1.0, \omega_{\text{rms}} < 0.3, r < 0.3$
Intensity	`Still`	$a_{\text{rms}} < 0.5$
	`Light`	$0.5 \leq a_{\text{rms}} < 1.5$
	`Moderate`	$1.5 \leq a_{\text{rms}} < 3.0$
	`Vigorous`	$a_{\text{rms}} \geq 3.0$
Head Pitch	`Looking Down`	$\text{pitch} > 20$
	`Looking Ahead`	$\lvert\text{pitch}\rvert \leq 20$
	`Looking Up`	$\text{pitch} < -20$
Head Roll	`Tilt Left`	$\text{roll} > 20$
	`Upright`	$\lvert\text{roll}\rvert \leq 20$
	`Tilt Right`	$\text{roll} < -20$
Head Motion	`Still`	$\bar{\omega}_y + \bar{\omega}_x < 0.1$
	`Slow Scanning`	$0.1 \leq \bar{\omega}_y + \bar{\omega}_x < 0.3$
	`Active Scanning`	$0.3 \leq \bar{\omega}_y + \bar{\omega}_x < 0.6$
	`Quick Turns`	$\bar{\omega}_y + \bar{\omega}_x \geq 0.6$

sensors to summarize overall motion. The larger of the two step-count estimates is retained to reduce missed detections when a peak is identified by only one sensor. Pitch and roll are taken from the left IMU, providing a consistent sensor reference for orientation features. Differences between the two sensors are retained separately for the analysis of bilateral signal variation.

A.4.4 Discrete Motion Labels

Continuous features are mapped to a predefined set of labels before being represented as text. Table 9 lists five feature groups: locomotion, intensity, head pitch, head roll, and head motion, together with their categories and thresholds. The same thresholds are applied across sources after coordinate normalization, providing consistent label definitions for annotation and evaluation. For each clip, we report the most frequent window-level label and the fraction of windows assigned to it, alongside the mean of the corresponding continuous feature. These summaries describe the dominant category and average feature value, respectively. For orientation features, the dominant label and the category assigned to the mean agree on most clips; disagreements occur in a small proportion of clips with mean values close to a category threshold.

Episode Segmentation. Window-level predictions can fluctuate over short intervals, producing fragmented motion descriptions. We therefore convert locomotion labels into two states, moving and still, and merge runs shorter than three windows into an adjacent run. The procedure is repeated until no short run remains or the sequence contains only one state. Algorithm 1 summarizes this process. To obtain non-overlapping intervals from overlapping analysis windows, each episode ends at the start of the next episode. The final episode ends at the clip boundary. This temporal smoothing reduces brief label changes while retaining longer transitions for caption generation and temporal annotation.

A.4.5 Activity-Specific Characteristics

Table 11, Table 13, and Table 12 illustrate complementary IMU patterns across activities. The examples highlight variations in motion amplitude, angular motion, and head orientation. Periodic

Algorithm 1 Episode Segmentation. Conversion of window-level locomotion labels into moving and still episodes by merging runs shorter than L windows.

Require: per-window labels $\ell_{1..N}$, window start times $t_{1..N}$, clip end time T, minimum run length $L = 3$

```
for i = 1 to N do
    c_i ← MOVING if ℓ_i denotes locomotion, else STILL
end for
R ← the maximal runs of equal values in c
while |R| > 1 and a run in R has length below L do
    ϱ ← the first run in R with length below L
    Assign all entries of ϱ the value of the preceding run, or the following run if ϱ is first
    R ← the maximal runs of equal values in c
end while
return one episode per run in R, from its first window start to the next run's first window start, or T for the final run
```

patterns are visible in activities such as walking, but are less apparent in others. We summarize these qualitative observations and their interpretation below:

- **Posture and Motion.** The cooking examples show sustained downward head orientation, whereas dancing exhibits more pronounced angular motion. These patterns illustrate how orientation and movement provide complementary descriptions of wearer behavior, without implying that either uniquely identifies an activity.
- **Bilateral Signal Differences.** The left and right IMUs can produce different measurements, motivating inspection of sensor-specific signals alongside aggregated features. These differences may reflect sensor placement and measurement characteristics; they should not be interpreted as independent movement of the two sides of the head.
- **Drift in Integrated Yaw.** Unlike gravity-based pitch and roll, cumulative yaw is obtained by integrating angular velocity and can accumulate sensor bias. We therefore use it to visualize relative turning patterns within a clip rather than as a precise heading estimate.

From our motion pattern visualizations, for each activity, all features are shown over the same representative 30 s interval. Each plot includes its own axes, with shared limits across rows within a column to support comparisons between activities. Signals are expressed in physical units under the unified device coordinate system. For visualization, higher-rate signals are downsampled to 100 Hz; feature computation uses the analysis rates described above. These examples illustrate characteristic patterns without representing the full variation within each activity.

A.5 Label Generation

This subsection describes how we generate single-modality captions and multimodal labels from synchronized clips and source annotations. Figure 5 summarizes the pipeline, including independent captioning, multimodal fusion, and the construction of training targets and benchmark items. We further examine the caption vocabulary across activities and describe the constraints used to maintain consistency between labels and their supporting evidence.

A.5.1 Single-Modality Captioning

We first generate captions independently for `Vision`, `Audio`, and `IMU` modalities. Each prompt produces a detailed natural-language description and structured fields in JSON format, using only the evidence available to the corresponding modality. The three prompts are provided in Section F.1. These captions serve as single-modality training targets, allowing each input stream to be supervised without requiring information available only from another sensor.

Modality-Specific Evidence. The captioning prompts specify both the available input and the descriptive perspective. Visual captions describe the observed scene and actions in the third person. Audio captions describe audible events without assigning them to the wearer unless the acoustic

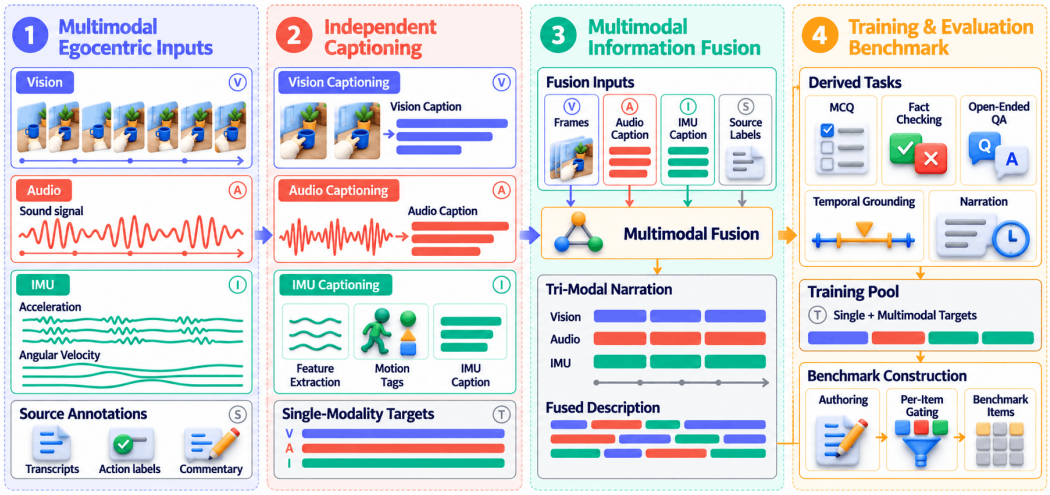


Figure 5: **Label Generation Pipeline.** Synchronized `Vision`, `Audio`, and `IMU` streams are captioned independently, with IMU captioning preceded by deterministic feature extraction and motion-tag generation. Multimodal fusion combines video frames, audio, and IMU captions, and available source annotations to produce a tri-modal narration; the independently generated video caption is retained as a training target rather than a fusion input. Single-modality captions, fused narrations, and derived task items provide training supervision. Benchmark items are constructed through a separate authoring process and undergo per-item quality checks.

evidence supports that attribution. IMU captions describe the wearer's motion in the first person, since the inertial measurements originate from the worn device. These conventions separate scene content, acoustic events, and wearer motion while maintaining a consistent format within each modality.

For IMU, captioning follows the deterministic feature extraction described in Section A.4. The extracted features are converted into the motion tags defined in Table 9, which constrain the resulting descriptions. Table 11, Table 13, and Table 12 illustrate the corresponding motion patterns through linear motion, angular motion, and head orientation, respectively. Together, these views show the signal properties underlying descriptions of movement intensity, turning, and head posture. They provide qualitative context for the captions without treating an inertial pattern as evidence of a specific object or activity.

Caption Vocabulary Across Activities. Table 10 presents word clouds for the three modalities across seven activity categories. The captions emphasize complementary information even when they describe the same clip. For cooking, visual captions frequently refer to a counter, board, or bowl; audio captions emphasize rustling, scraping, and noise; and IMU captions use terms such as downward, stationary, and brief. This difference reflects the evidence available to each sensor rather than alternative wording for a shared description.

The vocabulary also varies differently across activities. Comparing the top 25 terms between each pair of activity categories yields an average overlap of 9.5 terms for vision, 12.0 for audio, and 17.7 for IMU. Visual captions have lower overlap than IMU captions in all 21 category pairs, indicating greater activity specificity under this measure. The higher overlap for IMU is consistent with its constrained motion vocabulary: similar movements can occur across different activities, while their objects and environments differ. This shared vocabulary supports descriptions that remain tied to measurable motion properties.

A.5.2 Multimodal Fusion with Source Annotations

A straightforward baseline generates a caption for each modality and then merges the three captions with a language model. In our initial implementation, this approach introduced three recurring errors. First, transcript content was paraphrased even when a question required an exact quotation. Second, generic descriptions replaced source-provided entity references, making it difficult to associate participants and objects consistently across the resulting labels. Third, temporal answer spans became

misaligned with the annotated events. These errors were difficult to identify from fluency alone, since the generated descriptions remained linguistically plausible.

Fusion Inputs. We therefore incorporate source annotations directly into the fusion stage. As shown in Figure 5, the fusion model receives video frames, audio captions, IMU captions, and available source annotations, including transcripts, action labels, and expert commentary. The video caption remains a separate training target and is not used as a fusion input. Providing frames directly preserves visual details that may be omitted from a preliminary caption, while audio and IMU captions convey the evidence extracted from their respective streams.

The fusion stage produces a tri-modal narration that combines complementary observations while retaining their evidential basis. This narration supports the generation of multiple-choice questions, fact-checking items, open-ended questions, temporal grounding targets, and narration targets. The three single-modality captions remain available for modality-specific supervision, whereas the fused narration and derived items support training with the corresponding multimodal inputs.

Evidence Priorities. Source annotations provide references for fields with established ground truth, rather than additional text to be freely paraphrased during fusion. A transcript supports the exact wording of an utterance, but does not establish its emotional tone. Similarly, an action annotation supports the recorded event and its temporal extent, but does not independently verify every detail in a generated description. We therefore distinguish source-supported information from model inference at the field level.

This distinction also separates identity consistency from identity disclosure. Source-provided entity references help maintain correspondence across annotations, but the final labels use stable, observable descriptions rather than names or annotation-only identifiers. Likewise, source timestamps constrain temporal labels, while the resulting descriptions and answer spans must remain grounded in the presented clip. Source annotations guide generation without introducing information unavailable from the designated input modalities.

A.5.3 Label Constraints and Quality Control

We apply three constraints during label generation: audio descriptions must respect the intended acoustic scope, claims must be grounded in the available inputs, and participant identities must be anonymized. These constraints are specified in the prompts and assessed through targeted audits and consistency checks.

Acoustic Scope. The audio pathway targets sound-event understanding rather than speech transcription. Audio-supervised labels therefore describe acoustic properties rather than lexical content. They may describe speech presence, distinguishable voices, vocal tone, apparent distance, or language when supported by the audio, but exclude spoken words, conversational topics, and facts requiring utterance interpretation. Each audio-related phrase is reviewed for dependence on speech comprehension rather than acoustic evidence. In an audit of 218 clips, the rate of phrases containing speech content decreased from 56.5% to 1.8% after applying this rule, while the rate containing proper names decreased from 90% to 0%. Source transcripts support consistency checks but do not make lexical content an appropriate audio target.

Input Grounding. Every claim must be supported by the modalities supplied for its training target. Source annotations help identify and locate events but cannot justify facts absent from the clip or unavailable to the designated inputs. This applies to both individual captions and fused labels. Temporal descriptions refer to observable events within the clip, with answer spans constrained by source annotations when available.

Training Labels and Benchmark Verification. Training labels follow the prompt constraints above and undergo targeted audits rather than per-item model-based checks. Automated consistency checks are additionally applied where source references permit: retained transcript quotations must match the source verbatim, and temporal answers must fall within annotated action ranges. These checks assess source consistency without replacing the grounding and acoustic-scope requirements. Benchmark construction uses separate authoring prompts, provided in Section F.3, followed by per-item acceptance checks, since evaluation errors directly affect reported results. The acceptance thresholds, ablations, and agreement analyses are described in Section B.

Table 10: **Caption Vocabulary Across Activities.** Word clouds summarize the vocabulary of Vision, Audio, and IMU captions across seven activity categories, with larger words indicating higher frequency. The three modalities emphasize complementary information: visual objects and actions, acoustic events, and wearer motion.

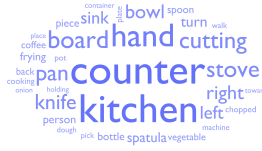
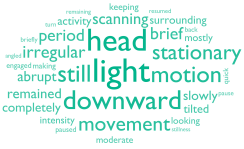
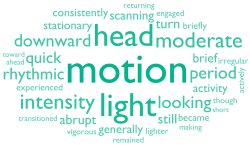
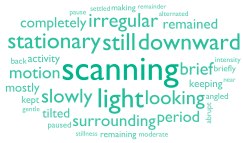
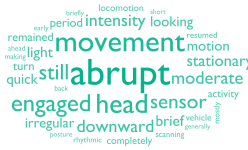
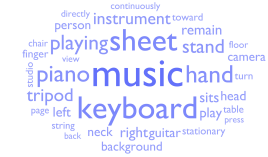
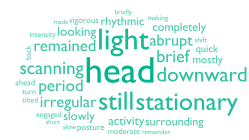
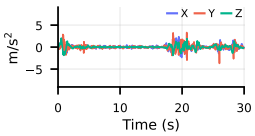
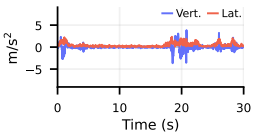
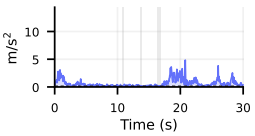
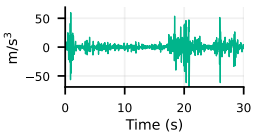
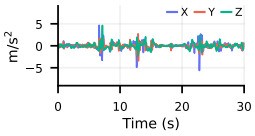
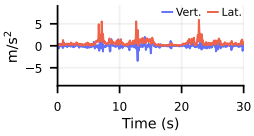
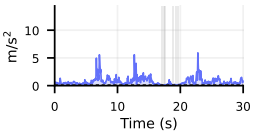
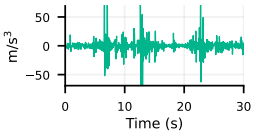
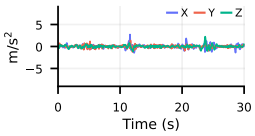
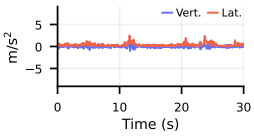
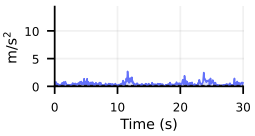
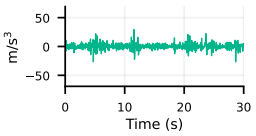
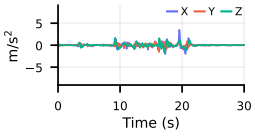
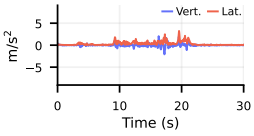
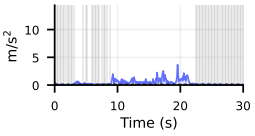
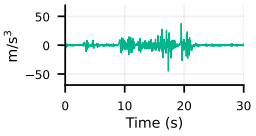
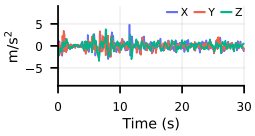
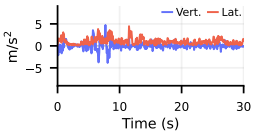
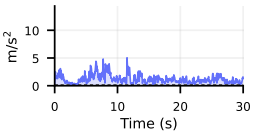
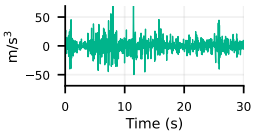
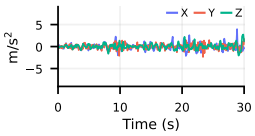
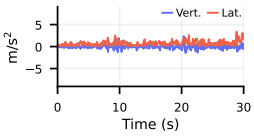
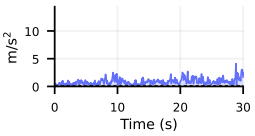
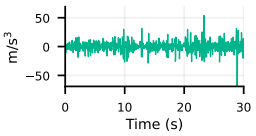
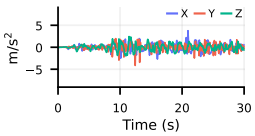
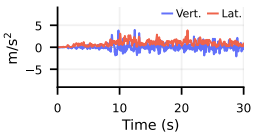
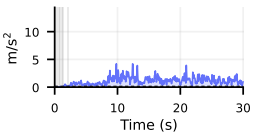
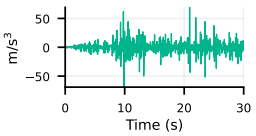
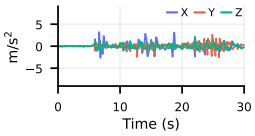
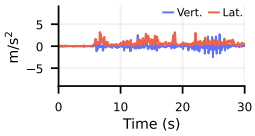
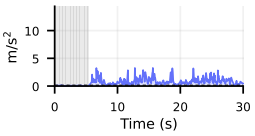
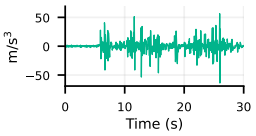
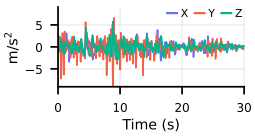
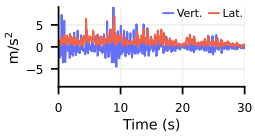
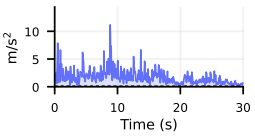
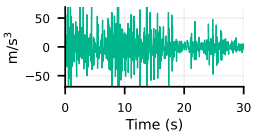
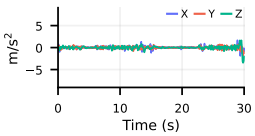
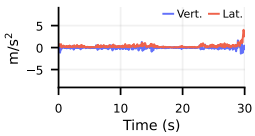
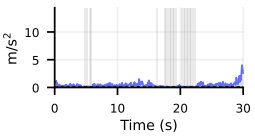
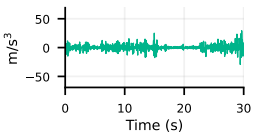
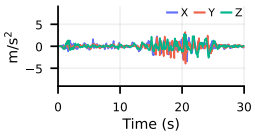
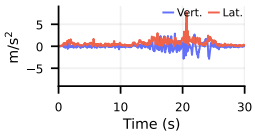
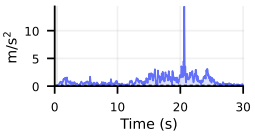
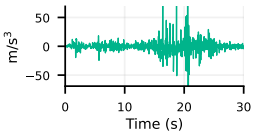
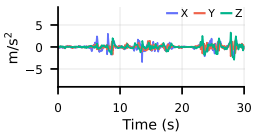
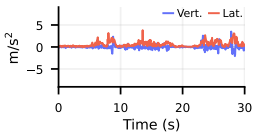
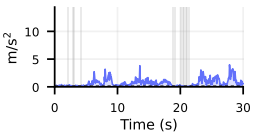
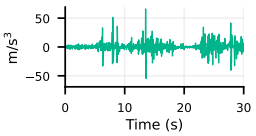
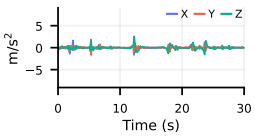
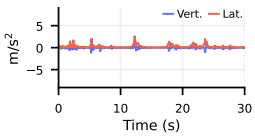
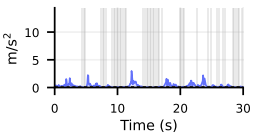
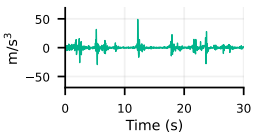
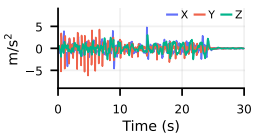
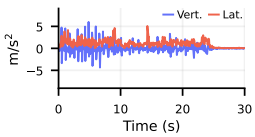
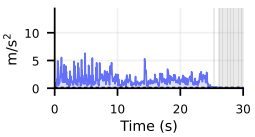
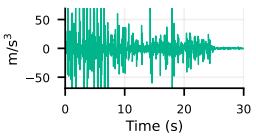
Activity	V Vision	A Audio	I IMU
Cooking & Eating			
Daily Living			
Making & Repair			
Workplace & Industry			
Sports & Fitness			
Social, Music & Play			
Health & Safety			

Table 11: **Head-Mounted Inertial Patterns: Motion.** Four views of body motion, one 30 s window per activity. "**Linear Acceleration**" is the three device axes with gravity removed; "**Bounce vs. Sway**" splits the same signal into the components along and across gravity; "**Motion Magnitude**" is its norm, shaded where a sub-window falls below a fixed stillness threshold; and "**Jerk**" is its time derivative. Kindly refer to our project page for animation examples.

Linear Acceleration	Bounce vs. Sway	Motion Magnitude	Jerk
Cooking & Eating: Cooking			
			
Cooking & Eating: Food Preparation			
			
Cooking & Eating: Making Coffee			
			
Cooking & Eating: Eating			
			
Daily Living: Cleaning			
			
Daily Living: Laundry			
			
Daily Living: Home Tour			
			

Continued on the next page.

Table 11 continued from the previous page.

Linear Acceleration	Bounce vs. Sway	Motion Magnitude	Jerk
Daily Living: Finding Item			
			
Daily Living: Commuting			
			
Making & Repair: Assembly & Disassembly			
			
Making & Repair: Electronics Setup			
			
Making & Repair: Woodworking			
			
Workplace & Industry: Office Work			
			
Workplace & Industry: Logistics			
			

Continued on the next page.

Table 11 continued from the previous page.

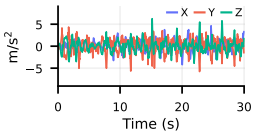
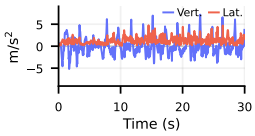
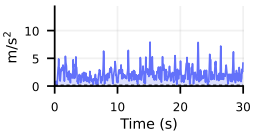
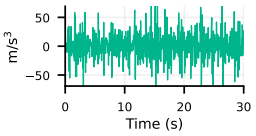
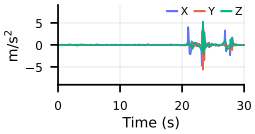
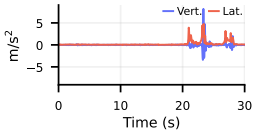
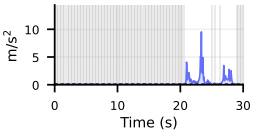
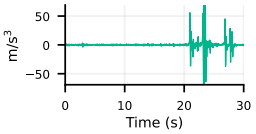
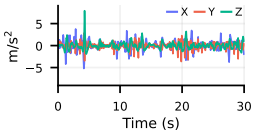
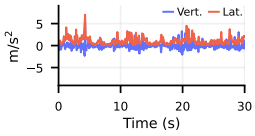
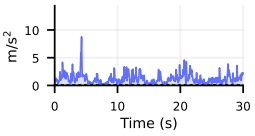
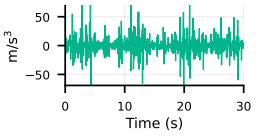
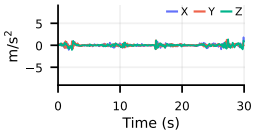
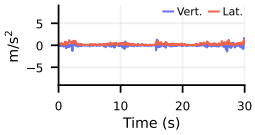
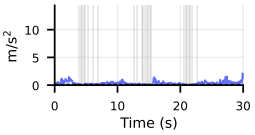
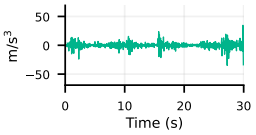
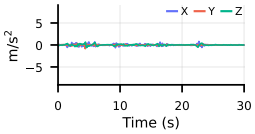
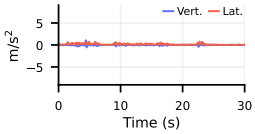
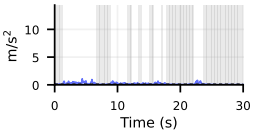
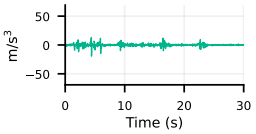
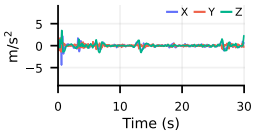
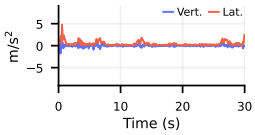
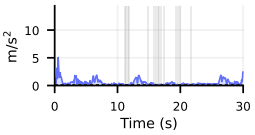
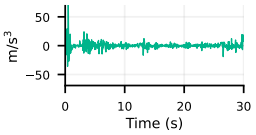
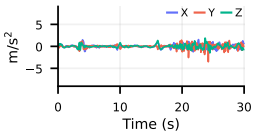
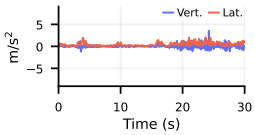
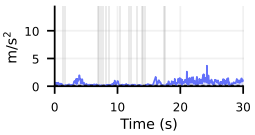
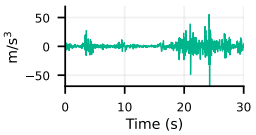
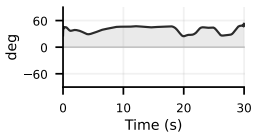
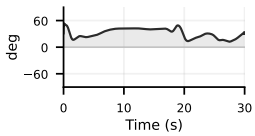
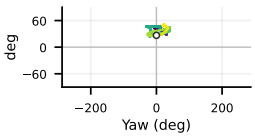
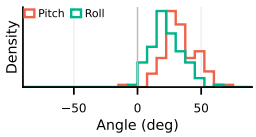
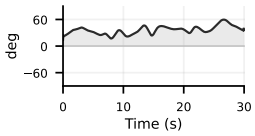
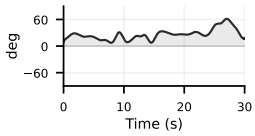
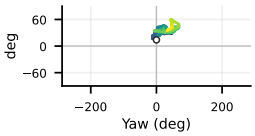
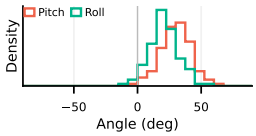
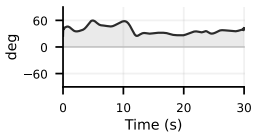
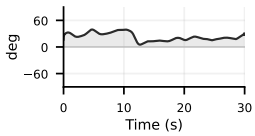
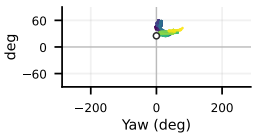
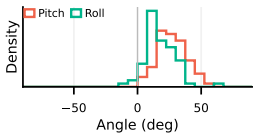
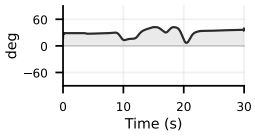
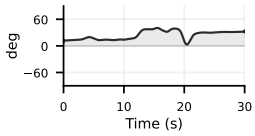
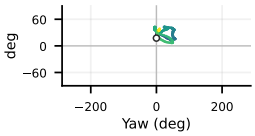
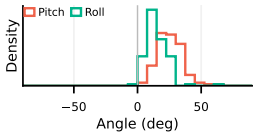
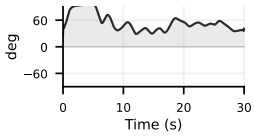
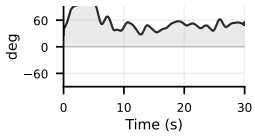
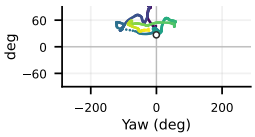
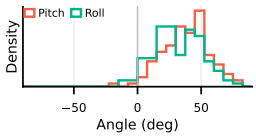
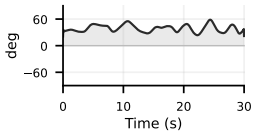
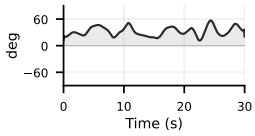
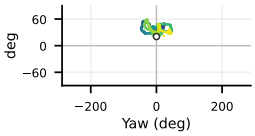
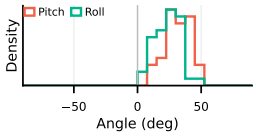
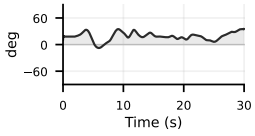
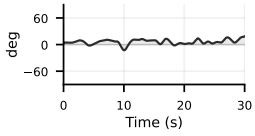
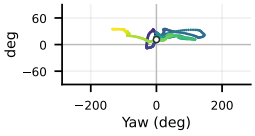
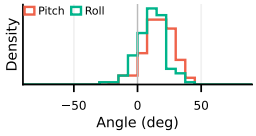
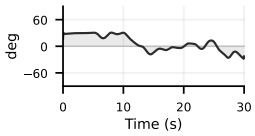
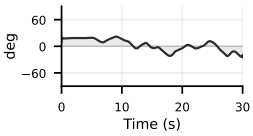
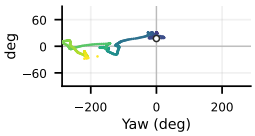
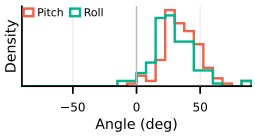
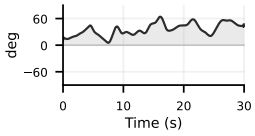
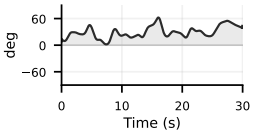
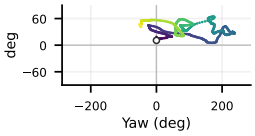
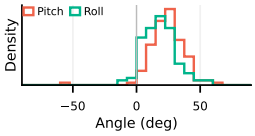
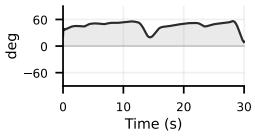
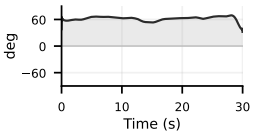
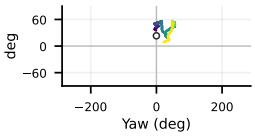
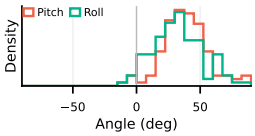
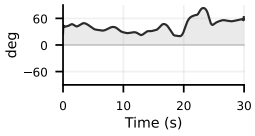
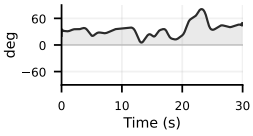
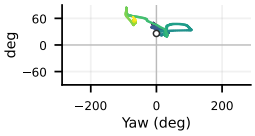
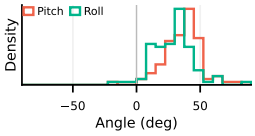
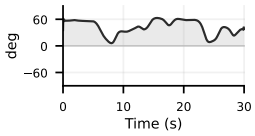
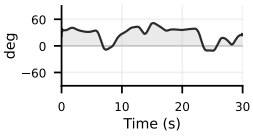
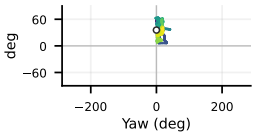
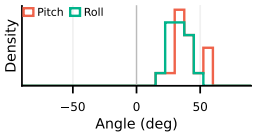
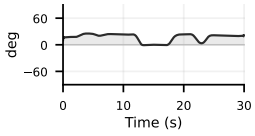
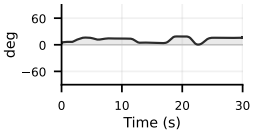
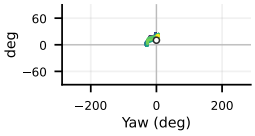
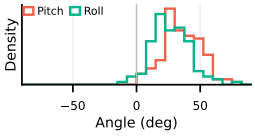
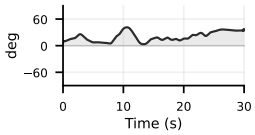
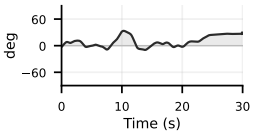
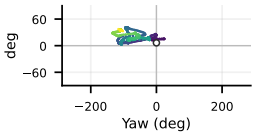
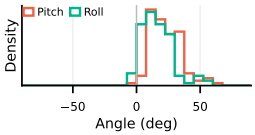
Linear Acceleration	Bounce vs. Sway	Motion Magnitude	Jerk
Sports & Fitness: Workout			
			
Social, Music & Play: Music			
			
Social, Music & Play: Dance			
			
Social, Music & Play: Games			
			
Social, Music & Play: Relaxing			
			
Social, Music & Play: Conversation			
			
Health & Safety: Visual Assistance			
			

Table 12: **Head-Mounted Inertial Patterns: Orientation.** Four views of head orientation, one 30 s window per activity. **Head Pitch** and **Head Roll** are recovered from the gravity direction; **Head Direction Path** traces the two against integrated yaw; and **Pitch & Roll Distribution** pools both angles over every candidate clip of the activity rather than over the single clip shown. Kindly refer to our project page for animation examples.

Head Pitch	Head Roll	Head Direction Path	Pitch & Roll Dist.
Cooking & Eating: Cooking			
			
Cooking & Eating: Food Preparation			
			
Cooking & Eating: Making Coffee			
			
Cooking & Eating: Eating			
			
Daily Living: Cleaning			
			
Daily Living: Laundry			
			
Daily Living: Home Tour			
			

Continued on the next page.

Table 12 continued from the previous page.

Head Pitch	Head Roll	Head Direction Path	Pitch & Roll Distribution
Daily Living: Finding Item			
			
Daily Living: Commuting			
			
Making & Repair: Assembly & Disassembly			
			
Making & Repair: Electronics Setup			
			
Making & Repair: Woodworking			
			
Workplace & Industry: Office Work			
			
Workplace & Industry: Logistics			
			

Continued on the next page.

Table 12 continued from the previous page.

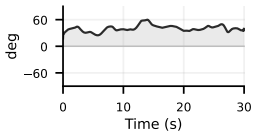
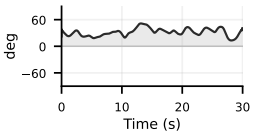
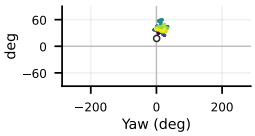
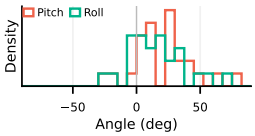
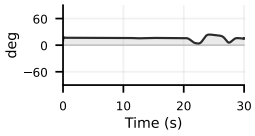
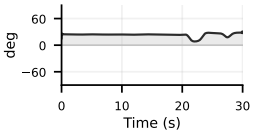
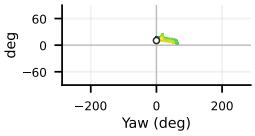
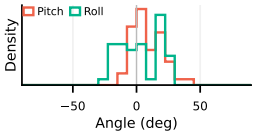
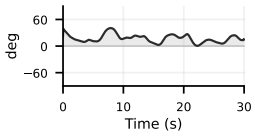
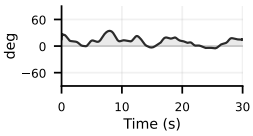
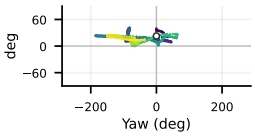
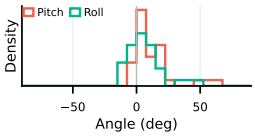
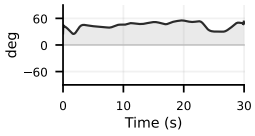
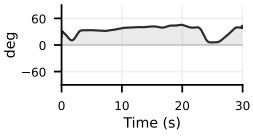
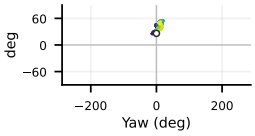
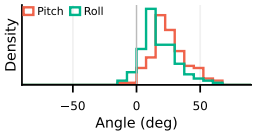
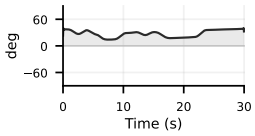
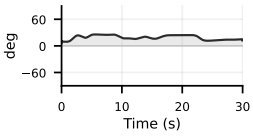
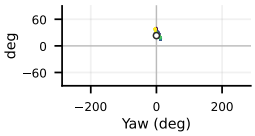
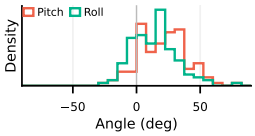
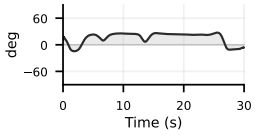
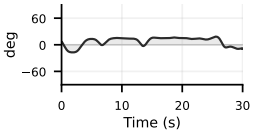
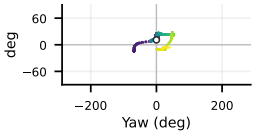
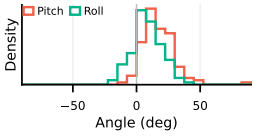
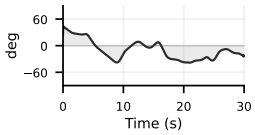
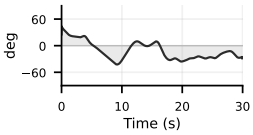
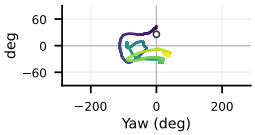
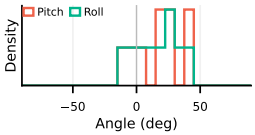
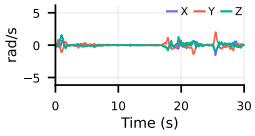
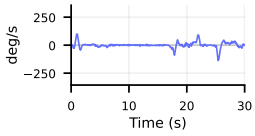
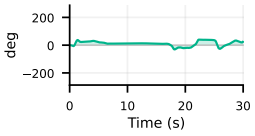
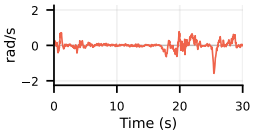
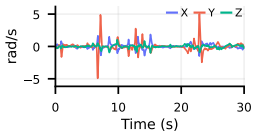
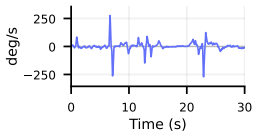
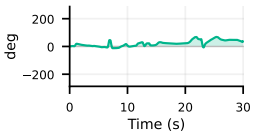
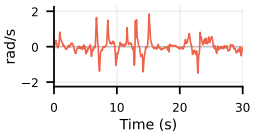
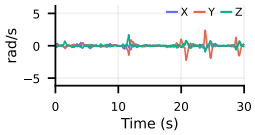
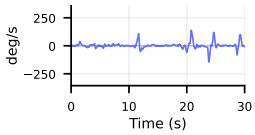
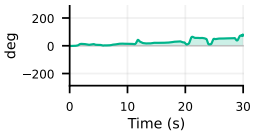
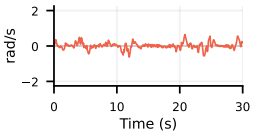
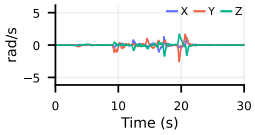
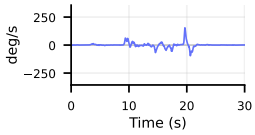
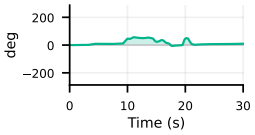
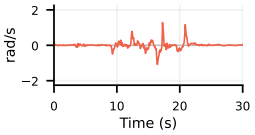
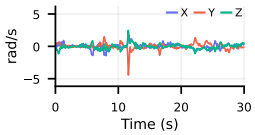
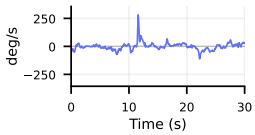
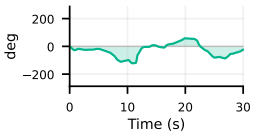
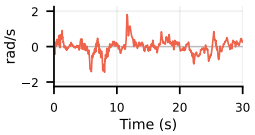
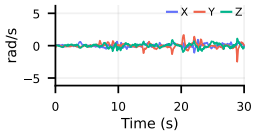
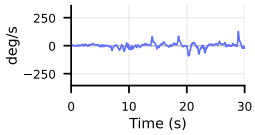
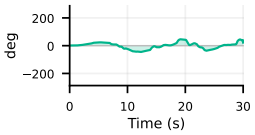
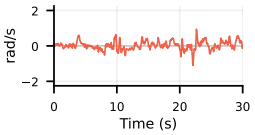
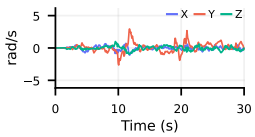
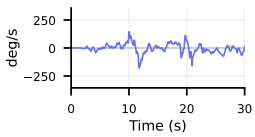
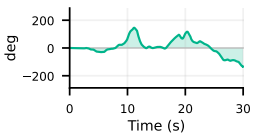
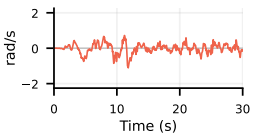
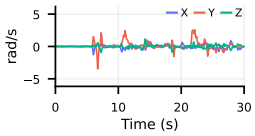
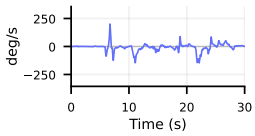
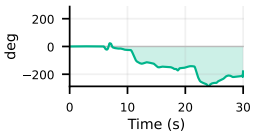
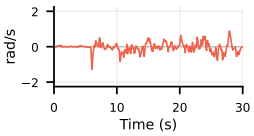
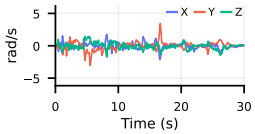
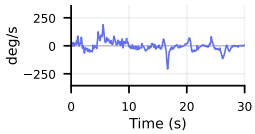
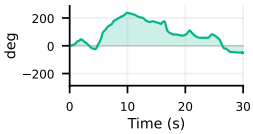
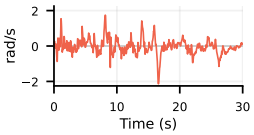
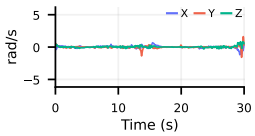
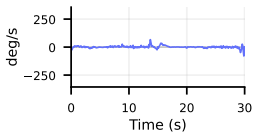
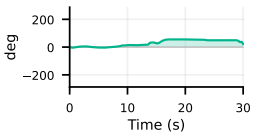
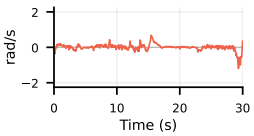
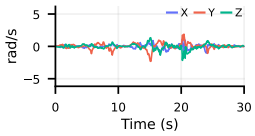
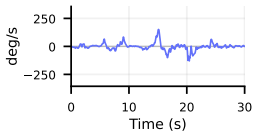
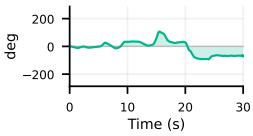
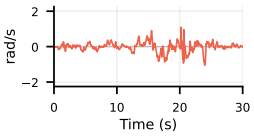
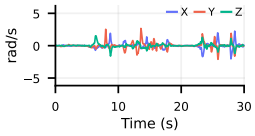
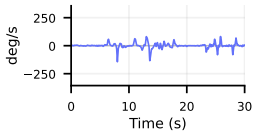
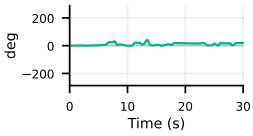
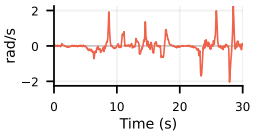
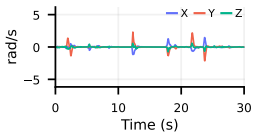
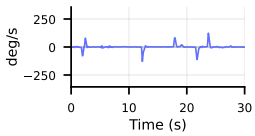
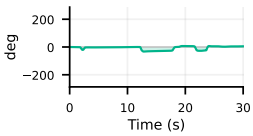
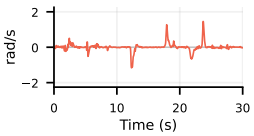
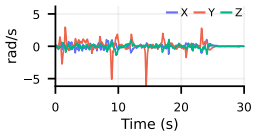
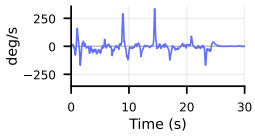
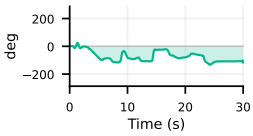
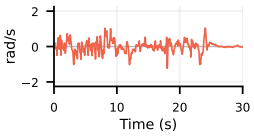
Head Pitch	Head Roll	Head Direction Path	Pitch & Roll Distribution
Sports & Fitness: Workout			
			
Social, Music & Play: Music			
			
Social, Music & Play: Dance			
			
Social, Music & Play: Games			
			
Social, Music & Play: Relaxing			
			
Social, Music & Play: Conversation			
			
Health & Safety: Visual Assistance			
			

Table 13: **Head-Mounted Inertial Patterns: Rotation.** Four views of angular rate, one 30 s window per activity. **Gyroscope** is the three device axes; **Turn Rate** is the component about the gravity axis; **Cumulative Yaw Sweep** is that component integrated from the start of the window; and **Nod Rate** is the component about the pitch axis. Kindly refer to our project page for animation examples.

Gyroscope	Turn Rate	Cumulative Yaw Sweep	Nod Rate
Cooking & Eating: Cooking			
			
Cooking & Eating: Food Preparation			
			
Cooking & Eating: Making Coffee			
			
Cooking & Eating: Eating			
			
Daily Living: Cleaning			
			
Daily Living: Laundry			
			
Daily Living: Home Tour			
			

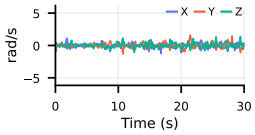
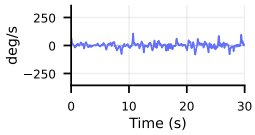
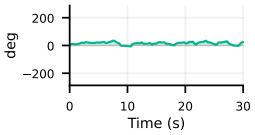
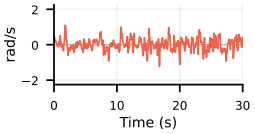
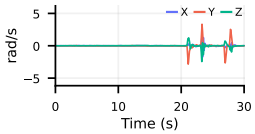
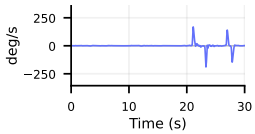
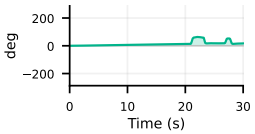
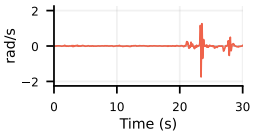
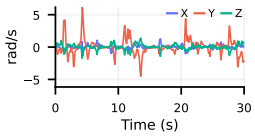
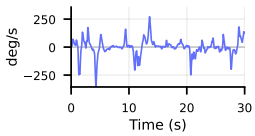
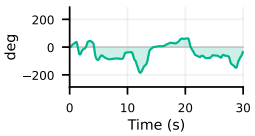
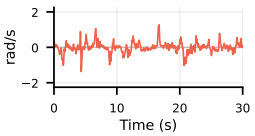
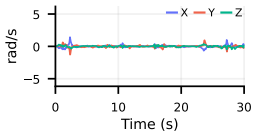
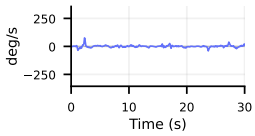
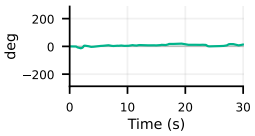
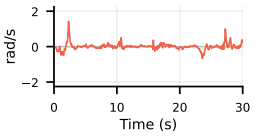
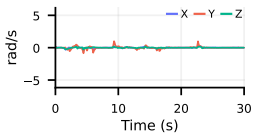
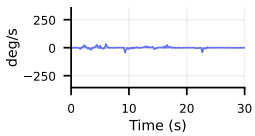
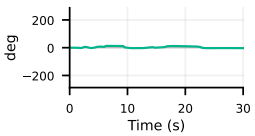
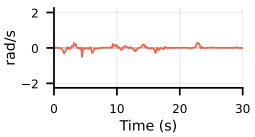
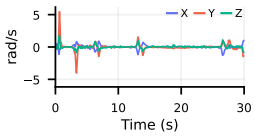
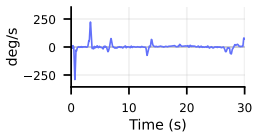
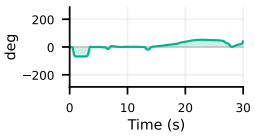
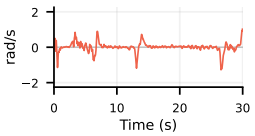
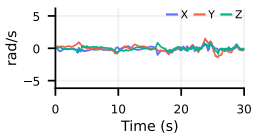
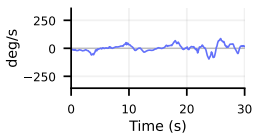
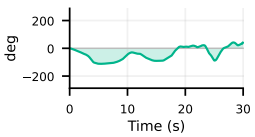
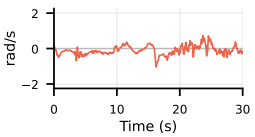
Continued on the next page.

Table 13 continued from the previous page.

Gyroscope	Turn Rate	Cumulative Yaw Sweep	Nod Rate
Daily Living: Finding Item			
			
Daily Living: Commuting			
			
Making & Repair: Assembly & Disassembly			
			
Making & Repair: Electronics Setup			
			
Making & Repair: Woodworking			
			
Workplace & Industry: Office Work			
			
Workplace & Industry: Logistics			
			

Continued on the next page.

Table 13 continued from the previous page.

Gyroscope	Turn Rate	Cumulative Yaw Sweep	Nod Rate
Sports & Fitness: Workout			
			
Social, Music & Play: Music			
			
Social, Music & Play: Dance			
			
Social, Music & Play: Games			
			
Social, Music & Play: Relaxing			
			
Social, Music & Play: Conversation			
			
Health & Safety: Visual Assistance			
			

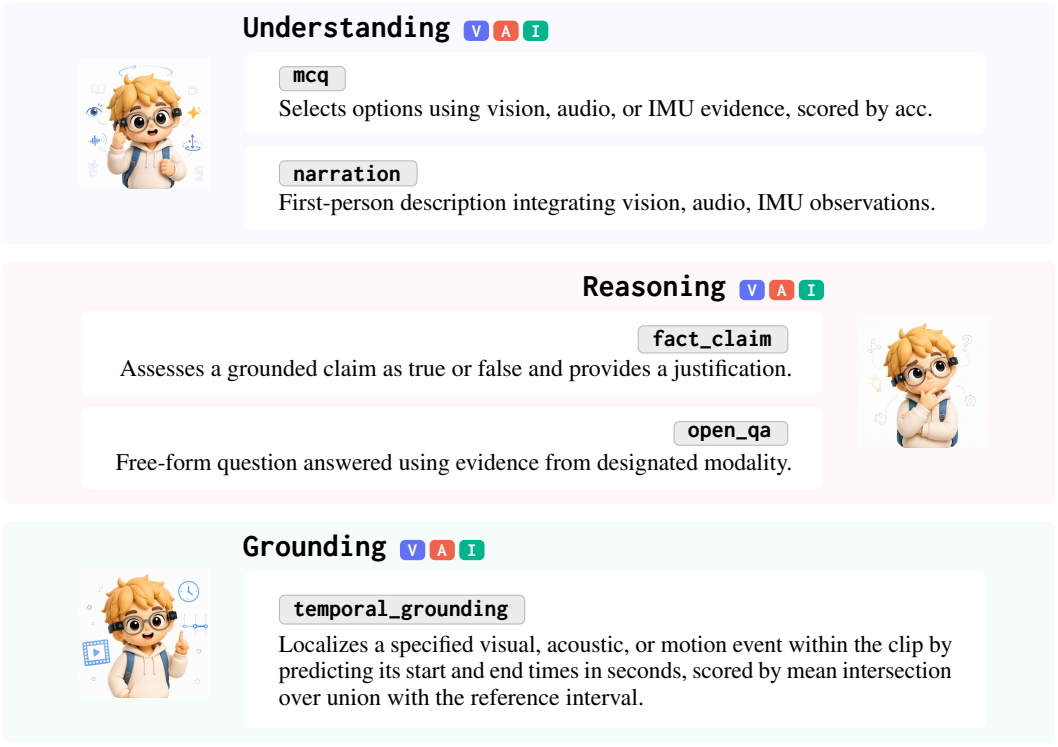
B THE EGOMMU BENCHMARK

This section describes the task structure, modality coverage, and dataset statistics of EgoMMU, followed by the task item construction process and quality control procedures. We also examine potential shortcuts, temporal priors, and modality dependence to assess whether the benchmark measures the intended capabilities.

B.1 BENCHMARK OVERVIEW

Table 14 compares modality coverage and annotation scope across instruction-tuning datasets and evaluation benchmarks. Existing resources establish several foundations for multimodal egocentric understanding: MMG-Ego4D (Gong et al., 2023) combines video, audio, and IMU for action recognition and modality generalization; EgoLifeQA (Yang et al., 2025) identifies questions requiring audio in long-horizon recordings; and EgoAVU-Bench (Seth et al., 2026) and EgoSound (Zhu et al., 2026) evaluate acoustic and audio-visual capabilities. These resources differ in both the streams they use and the forms of understanding their tasks assess.

Within this landscape, EgoMMU combines visual, acoustic, and IMU evidence with shared language-based task formats and explicit modality requirements. This organization supports comparisons across modalities within a task family and across task formats within a modality. Recording the required evidence also allows performance to be examined under controlled input withholding, even when the evaluated model receives additional streams. Item construction and the analyses in Sections B.3 and B.4 support this connection between task performance and the evidence each item is intended to assess.



EgoMMU comprises **five task families**: multiple-choice questions, narration, fact checking, open-ended question answering, and temporal grounding. Each family except narration is evaluated separately for vision, audio, and IMU, while narration integrates the available modalities into a single description, yielding 13 evaluation columns. The modality assigned to a column identifies the evidence required to answer its questions; the inputs supplied to a model depend on its supported modalities and the evaluation setting. For example, an audio multiple-choice item may be evaluated with video and audio available, while its answer still requires acoustic evidence.

Table 14: **Comparison of Egocentric Datasets & Benchmarks.** Modality coverage and annotation scope of representative resources. Modality columns indicate inputs used for the listed supervision or evaluation, rather than all sensors present in the source recordings. "✓": included; "✗": not included.

Dataset / Benchmark	Venue	V	A	I	Supervision / Evaluation Scope
Instruction-Tuning Datasets					
MM-Ego (Ye et al., 2025)	ICLR'25	✓	✗	✗	Egocentric video question answering
EgoIT-99K (Yang et al., 2025)	CVPR'25	✓	✓	✗	Visual and audio-visual captioning and QA
EgoAVU-Instruct (Seth et al., 2026)	CVPR'26	✓	✓	✗	Audio-visual narration and QA
EgoMM-Instruct	**Ours**	✓	✓	✓	Modality-specific and fused descriptions
Evaluation Benchmarks					
MMG-Ego4D (Gong et al., 2023)	CVPR'23	✓	✓	✓	Action recognition and modality generalization
EgoSchema (Mangalam et al., 2023)	NeurIPS'23	✓	✗	✗	Long-form and diverse visual-only multiple-choice QA
EgoLifeQA (Yang et al., 2025)	CVPR'25	✓	✓	✗	Long-horizon QA with audio-required labels
EgoAVU-Bench (Seth et al., 2026)	CVPR'26	✓	✓	✗	Audio-visual narration, reasoning, grounding
EgoSound (Zhu et al., 2026)	CVPR'26	✓	✓	✗	Sound perception and audio-visual reasoning
EgoMMU	**Ours**	✓	✓	✓	Five task families with explicit visual, acoustic, and IMU evidence requirements

The task families assess complementary aspects of egocentric understanding, from selecting an answer and describing observations to verifying claims and locating events in time. Multiple-choice questions use five options, with a chance accuracy of 20%. Fact checking evaluates both the verdict and its justification, while open-ended question answering and narration assess agreement with the reference answer or description. Temporal grounding measures the overlap between predicted and reference intervals. Metric definitions and grading protocols are provided in Section C.2.

The task formats are shared across modalities, while their content reflects the evidence each stream provides. Audio multiple-choice items, for example, include per-utterance speech-tone questions that assess how a statement is spoken rather than its lexical content. Figure 7 illustrates representative evidence requirements, from recognizing visual details and distinguishing sounds to describing wearer motion and locating events in time. Qualitative examples with model predictions from all 13 evaluation columns are provided in Section E.4.

B.2 Benchmark Statistics

EgoMMU contains 3,762 items from 1,196 distinct clips, drawn from eight datasets: EgoLife (Yang et al., 2025), Ego-Exo4D (Grauman et al., 2024), Ego4D (Grauman et al., 2022), Nymeria (Ma et al., 2024), HoloAssist (Wang et al., 2023), AEA (Lv et al., 2024), HD-EPIC (Perrett et al., 2025), and IndEgo (Chavan et al., 2025). All items use the 30 s windows described in Section A.2. Figure 6 illustrates the five task families, which together form 13 evaluation columns spanning visual, acoustic, and motion understanding.

Task and Source Coverage. Each column contains between 193 items for IMU open-ended question answering and 401 items for vision fact checking, supported by 178 to 370 distinct clips. A clip may contribute multiple items that examine different observations or task requirements, so item counts and clip counts describe complementary aspects of coverage. Narration contains 370 items, each associated with a distinct clip. Source composition varies across columns according to the availability and suitability of the corresponding modality. For example, Nymeria contributes no audio items, while HoloAssist contributes no IMU items to the benchmark. The columns therefore share a common clip duration and task structure without requiring identical source distributions.

Answer Distributions. The closed-form tasks use approximately balanced answer distributions. In each multiple-choice column, no answer position accounts for more than 24.5% of the items,

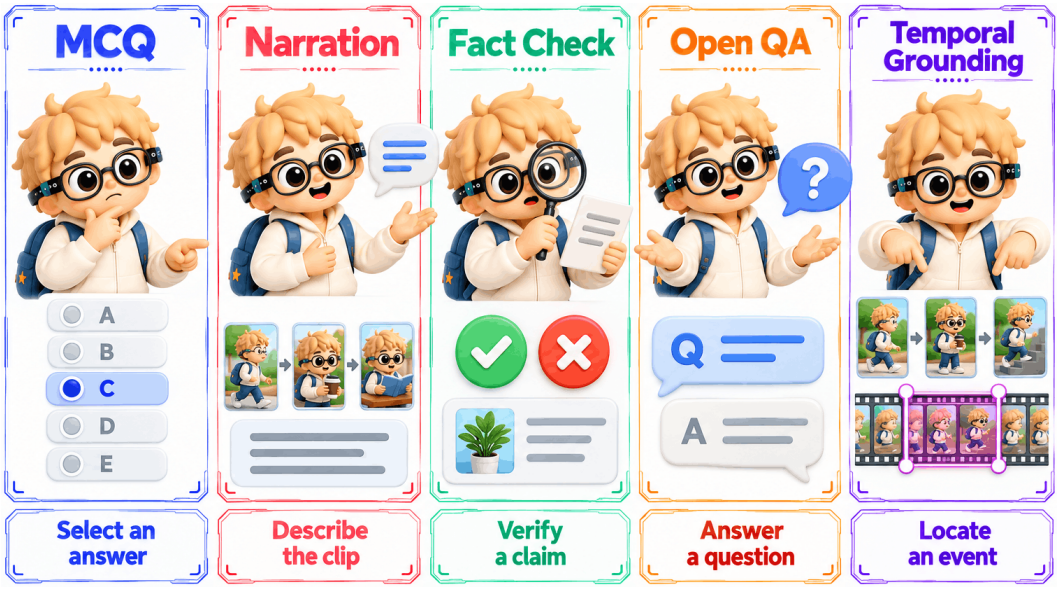


Figure 6: **Task Families of EgoMMU.** The benchmark covers multiple-choice questions, narration, fact checking, open-ended question answering, and temporal grounding. Narration integrates available modalities, while the other four families are evaluated separately for vision, audio, and IMU.

compared with 20% under a uniform distribution over five options. The vision fact-checking column contains 207 true and 194 false claims, while the audio and IMU columns are balanced at 200/200 and 122/122, respectively. These distributions limit the advantage of consistently selecting a frequent option or verdict. Fact-checking items additionally require a supporting explanation, whose quality contributes to the composite score defined in Section C.2.

Temporal Coverage. The vision, audio, and IMU temporal-grounding columns contain 201, 262, and 264 items, respectively. Their median target durations are 4.0 s, 2.8 s, and 2.9 s within the shared 30 s clip window. The target events thus typically occupy a limited portion of the available context, requiring models to identify when an event occurs in addition to recognizing its presence. Since target durations differ across modalities, we report temporal-grounding results separately and interpret them alongside the fixed-span baselines described in Section C.2.

B.3 Task Item Construction

We construct EgoMMU through source-clip selection, task item authoring, and quality control. The process combines clip media, source annotations, and signal-derived information to generate candidates, followed by automated checks and human verification. These checks assess whether each item is well formed, supported by the available evidence, and appropriate for its designated task and modality.

B.3.1 Source Clips

Benchmark clips are drawn from recording groups held out from our training data. For most sources, we assign groups, such as sessions or recordings, to the evaluation split by hashing their identifiers, selecting approximately one in five groups. We use source-specific partitions where the recording structure supports a more suitable separation. For EgoLife (Yang et al., 2025), we hold out the first recording day of every participant. Nymeria (Ma et al., 2024) uses a fixed list of held-out sequences, while HD-EPIC (Perrett et al., 2025) holds out two complete kitchens from its nine kitchens.

We define these partitions before constructing benchmark items and apply them at the recording-group level rather than independently to individual clips. All clips from a held-out group remain outside the training split, preventing neighboring windows from the same recording group from appearing in both training and evaluation. The selected recordings are processed into the same 30 s windows used throughout the dataset, maintaining a consistent temporal context across task families.



Figure 7: **Illustrative Use Cases of EgoMMU.** Examples of visual detail recognition, action ordering, sound-source identification, voice discrimination, wearer motion and posture, temporal localization, and multimodal narration. Each row pairs a question or claim with a schematic illustration of the relevant evidence. Scenes and signals are illustrative rather than actual benchmark samples.

B.3.2 Item Authoring

Task items are generated through prompt-based authoring, with their evidence supplied by the clip media, available source annotations, and signal-derived features. The authoring prompts specify

the task format, designated modality, and required output fields, as provided in Section F.3. Source annotations help identify relevant events and their temporal context, while the resulting questions and reference answers must remain grounded in the inputs available for the task.

Signal-Derived Supervision. For IMU multiple choice, fact checking, and open-ended question answering, we first extract the features described in Section A.4 and convert them into structured motion descriptions using predefined rules. These descriptions provide the evidence and answer constraints for the authoring prompt, which generates the corresponding questions, options, claims, and reference answers. The rules establish the motion attributes supported by the signal, while prompt-based generation expresses them in the required task format. The resulting references therefore follow signal-derived constraints rather than unrestricted model interpretation.

Temporal grounding also uses signal-derived information where applicable. Acoustic events and level transitions in the waveform provide candidate audio intervals, while transitions in the IMU signal provide candidate motion intervals. These intervals guide the generation of event descriptions and temporal references, which are subsequently checked for consistency with the underlying evidence. The resulting annotations represent events identified under the corresponding detection rules.

Media-Grounded Authoring. For other items, the authoring model uses the clip media together with relevant source annotations to formulate questions and references. Audio multiple-choice items, for example, are authored with access to both video and audio, allowing the recording context to inform question construction while keeping the requested answer grounded in acoustic evidence. Across task families, prompts constrain the item to the designated modality and discourage details unsupported by the clip. Narration instead combines observations from the available modalities into a coherent first-person description.

B.3.3 Quality Control

Candidate items undergo automated checks and human verification tailored to their task format and designated modality. These checks assess structural validity, reference consistency, answerability, and dependence on the intended evidence.

Structure and Reference Consistency. We validate required fields and task-specific formats, including option counts, answer indices, verdict labels, and temporal intervals. References are checked against the authoring evidence, including source annotations and rule-based motion descriptions where available. We monitor answer-position and true/false distributions within each column to limit label imbalance. Temporal references must fall within the clip and match the described event.

Answerability and Modality Dependence. We examine whether candidates admit shortcuts that bypass the intended evidence. For tasks with closed answer vocabularies, we check whether item fields or question patterns predict the reference. Text-only probes assess whether multiple-choice items can be solved from their questions and options without media, helping filter candidates whose wording or structure reveals the answer.

We also assess whether the designated modality supports the answer. IMU items are checked for well-posedness and consistency with structured signal-derived descriptions. These descriptions retain IMU information and serve as evidence for checking references. For audio multiple-choice items, evaluations without audio identify candidates answerable through visual context alone. These checks favor questions whose intended evidence contributes meaningfully to the answer.

Difficulty and Temporal Validity. A model panel assesses candidate difficulty. Items answered correctly by every panel model are filtered to retain questions that distinguish model capabilities. For temporal grounding, we check event descriptions against reference intervals, filtering candidates with unsuitable boundary conditions, insufficient duration, or unsupported localization targets. Model-based localization checks complement the signal-derived references and subsequent human review.

Human Verification. Human verification examines question clarity, reference correctness, modality grounding, and ambiguity, including whether the evidence supports the reference and the wording permits a clear interpretation. Task-specific checks cover the plausibility of multiple-choice alternatives, consistency between fact-checking verdicts and explanations, and correspondence between temporal descriptions and annotated intervals. Identified issues are addressed through revision or exclusion before finalization.

The combined process removes a substantial portion of the candidates. The IMU fact-checking column retains 244 of 1,762 candidates, while IMU open-ended question answering retains 193 of 2,160. These counts reflect cumulative selection rather than any individual check. Section B.4 further examines shortcuts, temporal priors, and modality dependence in the resulting benchmark.

B.4 Construct Validity

EgoMMU assesses whether models use the evidence required by each task and modality. We analyze language and metadata shortcuts, temporal priors, and modality withholding to support quality control and the interpretation of benchmark scores.

Language and Metadata Shortcuts. Text-only probes examine whether questions and options reveal the answer without access to the clip. For tasks with closed answer vocabularies, field-conditioned majority predictions assess associations between item fields and reference answers. Comparisons with shuffled-answer baselines help distinguish these associations from chance patterns. Together, these checks identify shortcuts in wording, answer distributions, and item structure.

The probes guide candidate filtering to reduce unintended cues; they do not establish that every retained item requires media. Performance on a pool selected using a given probe reflects that selection process and is interpreted alongside the other analyses.

Temporal Priors. Temporal-grounding scores can benefit from common patterns in target duration and position. The best constant-span baseline in Eq. 9, which uses one fixed interval for all items in a column, achieves 20.5, 11.1, and 11.4 mIoU for vision, audio, and IMU, respectively. These reference levels quantify overlap attainable without conditioning on the clip and provide context for interpreting event-localization performance.

Modality Dependence. Section E.2 withholds individual modalities at inference while keeping the checkpoint, evaluation items, and remaining inputs unchanged. Performance decreases primarily in the corresponding columns, supporting the intended relationship between task labels and evidence. These comparisons complement construction-time checks by examining the evaluated model's use of each stream. IMU withholding removes both learned signal tokens and tag text, assessing their combined contribution.

Task Difficulty and Scoring. Model-panel screening filters candidates answered correctly by every participating model to retain items that distinguish capabilities. Task-specific metrics further assess response quality. For fact checking, the composite score in Eq. 13 evaluates both verdict and justification: an incorrect verdict receives no credit, while a correct verdict is weighted by the quality of its supporting reason. The score therefore distinguishes responses with the same verdict but different justification quality. Section C describes the scoring procedures and grading configurations.

C Evaluation Settings & Protocols

This section describes the benchmarks used in our evaluation, the metrics used to assess model performance, and the prompts used for automated grading. Table 15 summarizes the evaluation settings for each benchmark, including the available modalities, the temporal extent of the input, the number of evaluation items, and the reported metrics. We detail the benchmark-specific settings below, followed by the metric definitions and references to the grading prompts.

Model Capabilities & Evaluation Inputs. In the benchmark result tables, the modality badges in the "Type" column indicate the modalities supported by the listed model configuration. EgoMind is therefore marked as "V + A + I", while its actual inputs depend on the modalities available in each benchmark and the evaluation protocol. Vision-only evaluations supply video, and video-audio evaluations supply the corresponding streams without IMU signals or tag text. An IMU badge thus denotes model capability, not IMU availability in every benchmark. For configurations explicitly evaluated with a modality withheld, the badges indicate the remaining modalities; the modality-ablation tables likewise identify the inputs retained in each setting.

C.1 Benchmark Statistics

C.1.1 EgoAVU-Bench

EgoAVU-Bench (Seth et al., 2026) evaluates audio-visual understanding through free-form questions about events and sounds in egocentric recordings. Its tasks cover segment narration, sound-source attribution, and three categories of hallucination, providing complementary assessments of how models interpret visual and auditory information. Among our external clip-level benchmarks, it provides paired audio and video inputs, allowing us to evaluate the audio pathway beyond EgoMMU. The evaluation collection contains 1,571 questions over 1,404 clips, each spanning 30 seconds. We report five metrics: one based on text overlap and four based on language-model grading.

The collection is constructed by re-cutting videos from a larger parent benchmark around the intervals referenced by individual questions. This process retains different question types at different rates, changing the task distribution relative to the parent benchmark. We therefore use the parent composition when computing the relevant weighted aggregates, as detailed in Section C.2. The question text retains its original timestamps; for 122 items, however, the referenced interval falls outside the re-cut clip. We exclude these items because their required evidence is unavailable in the supplied input, leaving 1,449 questions for the pooled evaluation.

C.1.2 EgoLife-QA

EgoLife-QA (Yang et al., 2025) evaluates long-horizon question answering over week-scale egocentric recordings. Answering a question requires retrieving relevant evidence from past observations, which may occur hours or days before the question is asked. It therefore assesses memory construction, retrieval, and answer generation beyond the understanding of an individual clip. We use this benchmark to evaluate the caption-then-retrieve pipeline under the frame budgets described in Section D.6. Our evaluation includes 475 four-way multiple-choice questions across five question types, with accuracy reported for each type and averaged across the evaluation set.

To distinguish retrieval errors from answer-generation errors, we additionally measure how accurately the pipeline localizes the relevant evidence. These measures include hit rates at three temporal resolutions and the distance to the nearest retrieved clip, as defined in Section C.2. Localization is evaluated on 270 of the 475 questions with a fixed reference moment. The remaining questions, including those concerning the most recent occurrence of an event, do not provide a fixed target for this comparison. All reported EgoLife-QA results cover a single participant, corresponding to the scope of our reproduction of the published caption-then-retrieve pipeline.

C.1.3 EgoSchema

EgoSchema (Mangalam et al., 2023) evaluates long-form egocentric video understanding using five-way multiple-choice questions. Each question is associated with a video of approximately three minutes and requires integrating information across events within the recording. Compared with

Table 15: **Summary of Evaluation Benchmarks and Metrics.** "**Type**" denotes the egocentric streams an answer draws on (V vision, A audio, I IMU). "**Extent**" denotes the temporal support of one question. "**Items**" denotes the number of questions or task items scored in total.

Benchmark	Type	Extent	Items	Metric	Range	Scoring	What It Captures
Ours							
EgoMMU	V A I	30 s	3,762	[1]Accuracy	[0, 1]	Closed Form	The exact match on the closed-form MCQ tasks
				[2]IoU & mIoU	[0, 1]	Closed Form	Span overlap, read against Eq. 9
				[3]Composite	[0, 1]	LLM-as-Judge	Verdict gated on its reason, Eq. 13
				[4]Open-Ended	[0, 1]	LLM-as-Judge	Meaning match against a reference for an open-ended QA task
				[5]Narration	[0, 1]	LLM-as-Judge	Meaning match against a reference for a narration task
External							
EgoAVU-Bench	V A	30 s	1,449	[1]AVSN-R	[0, 1]	Closed Form	Subsequence overlap with the reference narration, Eq. 11
				[2]AVSN-S	1–5	LLM-as-Judge	Rated faithfulness of the segment narration
				[3]SSA	1–5	LLM-as-Judge	Rated sound-source attribution, which sound came from what
				[4]AVH	1–5	LLM-as-Judge	Freedom from hallucinated actions, sounds and objects, combined by Eq. 14
				[5]Overall	1–5	LLM-as-Judge	Mean rating over all rated items
EgoLife-QA	V A	Days	475	[1]Accuracy	[0, 1]	Closed Form	Four-way choice, evidence days earlier, reported per question type
				[2]HIT @ DAY	[0, 1]	Closed Form	Selected day is the target's day, read beside a floor of always choosing the query day
				[3]HIT @ HOUR	[0, 1]	Closed Form	Target time falls inside the selected window, whose null moves with window width
				[4]HIT @ CLIP	[0, 1]	Closed Form	A retrieved clip brackets the target time, strict, 30 s
				[5]medΔ	min	Closed Form	Unsigned minutes to the nearest edge of the nearest retrieved clip, zero when it brackets
				[6]Acc | HIT	[0, 1]	Closed Form	Accuracy on items where HIT @ CLIP succeeds
				[7]Acc | MISS	[0, 1]	Closed Form	Accuracy on items where HIT @ CLIP fails, isolating retrieval from answering
EgoSchema	V	∼180 s	500	[1]Accuracy	[0, 1]	Closed Form	Five-way choice over long-form clips
EgoThink	V	Still	700	[1]Correctness	{0, 1}	LLM-as-Judge	Meaning match, single reference
EgoBlind	V	Full Clip	1,283	[1]Correctness	{0, 1}	LLM-as-Judge	Meaning match, any accepted reference, Eq. 12
EgoPolice	V	1/10/60 s	1,993	[1]Accuracy	[0, 1]	Closed Form	Choice, reported per window length

short-clip evaluations, this setting places greater emphasis on temporal context and the relationships between observations at different moments. It provides the longest visual input among our external clip-level benchmarks, complementing the shorter observation windows used in the other suites and allowing us to assess performance over a broader temporal extent.

We evaluate the 500-question subset with publicly available ground-truth labels; evaluation on the remaining questions requires submission to the benchmark's evaluation service. We use video-only inputs in this setting, without an audio branch. For a fixed number of input frames, sampling a three-minute video produces a lower temporal sampling density than sampling a 30-second clip. The results therefore characterize long-form understanding under the specified frame budget. We record the inference frame budget associated with each result so that performance can be interpreted together with the amount of visual input supplied.

C.1.4 EGOTHINK

EgoThink (Cheng et al., 2023) evaluates first-person understanding from still images through free-form questions spanning twelve dimensions, including activity recognition, forecasting, localization, and reasoning about the observed scene. Unlike the video-based suites, it provides no sequence of observations from which a model can directly recover temporal changes. This setting complements our video evaluations by assessing how well models interpret egocentric visual content from a single observation. It also allows us to examine whether the trained model retains useful image-based capabilities when temporal information is unavailable.

We evaluate all 700 released questions, each paired with one image, and score answer correctness against the reference provided by the benchmark. To use the same input interface as the video evaluations, we repeat the image to meet the configured frame count. This preserves the frame-count setting without introducing additional visual or temporal information. Answers are evaluated using a binary grading rubric. Although the rubric follows the same general form as the source benchmark, our grader model differs from the original one. The resulting scores support comparisons among models evaluated under our shared protocol, but are not directly comparable to the originally published scores.

C.1.5 EGOBLIND

EgoBlind (Xiao et al., 2025) evaluates egocentric assistance using recordings and questions from blind and low-vision users. Its questions span six categories, with information seeking most common, followed by safety and navigation. These user-generated questions introduce practical information needs that complement the perception and reasoning tasks in our other benchmarks. The evaluation allows us to examine whether models can provide relevant answers in assistive scenarios, where the information requested by the user may differ from the visually prominent content of a recording.

We evaluate all 1,283 released questions over 1,185 clips, using the full clip as the visual input. Each question includes between one and four accepted reference answers, reflecting the fact that multiple responses may satisfy the same information need. We consider both matching against any accepted reference and matching against one designated reference, with the two scoring variants defined in Section C.2. In the current evaluation protocol, the designated-reference variant is used when the evaluated system was trained against one specified reference; otherwise, predictions are compared against the accepted reference set.

C.1.6 EGOPOLICE

EgoPolice (Gonzalez Saez-Diez et al., 2026) evaluates understanding of body-camera footage using multiple-choice questions over observation windows of 1, 10, and 60 seconds. The question format remains consistent across these temporal settings, allowing us to examine model performance at different observation lengths. Short windows emphasize information available from a brief observation, while longer windows provide additional context about actions and events. This benchmark complements the fixed-duration suites by supporting separate assessments across three temporal extents within the same evaluation framework.

From the 12,000 released questions, we evaluate a subset of 1,993 items sampled jointly by window length and answer option. This stratification controls the composition of the evaluation subset and reduces answer-position imbalance within the temporal settings. We report accuracy separately for the three window lengths rather than combining them into a single score. Keeping these results separate makes differences across temporal settings visible and avoids allowing the relative frequency of each window length to determine an aggregate ranking. All models are evaluated on the same selected items within each setting.

C.2 METRICS & JUDGING

C.2.1 ANSWER CORRECTNESS

Closed-form tasks are evaluated using `Accuracy`, defined as the fraction of predictions that match the reference answer. Let $\mathcal{I}$ denote the items with a successfully parsed answer, $\hat{a}_i$ the predicted option, and a_i the reference. `Accuracy` is computed as $\texttt{Acc} = |\mathcal{I}|^{-1} \sum_{i \in \mathcal{I}} \mathbf{1}[\hat{a}_i = a_i]$. The parser accepts

lettered options, affirmative and negative responses where appropriate, and answers accompanied by explanations. Responses without a recoverable answer are excluded from $\mathcal{I}$ and reported separately. The resulting score therefore measures correctness among valid responses, and should be interpreted together with the proportion of outputs successfully parsed.

Chance accuracy depends on the number of options: 20% for EgoMMU and EgoSchema (Mangalam et al., 2023), and 25% for EgoLife-QA (Yang et al., 2025). These differences should be considered when interpreting results across benchmarks. For EgoPolice (Gonzalez Saez-Diez et al., 2026), accuracy is computed separately for the three observation lengths described in Section C.1.

C.2.2 Temporal Localization

We evaluate temporal localization at two scales: event intervals within individual clips and relevant evidence within long-horizon recordings. For clip-level grounding, IoU measures the intersection over union between a predicted interval $\hat{\varsigma}_i$ and its reference γ_i. We report the mean over N evaluated items as $\texttt{mIoU} = N^{-1} \sum_{i=1}^{N} \texttt{IoU}(\hat{\varsigma}_i, \gamma_i)$.

For long-horizon retrieval, each eligible question is associated with a reference moment. Let t_i denote this moment, $d(\cdot)$ map a timestamp to its day, $\hat{d}_i$ denote the selected day, W_i the selected temporal window, and $\mathcal{R}_i$ the set of retrieved clips. Each clip u spans $[u^-, u^+]$. Over the items $\mathcal{I}$ with a reference moment, we define three hit rates:

$$\texttt{HIT} = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} h_i, \qquad \begin{aligned} h_i^{\texttt{DAY}} &= \mathbf{1}\big[\hat{d}_i = d(t_i)\big], \\ h_i^{\texttt{HOUR}} &= \mathbf{1}\big[t_i \in W_i\big], \\ h_i^{\texttt{CLIP}} &= \mathbf{1}\big[\exists\, u \in \mathcal{R}_i :\ t_i \in [u^-, u^+]\big]. \end{aligned} \tag{7}$$

These metrics measure whether the selected day, selected window, or any retrieved clip contains the reference moment, respectively. To complement these binary measures, we report medΔ, the median distance from the reference moment to the nearest retrieved clip:

$$\texttt{med}\Delta = \operatorname*{median}_{i \in \mathcal{I}} \min_{u \in \mathcal{R}_i} \delta(t_i, u), \qquad \delta(t, u) = \begin{cases} 0, & t \in [u^-, u^+], \\ \min\big(|t - u^-|,\, |t - u^+|\big), & \text{otherwise.} \end{cases} \tag{8}$$

The distance is reported in minutes and equals zero when a retrieved clip contains the target. It measures temporal proximity without distinguishing whether the retrieved evidence precedes or follows the reference moment.

We interpret localization scores relative to baselines that account for the temporal extent accepted by each metric. For clip-level grounding, we use the best constant interval selected retrospectively over the reference spans $\{\gamma_i\}_{i=1}^{N}$:

$$\texttt{null} = \max_{\varsigma} \frac{1}{N} \sum_{i=1}^{N} \texttt{IoU}\big(\varsigma, \gamma_i\big). \tag{9}$$

This baseline uses the same interval for every item and measures how much overlap can be obtained without conditioning on the clip. Scores above it indicate an advantage over this constant prediction. For long-horizon retrieval, baselines are recomputed for each configuration: the day-level baseline selects the question's own day, while window- and clip-level baselines use random placements matched to the temporal widths and retrieval counts. An analogous measure can also be defined at ten-minute-block resolution.

To examine the relationship between retrieval success and answer correctness, we additionally report accuracy conditioned on whether a retrieved clip contains the reference moment. Using $h_i^{\texttt{CLIP}}$ from Eq. 7, we compute the following over items with both a reference moment and a valid predicted answer:

$$\texttt{Acc} \mid \texttt{HIT} = \frac{\sum_{i \in \mathcal{I}} h_i^{\texttt{CLIP}}\, \mathbf{1}[\hat{a}_i = a_i]}{\sum_{i \in \mathcal{I}} h_i^{\texttt{CLIP}}}, \qquad \texttt{Acc} \mid \texttt{MISS} = \frac{\sum_{i \in \mathcal{I}} (1 - h_i^{\texttt{CLIP}})\, \mathbf{1}[\hat{a}_i = a_i]}{\sum_{i \in \mathcal{I}} (1 - h_i^{\texttt{CLIP}})}. \tag{10}$$

The two scores characterize answer performance with and without successful clip-level retrieval. They are computed over separate subsets and need not sum to a fixed value. Their weighted average, using the respective subset sizes, recovers accuracy on the same evaluated item set. This comparison describes the association between retrieval and answering, rather than establishing a causal effect.

C.2.3 FREE-TEXT AGREEMENT

Free-text outputs are evaluated through lexical overlap or language-model grading. The overlap-based metric, `AVSN-R`, measures agreement between a prediction c and reference g using their longest common subsequence. Both strings are tokenized on whitespace. With ℓ denoting the subsequence length, the score is

$$\texttt{AVSN-R} \;=\; \frac{(1+\beta^2)\,pr}{r+\beta^2 p}, \qquad p=\frac{\ell}{|c|}, \quad r=\frac{\ell}{|g|}, \quad \beta=1.2, \tag{11}$$

where p and r represent subsequence precision and recall. The choice $\beta > 1$ assigns greater weight to recall, emphasizing coverage of the reference description.

Language-model grading evaluates semantic agreement while allowing paraphrases and compatible additional detail. `Open-Ended` and `Correctness` use binary judgments, `AVSN-S` evaluates segment narration on the source benchmark's five-point scale, and `Overall` averages ratings over the evaluated item set. EgoBlind additionally supports multiple accepted references. With R_i indexing the references for item i, its match-to-any score is

$$s_i \;=\; \max_{j\in R_i} \mathbf{1}\big[\ \texttt{Prediction}_i \equiv \texttt{Reference}_{ij}\ \big], \tag{12}$$

where $\equiv$ denotes semantic agreement under the grading rubric. A single-reference variant restricts the comparison to one designated answer. In the current protocol, this variant is used for systems trained against a designated reference, as described in Section C.1. The two variants define different scoring settings and should be distinguished when reporting comparisons.

C.2.4 GROUNDING AND FAITHFULNESS

We also evaluate the quality of supporting explanations and the consistency of generated content with reference annotations. For fact checking, `Composite` combines verdict correctness with a graded assessment of the accompanying reason. Let $b \in \{0,1\}$ indicate whether the verdict is correct and $r \in \{1,\dots,R\}$ denote the explanation rating, with $R=5$ under our rubric. The score is:

$$s \;=\; b\cdot\frac{r-1}{R-1}. \tag{13}$$

An incorrect verdict receives zero regardless of its explanation. A correct verdict without a supporting reason also receives zero, since the rubric assigns an absent explanation the minimum rating. The grader assesses the specificity, relevance, and consistency of the stated reason with the reference verdict. As it does not receive the media, this rating does not independently verify that every cited detail occurs in the clip.

For audio-visual evaluation, `SSA` assesses sound-source attribution, while `AVH` assesses consistency with reference answers concerning actions, sounds, and objects. Both use the grading rubric reproduced in Appendix F. Because re-cutting retains the three hallucination categories at different rates, we aggregate their mean ratings using weights from the parent benchmark:

$$\texttt{AVH} \;=\; \frac{\sum_{k\in K} w_k\,\bar{s}_k}{\sum_{k\in K} w_k}, \qquad w=\left(\tfrac{196}{576},\,\tfrac{196}{576},\,\tfrac{184}{576}\right), \tag{14}$$

where $\bar{s}_k$ is the mean rating for category k, and K contains the categories represented in the evaluated subset. The weights follow the order of actions, sounds, and objects, and are renormalized over the available categories. We also retain per-category means to distinguish category-specific performance from changes in aggregate composition. Higher scores indicate better performance throughout, including `AVH`, where a higher rating indicates greater consistency with the reference and fewer hallucination errors under the rubric.

C.3 GRADING CONFIGURATIONS & PROMPTS

We use language-model grading (*i.e.*, LLM-as-Judge) for four benchmarks: EgoMMU, EgoAVU-Bench (Seth et al., 2026), EgoThink (Cheng et al., 2023), and EgoBlind (Xiao et al., 2025). EgoMMU uses three task-specific grading prompts for fact checking, narration, and open-ended question

answering, while each external benchmark follows its respective rubric. The grader model and version are fixed across all evaluated tasks, supporting consistent comparisons within each protocol.

The complete grading prompts are reproduced in Section F.4 of Appendix F. The three EgoMMU prompts are provided in Section F.4.1, followed by the prompts for EgoAVU-Bench, EgoThink, and EgoBlind in Sections F.4.2, F.4.3, and F.4.4, respectively.

D Model Architecture & Training

This section describes the architecture and training of **EgoMind** and its integration into **EgoBuddy** for long-horizon egocentric assistance. We define the multimodal inputs, detail the sensor pathways and training stages, and explain how the model handles different modality combinations. We then describe memory construction, retrieval, and frame-budgeted capture over continuous recordings.

D.1 Problem Formulation

We consider synchronized egocentric recordings comprising video from an outward-facing camera, audio from onboard microphones, and IMU measurements from a head-mounted device. In the dual-IMU configuration, the left and right sensors each provide three-axis acceleration and angular velocity. Modality availability and the number of IMUs vary across data sources, as described in Section A.1.

Following Section A.2, we divide each recording into consecutive windows of length $T = 30\,\mathrm{s}$. A clip with all three modalities and both IMUs is represented as a synchronized tuple, that is,

$$x = (\mathbf{v}, \mathbf{a}, \mathbf{u}), \qquad \mathbf{u} = (\mathbf{u}^L, \mathbf{u}^R), \tag{15}$$

where $\mathbf{v}$ denotes video frames, $\mathbf{a}$ the audio waveform, and $\mathbf{u}^L$ and $\mathbf{u}^R$ the left and right IMU streams. All available streams correspond to the same recording interval.

We use V, A, and I to identify the three modalities and write $M^{\mathrm{av}} \subseteq \{V, A, I\}$ for those available in a clip. The model receives a non-empty subset $M \subseteq M^{\mathrm{av}}$, with x_M denoting the corresponding inputs. This distinction accommodates both missing modalities in source recordings and streams withheld during training or inference.

Prediction from Available Modalities. Given x_M and an instruction q, **EgoMind** generates a text response, that is,

$$\hat{y} = f_\phi(\mathbf{X}_M(x),\, q), \tag{16}$$

where $\mathbf{X}_M(x)$ contains the token representations of the supplied modalities and any associated IMU tag text, while q specifies the task. Blocks corresponding to absent modalities are omitted. The parameter set ϕ contains the trainable components; all remaining model parameters are frozen.

A training example associates the inputs and instruction with a reference response y_M supported by the supplied modalities. The subscript emphasizes that supervision must remain consistent with the evidence available in x_M. We denote the audio and IMU pathway parameters by θ_{A} and θ_{I}, respectively, and describe their pretraining in Section D.3. When IMU is supplied, its input includes both learned signal tokens and deterministic motion descriptors, denoted by $\mathrm{Tag}(\mathbf{u})$ and computed using the procedure in Section A.4.

D.2 Model Architecture

Open video-language models typically combine a vision encoder with a language model (Bai et al., 2025; Qwen Team, 2025b; Zhu et al., 2025; Wang et al., 2024; Zhang et al., 2025a;b; Li et al., 2025a; Yuan et al., 2025), while omni-modal systems extend this architecture to jointly process audio and video (Xu et al., 2025b; Qwen Team, 2025a; MiniCPM Team, 2025; Li et al., 2025c; Zhang et al., 2023; Tang et al., 2024; Wang et al., 2026a). Closed multimodal models (OpenAI, 2024; Gemini Team, Google, 2025; 2026; Anthropic, 2025) provide additional evaluation baselines, and recent surveys summarize developments in this area (Yin et al., 2024; Tang et al., 2025; Madan et al., 2024). Building on this encoder-language-model design, **EgoMind** incorporates audio and IMU pathways to support egocentric understanding across the three modalities.

We retain the backbone architecture and project audio and IMU features into its embedding space as additional token blocks. The language model's base weights remain frozen, with adaptation performed through low-rank updates. This modular design separates sensor encoding from language generation and allows the same integration strategy to be applied across backbone sizes, with projections matched to each embedding width. Figure 2 summarizes the architecture.

D.2.1 TOKEN ASSEMBLY

Each available modality contributes a block of tokens at the backbone's embedding width d. These blocks, together with any IMU tag text, form $\mathbf{X}_M(x)$ and are concatenated ahead of the instruction q. Under the standard input setting, video is sampled at one frame per second, yielding 30 frames per 30 s clip.

To provide a shared temporal reference, we add a learned temporal embedding to each sensor token according to its clip-relative time. Tokens assigned to the same temporal index receive the same temporal embedding across modalities, while the backbone's positional encoding remains unchanged. This encoding provides explicit timing information for learning cross-modal relationships. Each modality retains a separate token block; absent modalities and their associated inputs are omitted, allowing the same architecture to process different modality subsets (Section D.4).

D.2.2 AUDIO PATHWAY

Audio-visual representation learning provides encoders that associate acoustic content with visual observations and actions (Shi et al., 2022; Guzhov et al., 2022; Chen et al., 2022; Elizalde et al., 2023; Chen et al., 2024a;b). Our audio pathway processes the waveform $\mathbf{a}$ using a frozen BEATs encoder (Chen et al., 2022), followed by a trainable window Q-Former that compresses its output into a fixed number of tokens while retaining temporal order. The resulting tokens are projected into the backbone's embedding space.

We use this pathway for sound-event understanding rather than speech transcription. Accordingly, audio supervision describes audible events, sound sources, and acoustic or paralinguistic properties, while excluding transcribed speech and conversational content. Section A.5 details the corresponding annotation constraints.

D.2.3 IMU PATHWAY

The IMU pathway combines learned inertial signal representations with deterministic motion descriptors. The learned branch resamples $\mathbf{u}^L$ and $\mathbf{u}^R$ to a common rate, processes them through a convolutional stem and a Transformer, and projects the resulting features into the backbone's embedding space. In parallel, the feature extraction procedure in Section A.4 produces $\mathrm{Tag}(\mathbf{u})$, a short text block based on the motion labels in Table 9.

These two representations provide complementary inputs: learned tokens encode temporal signal patterns, while the tag text explicitly summarizes predefined motion properties. The IMU encoder has the same configuration across all four model sizes, whereas its projection depends on the backbone's embedding width. Together, the encoder and projection account for approximately 35–37% of the trainable parameters, as reported in Table 16.

D.2.4 MODEL FAMILY

We instantiate our tri-modal egocentric understanding model at four backbone sizes to examine performance under different model-capacity budgets. Table 16 lists the backbone configurations and trainable parameter counts. All variants follow the same sensor-integration design, with projections adapted to the corresponding embedding width.

The two larger backbones contain 32 layers, while the two smaller ones contain 24. Comparisons across the full family therefore reflect changes in both width and depth. Models within each same-depth pair control for layer count, but the overall comparison is between backbone configurations rather than an isolated ablation of width.

D.3 TRAINING SETUP

We train **EgoMind** through sensor-specific pretraining followed by joint instruction tuning. Pretraining aligns each sensor pathway with the language backbone before introducing multimodal task supervision. The audio pathway builds on a pretrained encoder, whereas the IMU pathway first learns a signal representation and then aligns it with language generation.

Table 16: **Summary of Model Configurations.** Backbone dimensions and trainable parameters across four model sizes. Audio and IMU counts include their respective projections; "IMU Share" denotes the proportion of total trainable parameters assigned to the IMU pathway.

Model	Hidden Dim.	Depth (Attn. + SSM)	Audio	IMU	LoRA	IMU Share
EgoMind-9B	4096	32 (8 + 24)	58.3M	43.3M	15.8M	36.9%
EgoMind-4B	2560	32 (8 + 24)	46.9M	31.8M	12.5M	34.9%
EgoMind-2B	2048	24 (6 + 18)	44.2M	29.0M	6.4M	36.4%
EgoMind-1B	1024	24 (6 + 18)	40.2M	25.0M	4.8M	35.7%

We denote the audio pathway, IMU pathway, and language-model adapter parameters by θ_{A}, θ_{I}, and θ_{LoRA}, respectively. Phase "A" trains the audio pathway, while Phase "I" trains the IMU pathway through contrastive learning and caption alignment. Two auxiliary projection heads are used only during IMU contrastive pretraining and are subsequently discarded. Joint instruction tuning updates both sensor pathways and the low-rank adapters. The vision encoder, BEATs encoder, and base language-model weights remain frozen throughout.

D.3.1 Pretraining Phase A: Audio

We retain the pretrained BEATs encoder (Chen et al., 2022) and train the window Q-Former and its output projection to produce audio tokens compatible with the language backbone. Given an audio waveform $\mathbf{a}$ and its caption y, we minimize the caption's negative log-likelihood conditioned on the projected audio tokens and a fixed captioning instruction, that is,

$$\begin{aligned} \mathcal{L}_{\mathrm{A}}(\theta_{\mathrm{A}}) &= -\,\mathbb{E}_{(\mathbf{a},\,y)} \sum_{k} \log p_{\theta_{\mathrm{A}}}\left(y_k \mid y_{<k},\, \mathbf{z}_A\right), \\ \mathbf{z}_A &= \mathrm{QF}_{\theta_{\mathrm{A}}}\left(\mathrm{BEATs}(\mathbf{a})\right). \end{aligned} \tag{17}$$

Here, k indexes caption tokens, QF includes the output projection, and the fixed instruction is omitted from the conditioning notation for brevity. Although the language model is frozen, gradients propagate through it to update θ_{A}. Caption supervision directly aligns the audio pathway with its downstream role in language generation and provides descriptions of acoustic events beyond a fixed classification vocabulary.

Specificity Criterion. Captioning loss alone does not establish how strongly predictions depend on the input signal. We therefore compare captions generated from a clip's own audio with those generated after substituting another clip's waveform, while keeping the reference unchanged. Let $\hat{y}(\mathbf{a})$ denote the generated caption and Match a caption-agreement metric. For a mismatched clip $j' \neq j$, we measure the agreement difference, that is,

$$\mathrm{Spec} = \mathbb{E}_j\Big[\mathrm{Match}\big(\hat{y}(\mathbf{a}_j),\, y_j\big) - \mathrm{Match}\big(\hat{y}(\mathbf{a}_{j'}),\, y_j\big)\Big]. \tag{18}$$

A positive difference indicates that the matched input supports better agreement with the reference than the substituted signal. A pathway that ignores its input is expected to show little difference. We apply the same comparison to Phase "I" using $\mathbf{u}$ in place of $\mathbf{a}$. Section E reports the evaluation metrics and results across model sizes.

D.3.2 Pretraining Phase I: IMU

We initialize the IMU encoder from scratch and train it in two steps. Contrastive pretraining associates signal representations with caption semantics, followed by caption alignment that connects the learned encoder to the language backbone.

Contrastive Objective. For each IMU-caption pair, we construct normalized embeddings in a shared contrastive space. The IMU encoder is shared between the left and right sensors. Its outputs are averaged over time for each sensor and then across sensors. On the text side, we average the caption's token embeddings from a frozen Qwen3.5 embedding table, denoted by Emb. Two trainable heads

map these pooled representations into the shared space, that is,

$$\bar{\mathbf{s}} = \tfrac{1}{2}\Big(\overline{\mathrm{Enc}_{\theta_I}(\mathbf{u}^L)} + \overline{\mathrm{Enc}_{\theta_I}(\mathbf{u}^R)}\Big),$$
$$\mathbf{e}_I = \frac{\mathrm{Head}_I(\bar{\mathbf{s}})}{\|\mathrm{Head}_I(\bar{\mathbf{s}})\|_2}, \qquad \mathbf{e}_{\mathrm{txt}} = \frac{\mathrm{Head}_{\mathrm{txt}}\big(\overline{\mathrm{Emb}(y)}\big)}{\big\|\mathrm{Head}_{\mathrm{txt}}\big(\overline{\mathrm{Emb}(y)}\big)\big\|_2}. \tag{19}$$

An overline denotes averaging over temporal or token positions. For a batch $\mathcal{B}$ of paired examples and a learned positive temperature κ, we optimize a symmetric contrastive objective, that is,

$$\mathcal{L}_{\mathrm{con}} = -\frac{1}{2|\mathcal{B}|}\sum_{j\in\mathcal{B}}\left[\log\frac{\exp\big((\mathbf{e}_I^j)^\top \mathbf{e}_{\mathrm{txt}}^j/\kappa\big)}{\sum_{l\in\mathcal{B}}\exp\big((\mathbf{e}_I^j)^\top \mathbf{e}_{\mathrm{txt}}^l/\kappa\big)} + \log\frac{\exp\big((\mathbf{e}_{\mathrm{txt}}^j)^\top \mathbf{e}_I^j/\kappa\big)}{\sum_{l\in\mathcal{B}}\exp\big((\mathbf{e}_{\mathrm{txt}}^j)^\top \mathbf{e}_I^l/\kappa\big)}\right]. \tag{20}$$

The two terms match each IMU example to its caption and each caption to its IMU example, respectively. Since the embeddings are normalized, their inner products are cosine similarities. This step updates the IMU encoder, both projection heads, and the temperature, while keeping the text embedding table frozen. The corresponding retrieval evaluations use within-batch candidates, so their accuracy is interpreted relative to $|\mathcal{B}|$ rather than the full dataset.

Alignment Objective. We next retain the pretrained IMU encoder, discard the contrastive heads, and introduce a randomly initialized projection into the language model's embedding space. The encoder and projection are jointly optimized to generate IMU captions through the frozen language model, that is,

$$\mathcal{L}_{\mathrm{I}}(\theta_{\mathrm{I}}) = -\,\mathbb{E}_{(\mathbf{u},\,y)}\sum_k \log p_{\theta_{\mathrm{I}}}\big(y_k \mid y_{<k},\,\mathbf{z}_I\big),$$
$$\mathbf{z}_I = \mathrm{Proj}_{\theta_{\mathrm{I}}}\Big(\mathrm{Enc}_{\theta_{\mathrm{I}}}\big(\mathbf{u}^L,\,\mathbf{u}^R\big)\Big). \tag{21}$$

As in Phase "A", the fixed captioning instruction is omitted from the notation. Rule-based tag text is not supplied during this step, ensuring that caption generation is conditioned on the learned signal representation.

The contrastive encoder has the same configuration across the model family and is pretrained once. Its weights initialize caption alignment at all four model sizes in Table 16, with a separate projection matched to each backbone's embedding width.

D.3.3 Instruction Tuning

After sensor-specific pretraining, we jointly instruction-tune the model on the task mixture. This stage combines the aligned pathways and introduces training with multiple modalities. Following Section D.4, the subset distribution π retains the modalities required by each target and may omit optional streams. For an input subset M and its supported reference response y_M, we minimize the mean negative log-likelihood over supervised token positions $\mathcal{S}$, that is,

$$\mathcal{L}(\phi) = -\,\mathbb{E}_{(x,\,q)}\,\mathbb{E}_{M\sim\pi}\left[\frac{1}{|\mathcal{S}|}\sum_{k\in\mathcal{S}}\log p_\phi\big(y_{M,k} \mid y_{M,<k},\,\mathbf{X}_M(x),\,q\big)\right]. \tag{22}$$

The set $\mathcal{S}$ includes the answer tokens and the end-of-sequence token, while excluding the prompt. Supervising the end-of-sequence token trains the model to terminate its response. This expression specifies the token-level averaging used for the objective in Eq. 3. Section E evaluates the contribution of each pretraining phase, including a variant that omits both phases and randomly initializes the corresponding trainable sensor components.

Trainable Components. Joint instruction tuning updates the audio Q-Former and output projection, the IMU encoder and projection, and low-rank adapters applied to nine module types of the language backbone. The vision encoder, BEATs encoder, and base language-model weights remain frozen. Table 17 summarizes the components and their parameter counts.

Table 17: **Summary of Model Components.** Architecture, parameter counts, token representations, and training status for EgoMind-9B under the standard 30-frame (FPS1) input setting.

Component	Architecture	Params	Tokens	Training
Vision Encoder	Qwen3.5 ViT	$\sim$400M	678/clip	Frozen
Audio Encoder	BEATs	$\sim$90M	1500/clip	Frozen
Audio Projector	Window Q-Former + Output Projection	58.3M	89/clip	Trainable
IMU Encoder	Conv1d + 4-Layer Transformer	23.1M	60/side	Trainable
IMU Projector	$768 \to d$ + Side/Absence Embeddings	20.2M	120/clip	Trainable
IMU Tags	Rule-Based Motion Descriptors	—	Text	Fixed
LLM Backbone	Qwen3.5 (8 Attn. + 24 SSM Layers)	9B	Text	LoRA ($r = 14$)

Schedule. Instruction tuning runs for three epochs using AdamW, with linear warmup over the first 3% of updates followed by cosine learning-rate decay. We compare training losses under matched supervision, normalization, and data-mixture settings, since differences in these choices affect their absolute values.

D.4 Modality Handling

Our model uses a single checkpoint to process any non-empty subset of vision, audio, and IMU inputs. As defined in Eq. 16, the supplied subset M determines which modality blocks appear in $\mathbf{X}_M(x)$. This interface accommodates differences in source-data availability and allows streams to be withheld at inference without changing the model architecture or retraining its parameters.

Modality-Robust Training. We train with varying input combinations while retaining the evidence required by each supervision target. Let M^{req} denote the modalities needed to support a target and M^{av} those available in its source clip. The subset distribution π in Eq. 22 always retains the required modalities and independently retains each optional modality with probability $1 - p_{\text{drop}}$. With $p_{\text{drop}} = 0.15$, the sampling rule is defined as follows, that is,

$$M^{\text{req}} \subseteq M \subseteq M^{\text{av}}, \qquad M \neq \varnothing,$$
$$\Pr\left[m \in M \mid M^{\text{req}}, M^{\text{av}}\right] = \begin{cases} 1, & m \in M^{\text{req}}, \\ 1 - p_{\text{drop}}, & m \in M^{\text{av}} \setminus M^{\text{req}}. \end{cases} \tag{23}$$

This procedure varies the optional context without removing inputs needed to answer the instruction. The training mixture also includes examples prepared for specific input combinations, including vision-text, vision-audio-text, audio-text, and IMU-text. Here, text denotes the instruction and response, rather than an additional sensor modality. These examples provide direct supervision for single-modality and two-modality inputs alongside the combinations produced through dropout.

Modality-Matched Targets. The reference response y_M must be supported by the supplied inputs and remain appropriate for the instruction q. For captioning, a modality-specific target describes evidence from that modality, while a fused target requires the modalities supporting its content. A narration that depends on all three streams therefore retains all three during training; supplying two streams alone does not justify using a target that requires the third.

For task-specific supervision, including multiple-choice questions, fact checking, open-ended question answering, and temporal grounding, the target remains the corresponding answer or temporal annotation. Its required modality set determines which streams can be omitted. The notation y_M expresses this compatibility requirement: it does not imply that every sampled subset needs a different reference. The same target can supervise several subsets when each retains the evidence it requires.

Inference with Missing Modalities. At inference, the available inputs determine M, and token blocks for omitted modalities are excluded from the sequence. IMU-associated tag text is included only when IMU is supplied, keeping both forms of motion input consistent with the selected subset. The instruction is retained across input configurations, and no separate checkpoint is required for each combination.

Table 18: **Comparison of Memory and Capture Frameworks.** Input modalities, memory organization, and camera settings. **Symbols:** ✓: supported; ✗: not supported; –: not reported. EgoBuddy lists framework capabilities; the reported memory experiments use video and audio.

Framework	Venue	Memory	IMU Input	Hierarchical	Streaming	Camera
Vinci (Huang et al., 2024b)	arXiv'24	V	✗	–	✓	Always on
OmniQuery (Li et al., 2025b)	arXiv'24	V	✗	–	✗	User-initiated
EgoLife (Yang et al., 2025)	CVPR'25	V A	✗	✓	–	Always on
EgoTrigger (Paruchuri et al., 2025)	TVCG'25	V	✗	–	–	Audio-triggered
Vinci2 (Gong et al., 2026)	arXiv'26	V	✗	✓	✓	–
LightMem-Ego (Chen et al., 2026)	arXiv'26	V A	✗	✓	✓	Always on
MobiDiary (Deng et al., 2026)	arXiv'26	I	✓	–	–	None
EgoBuddy	**Ours**	V A I	✓	✓	✓	Per-clip frame budget

This interface also supports controlled modality ablations using the same trained weights, as reported in Section E. Unlike training-time dropout, an evaluation ablation may deliberately remove a modality required by the question to measure the resulting performance change. Accepting an input subset specifies the model's interface; whether that subset provides sufficient evidence depends on the task.

D.5 Streaming Memory

Personal capture and memory-assisted interaction have been explored through SenseCam (Hodges et al., 2006), egocentric assistants, and personal-memory agents (Huang et al., 2024b; Pu et al., 2025; Zhang et al., 2025c; Li et al., 2025b; Zhou et al., 2026). Recent work extends these settings to week-scale recordings (Yang et al., 2025), streaming interaction (Yan et al., 2025), egocentric perception (Ye et al., 2025), and hierarchical memory (Chen et al., 2026; Gong et al., 2026), supported by lifelog, streaming, and long-video benchmarks (Tran et al., 2025; Zhou et al., 2025b; Forte et al., 2026; Lando et al., 2025; Wang et al., 2026b; Fu et al., 2024; Chandrasegaran et al., 2024; Zhou et al., 2025a; Wu et al., 2024; Zou et al., 2024).

EgoBuddy combines multimodal perception from EgoMind with a hierarchical text memory. The framework supports different input subsets; the memory experiments reported here use video and audio. Visual input is controlled through a per-clip frame budget, as described in Section D.6.

Table 18 summarizes the comparison with prior memory and capture frameworks. Below, we detail the record generation, memory consolidation, and retrieval procedures outlined in Section 3.4.

D.5.1 Clip-Level Records

We process a continuous recording as a sequence of fixed-duration clips using the same input interface as clip-level inference. We use $o_{\le t}$ to denote observations available up to time t, with $o^{A}_{\le t}$ and $o^{I}_{\le t}$ denoting the audio and IMU streams. Each completed $T = 30$ s window produces a text record, *i.e.*,

$$x_i = o_{[(i-1)T,\, iT)}, \quad c_i \;\; = f_\phi\big(\mathbf{X}_{M_i}(x_i),\, q_{\text{rec}}\big), \quad \mathcal{C} = \{\, c_i : i \in \mathcal{I}_{\text{rec}} \,\}, \tag{24}$$

where M_i identifies the modalities supplied for clip i, q_{rec} is a fixed recording instruction, and $\mathcal{I}_{\text{rec}}$ indexes the processed clips. Each record is associated with its source interval for temporal grouping and retrieval. One hour of continuous recording yields 120 clip records.

We use EgoMind-9B to generate c_i, reusing the perception model described in Sections D.2 to D.4. The model supports records generated from different modality subsets. In the memory experiments

reported in Table 6, we supply video and audio, so each record summarizes visible scenes and actions together with acoustic events and vocal characteristics. Speech is described through its acoustic properties rather than transcribed content, following Section A.5. IMU inputs and their associated tag text are not used in this evaluation.

The input subset and visual frame budget may vary across configurations, but record generation retains the same clip duration and model interface. Long-horizon operation therefore extends the sequence of records rather than the temporal context of a single EgoMind invocation.

D.5.2 Hierarchical Memory

We organize the records into three temporal levels: clips, hourly episodes, and days. Clip records from the same clock hour are consolidated into an episode summary, and the episode summaries for a day are consolidated into a day summary. We denote this precomputed memory by $\mathcal{M} = (\mathcal{C}, \mathcal{E}, \mathcal{D})$, with $\mathcal{E} = \sigma(\mathcal{C})$ and $\mathcal{D} = \sigma'(\mathcal{E})$. Here, an episode denotes a fixed hourly group rather than an activity-dependent segment. For comparison, $\mathcal{M}_t$ denotes a strictly causal memory state containing only information from observations available by time t.

Gemini-3.5-Flash performs episode- and day-level consolidation using the text records. Coarser summaries support temporal selection, while the clip-level entries retain more detailed descriptions for answering. This organization follows established hierarchical-memory designs (Yang et al., 2025; Chen et al., 2026; Gong et al., 2026), with EgoMind providing the multimodal content stored at the finest level. The text hierarchy provides a compact representation of the observations, although retaining all clip records still causes memory size to increase with recording duration.

Temporal Availability. Strict question-time causality requires all retrieved records and summaries to contain only information from observations available by the question time t. Our current implementation restricts day and hour selection using the question timestamp, but does not strictly truncate the memory at t. Consequently, a selected interval or its summary may include observations recorded after the question. We therefore distinguish this implementation from the strictly causal setting: the reported results evaluate retrieval over precomputed text memory under coarse temporal restrictions. Record generation is independent of the question, and answering does not revisit raw media, but these properties alone do not guarantee question-time causality.

Comparison Setting. Table 6 compares video-audio memory generated by EgoMind-9B with memory generated by the EgoLife captioner, EgoGPT (Yang et al., 2025), at matched frame budgets. Gemini-3.5-Flash performs consolidation, retrieval, and answering under the same procedures for both configurations. This comparison evaluates the effect of the generated video–audio records; it does not measure the contribution of IMU inputs to long-horizon retrieval. Both configurations use the temporal restrictions described above.

D.5.3 Retrieval and Answering

Given a question q at time t, EgoBuddy retrieves evidence from the precomputed memory $\mathcal{M}$. The hierarchical route first selects a day using the day summaries, then an hour within that day using the episode summaries. Candidate days and hours are restricted by the question timestamp, but records and summary contents are not strictly truncated at t, as discussed above. Clip records from the selected hour form the candidate evidence set, from which up to 20 entries are supplied for answering.

Questions asking about the most recent occurrence of an event follow a separate keyword-search route over $\mathcal{C}$. This route searches the clip records without first restricting candidates to a single selected hour. Both routes provide textual evidence to the same answer model.

AAlgorithm 2 summarizes evidence retrieval. The function score denotes the relevance assessment used for day and hour selection, and $\mathrm{SelectEntries}$ denotes candidate selection subject to the entry limit. We use $\mathrm{start}(h)$ for the absolute start time of an hourly interval, avoiding ambiguity between hours on different days. Given the retrieved records, Gemini-3.5-Flash produces the final response, that is,

$$\hat{y} = \mathcal{A}\big(q,\, \mathcal{R}(q, \mathcal{M}, t)\big). \tag{25}$$

The separation between record generation and answering allows the same memory to support multiple questions without rerunning perception over the original recordings.

Algorithm 2 Memory Retrieval. Hierarchical retrieval with day/hour temporal restrictions and a separate keyword-search route.

Require: Precomputed memory $\mathcal{M} = (\mathcal{C}, \mathcal{E}, \mathcal{D})$; question q; question time t

if q asks for the most recent occurrence of an event **then**
  $\mathcal{U} \leftarrow \mathrm{KeywordSearch}(q, \mathcal{C})$
else
  $\hat{j} \leftarrow \arg\max_{\substack{j \in \mathrm{days}(\mathcal{D}) \\ j \le \mathrm{day}(t)}} \mathrm{score}(q, \mathcal{D}[j])$
  $\hat{h} \leftarrow \arg\max_{\substack{h \in \mathrm{hours}(\mathcal{E}, \hat{j}) \\ \mathrm{start}(h) \le t}} \mathrm{score}(q, \mathcal{E}[h])$
  $\mathcal{U} \leftarrow \{\, c_i \in \mathcal{C} : c_i \text{ belongs to hour } \hat{h} \,\}$
end if
$\mathcal{R}_q \leftarrow \mathrm{SelectEntries}(q, \mathcal{U}, 20)$
return $\mathcal{R}_q$

D.6 Frame-Budgeted Capture

Reducing visual input is a common approach to efficient video understanding, including frame selection and token reduction (Shen et al., 2025; Park et al., 2024; Huang et al., 2024a; Bagrov et al., 2025; Tu et al., 2026). Recent evaluations further examine performance under different frame budgets and temporal sampling settings (Shen et al., 2026; Tian, 2026). Complementary approaches use audio to guide capture (Paruchuri et al., 2025), transfer visual representations to other sensors (Chen et al., 2025; Radevski et al., 2023), or construct records without camera input (Deng et al., 2026). Motivated by these settings, we evaluate EgoBuddy with a fixed visual budget per clip while retaining the corresponding audio input.

Uniform Sampling. For each $T = 30\,\mathrm{s}$ clip, we uniformly sample B video frames, with $B \in \{30, 15, 6, 3, 1\}$. These budgets correspond to effective sampling rates of 1, $1/2$, $1/5$, $1/10$, and $1/30$ frames per second, respectively. The audio input spans the full clip at every budget, while IMU signals and tag text are excluded from the reported memory experiments, consistent with Section D.5.

We apply the same frame budgets to the EgoMind-9B and EgoGPT captioners and retain the shared consolidation, retrieval, and answering procedures described above. Table 6 reports the resulting memory evaluations. This setting measures performance as visual input is reduced; frame counts alone do not establish reductions in device power, heat, or end-to-end latency.

Causal Capture Formulation. Beyond uniform sampling, a causal policy could select capture times using previously observed audio and IMU signals. Let $\mathcal{T}_i$ denote the candidate capture times within clip x_i, and let η_t denote a decision threshold. A budget-constrained causal policy takes the following form, that is,

$$\begin{aligned} g(t) &= \psi\big(o^{A}_{\le t},\, o^{I}_{\le t}\big), \quad \text{acquire at } t \iff g(t) > \eta_t, \\ \sum\nolimits_{t \in \mathcal{T}_i} \mathbb{1}\big[\text{acquire at } t\big] &\le B, \qquad \mathcal{T}_i \subset [(i-1)T,\, iT). \end{aligned} \tag{26}$$

Here, ψ uses only sensor observations available by t, and the decision rule must enforce the remaining frame budget. This formulation specifies constraints for adaptive capture rather than the policy used in the reported experiments. Our uniform sampling schedule selects frames at predetermined intervals without evaluating a sensor-dependent score.

Adaptive Selection. Adaptive capture would allocate the same frame budget according to the observed sensor content. Learning a score that predicts the value of an additional visual observation, and evaluating it under causal acquisition constraints, remain future work. The current experiments assess uniform frame-budget reduction and do not establish an advantage for audio- or IMU-triggered capture.

D.7 Symbols & Notations

Table 19 summarizes the notation for model inputs, architecture, training, and memory. Modality-specific symbols follow the same color convention throughout: **Vision**, **Audio**, and **IMU**.

Table 19: **Summary of Symbols & Notations.** Key symbols for multimodal inputs, model architecture, training, memory, and frame-budgeted capture, which are used throughout the paper.

Symbol	Description
Inputs	
T	Clip duration, fixed at 30 s.
$\mathbf{v}$, $\mathbf{a}$, $\mathbf{u}$	Video frames, audio waveform, and IMU signals.
$\mathbf{u}^L$, $\mathbf{u}^R$	Left and right IMU streams.
x, x_M	A clip and its inputs restricted to modality subset M.
V, A, I	Modality identifiers for vision, audio, and IMU.
M^{av}, M^{req}, M	Available, required, and supplied modality sets, with $M^{\mathrm{req}} \subseteq M \subseteq M^{\mathrm{av}}$ during training.
q, y, y_M	Instruction, reference response, and a reference supported by subset M.
Model	
f_ϕ, $\hat{y}$	Model with trainable parameters ϕ, and its generated response.
$\mathbf{z}_* \in \mathbb{R}^{n_* \times d}$	Token block with n_* tokens and embedding width d; $*$ identifies a modality, tag text, or instruction text.
$\mathbf{X}_M(x)$	Input token sequence for modalities in M, including associated IMU tag text; the instruction q is supplied separately.
BEATs, QF	Frozen audio encoder and trainable window Q-Former with its output projection.
Enc, Proj	Learned IMU encoder and its projection into the backbone embedding space.
$\mathrm{Tag}(\mathbf{u})$	Deterministic motion descriptors represented as text.
θ_A, θ_I, θ_{LoRA}	Parameters of the audio pathway, IMU pathway, and language-model adapters.
Training	
$\mathcal{L}_A$, $\mathcal{L}_{\mathrm{con}}$, $\mathcal{L}_I$, $\mathcal{L}$	Audio alignment, IMU contrastive pretraining, IMU caption alignment, and instruction-tuning objectives.
$\mathcal{B}$, κ	Contrastive batch and learned positive temperature.
$\bar{\mathbf{s}}$	IMU representation averaged over time and across sensors.
$\mathbf{e}_I$, $\mathbf{e}_{\mathrm{txt}}$	Unit-normalized IMU and text embeddings for contrastive learning.
Emb	Frozen Qwen3.5 token-embedding table.
Head_I, $\mathrm{Head}_{\mathrm{txt}}$	Auxiliary IMU and text projection heads used in contrastive pretraining.
Spec, Match	Matched-versus-mismatched agreement difference and caption metric used to compute it.
$\mathcal{S}$	Supervised answer and end-of-sequence token positions.
π, p_{drop}	Modality-subset distribution and dropout probability for optional modalities.
Memory	
t, $o_{\le t}$	Observation or question time, and observations accumulated up to that time.
$o^A_{\le t}$, $o^I_{\le t}$	Audio and IMU observations accumulated up to t.
x_i, M_i	The i-th clip and its supplied modality subset.
q_{rec}, c_i, $\mathcal{I}_{\mathrm{rec}}$	Recording instruction, generated clip record, and indices of processed clips.
$\mathcal{C}$, $\mathcal{E}$, $\mathcal{D}$	Clip records, hourly episode summaries, and day summaries.
σ, σ'	Consolidation from clips to hours and from hours to days.
$\mathcal{M}$, $\mathcal{M}_t$	Precomputed memory and a strictly causal memory state containing only information available by t.
$\mathcal{R}(q, \mathcal{M}, t)$, $\mathcal{A}$	Retrieval under the implemented temporal restrictions, and answer generation.
$\hat{j}$, $\hat{h}$	Selected day and hourly interval.
$\mathcal{U}$, $\mathcal{R}_q$	Candidate clip records and the selected evidence returned for question q.
Capture	
B, $\mathcal{T}_i$	Per-clip frame budget and candidate capture times in clip i.
ψ, $g(t)$, η_t	Causal scoring function, capture score, and decision threshold in the adaptive-capture formulation.

E Additional Experiments & Analyses

This section expands the main evaluation with complete benchmark results and analyses of modality dependence, encoder pretraining, and qualitative model behavior. We compare performance across model families and sizes, examine whether task scores reflect access to the relevant sensor inputs, and assess how pretraining affects signal specificity and downstream performance. Qualitative examples across all EgoMMU task families further illustrate differences between scene recognition, evidence-grounded answering, and temporal localization.

E.1 Complete Benchmark Results

We report all evaluated model variants on EgoMMU in Table 20, EgoAVU-Bench (Seth et al., 2026) in Table 21, and the external egocentric benchmarks (Mangalam et al., 2023; Xiao et al., 2025; Cheng et al., 2023; Gonzalez Saez-Diez et al., 2026) in Table 22.

These tables extend the representative results in the main text; shared entries use the same measurements, evaluation items, and inference settings. Comparisons are made within each column under the protocols of Section C.2, since option counts, rating scales, and reference formats differ across tasks.

Scaling Across Task Families. Within the EgoMind family, performance increases consistently from 1B to 9B across all 13 EgoMMU columns. The improvement extends beyond answer selection to fact checking, open-ended answering, narration, and temporal grounding, indicating that the benefit of larger models is distributed across several forms of egocentric understanding. The gains are nevertheless uneven: MCQ accuracy increases by 19.8, 27.9, and 27.2 percentage points for vision, audio, and IMU, respectively. Scaling therefore improves all three modalities without reducing their performance differences to a uniform offset. As described in Section D.2.4, these comparisons vary both backbone width and, between the smaller and larger pairs, depth.

Modality-Dependent Model Rankings. Strong visual performance does not imply equally strong performance on questions about other modalities. On EgoMMU, Gemini 3.1 Pro (Gemini Team, Google, 2026) exceeds EgoMind-9B on visual MCQ accuracy (78.2% versus 70.3%), whereas the ordering reverses on IMU MCQ accuracy (45.4% versus 66.2%). This reversal illustrates why modality-specific reporting is necessary: a model's ability to interpret visible scenes does not fully characterize its ability to answer questions about the wearer's motion.

Cross-model comparisons reflect differences in both available inputs and training, however, so they do not isolate the contribution of an individual pathway. Section E.2 examines that contribution by withholding streams from the same checkpoint.

Complementary Audio-Visual Metrics. EgoAVU-Bench further separates narration quality, sound-source attribution, and hallucination. These measures reveal differences that an overall score can conceal. For example, the unmarked EgoMind-9B configuration improves the object-hallucination rating from 1.87 to 2.58 relative to Qwen3.5-9B (Qwen Team, 2026), while the action-hallucination rating remains close (3.32 versus 3.27). Higher ratings indicate fewer hallucinations. Thus, improved faithfulness in one category need not extend equally to another, even within the same benchmark. We retain the category-level results alongside the aggregate to distinguish these behaviors.

Transfer Across External Tasks. The external evaluations assess whether the improvements extend beyond EgoMMU. Relative to Qwen3.5-9B (Qwen Team, 2026), EgoMind-9B improves EgoSchema accuracy from 67.1% to 72.1%, while EgoThink scores remain close (73.4% versus 73.6%). EgoBlind results are also nonuniform: information-seeking accuracy increases from 52.3% to 54.3%, whereas the tool category decreases from 45.7% to 43.1%. The adaptation therefore benefits several external tasks while leaving some capabilities largely unchanged and introducing regressions in others. These differences support evaluating transfer across both benchmarks and question categories, rather than treating egocentric understanding as a single capability.

Performance Across Temporal Extents. On EgoPolice, our model exceeds Qwen3.5-9B (Qwen Team, 2026) by 7.6, 8.3, and 11.3 percentage points at the 1 s, 10 s, and 60 s window lengths, respectively. The larger margin at 60 s coexists with lower absolute accuracy: our scores decrease from 66.7% at 1 s to 54.3% at 60 s. Relative improvement therefore does not imply that longer-window understanding is resolved. Since the three settings contain different evaluation items, this

Table 20: **The Complete Results on EgoMMU Benchmark.** We evaluate all unimodal and multimodal egocentric scene understanding models and their variants across five task families and different task-modality combinations. Higher scores indicate better performance in every column.

Model	Size	Type	MCQ			Narr	Fact Check			Open QA			Temporal		
			V	A	I		V	A	I	V	A	I	V	A	I
Group: Open-Source Omni Models															
Qwen3-Omni	30B	V A	63.9	59.8	37.6	39.3	54.2	54.5	53.6	56.4	42.8	47.7	33.3	11.6	13.3
Gemma 4	12B	V A	60.6	41.3	35.2	36.9	52.7	43.3	50.9	53.2	33.0	38.3	14.2	8.6	8.3
Gemma 4	4B	V A	53.1	33.2	32.3	32.9	43.6	31.1	40.3	33.9	32.5	30.9	11.5	7.1	6.3
Gemma 4	2B	V A	43.9	29.7	29.7	31.0	36.7	26.8	25.3	27.0	25.6	29.9	11.0	5.2	4.2
Intern-Omni	9B	V A	59.1	53.7	36.1	40.9	51.6	45.5	39.7	44.7	36.1	43.7	10.6	7.9	5.7
MiniCPM-o	8B	V A	54.9	45.8	36.1	40.0	55.0	44.3	32.7	50.1	35.9	34.3	15.4	7.9	4.5
Qwen2.5-Omni	7B	V A	56.9	46.6	39.4	41.7	43.7	46.7	33.9	44.1	36.7	40.9	14.1	10.8	11.5
Qwen2.5-Omni	3B	V A	51.9	39.7	31.0	34.3	36.5	40.1	34.2	29.8	30.3	31.6	9.8	10.0	10.8
VideoLLaMA2.1	7B	V A	54.7	44.1	35.9	36.8	50.2	37.2	43.1	36.7	35.1	43.5	16.2	8.7	10.8
Baichuan-Omni-1.5	7B	V A	52.1	43.7	34.7	33.1	49.9	42.6	33.1	39.1	34.3	40.7	9.3	5.4	6.1
EgoGPT	7B	V A	57.3	52.8	35.1	42.7	46.8	52.1	39.6	45.3	40.7	38.0	19.5	12.1	8.7
Phi-4-Multimodal	6B	V A	50.3	36.7	34.1	25.3	37.7	38.6	40.1	24.7	35.8	36.6	9.1	9.8	10.0
EgoMind	9B	V A I	70.3	67.3	66.2	52.1	61.6	62.0	66.7	61.9	52.6	62.2	34.5	16.7	20.3
EgoMind	4B	V A I	61.5	55.1	47.2	44.7	54.7	53.8	57.1	56.2	44.8	52.8	29.5	14.0	18.5
EgoMind	2B	V A I	56.7	48.9	43.8	39.3	47.3	50.7	49.2	42.3	38.0	43.5	25.2	10.5	14.7
EgoMind	1B	V A I	50.5	39.4	39.0	36.1	41.9	46.1	44.7	37.2	35.7	39.4	18.1	10.3	12.5
Group: Open-Source Vision-Only Models															
InternVL3	78B	V	55.6	30.2	38.4	37.2	52.0	35.9	45.8	46.8	30.4	43.5	17.4	11.3	9.8
InternVL3	38B	V	53.1	30.5	37.2	36.7	46.7	30.1	44.7	42.1	29.8	44.1	14.7	9.0	10.8
InternVL3	8B	V	51.4	23.6	40.7	35.7	39.7	21.0	47.2	43.6	33.6	42.0	13.8	5.9	5.2
InternVL3	2B	V	49.2	23.0	30.1	35.0	39.0	18.2	31.5	42.8	31.2	41.5	11.9	4.7	2.5
Qwen3.5	35B	V	63.7	24.3	38.7	34.7	54.7	32.4	46.1	50.7	30.8	43.8	28.1	7.0	5.3
Qwen3.5	9B	V	56.2	25.9	37.2	30.1	49.1	29.4	46.7	45.6	29.8	44.6	24.7	5.1	4.9
Qwen3.5	4B	V	50.3	23.9	34.5	26.7	44.2	24.4	35.0	32.0	22.8	39.9	14.0	4.7	4.0
Qwen3.5	2B	V	40.8	21.6	29.2	28.1	38.2	24.1	32.3	29.6	25.2	34.2	11.9	4.1	2.1
Qwen3-VL	32B	V	54.7	24.4	38.5	30.5	51.5	36.8	51.3	32.4	32.0	43.7	18.7	6.8	4.9
Qwen3-VL	8B	V	52.5	22.7	38.3	29.3	50.8	36.4	45.4	30.2	30.4	44.0	15.0	6.1	4.6
Qwen3-VL	4B	V	49.0	20.4	38.1	28.8	44.0	29.2	43.6	30.0	24.4	42.0	13.8	5.7	3.1
Qwen3-VL	2B	V	40.5	19.2	31.9	27.7	34.8	22.7	25.4	26.8	22.4	38.3	12.2	3.1	2.7
Qwen3-Omni	30B	V	63.1	29.9	42.0	34.1	52.7	36.1	53.1	57.1	29.7	44.3	29.7	8.5	11.4
Qwen2.5-VL	72B	V	54.7	21.6	42.7	35.5	51.5	26.0	45.2	38.8	34.8	35.8	17.9	8.6	8.8
Qwen2.5-VL	7B	V	51.2	21.0	40.3	33.7	50.2	19.0	42.1	36.8	30.8	30.0	16.8	7.0	4.7
Qwen2.5-VL	3B	V	46.4	20.7	33.6	22.6	45.1	19.9	40.3	22.0	22.0	24.9	10.6	6.6	3.0
Qwen2.5-Omni	7B	V	55.4	19.5	38.1	39.4	44.1	17.7	36.8	42.3	26.7	40.3	15.3	7.7	11.2
Qwen2.5-Omni	3B	V	52.3	20.6	33.7	32.9	36.0	16.8	34.2	30.0	25.1	32.7	10.3	6.1	11.2
Gemma 4	12B	V	56.1	19.0	29.8	34.2	44.3	31.8	43.9	49.8	28.4	35.9	13.8	4.8	8.7
Gemma 4	4B	V	49.3	18.1	27.3	29.7	42.1	17.9	39.6	30.6	12.0	32.1	12.3	5.3	6.8
Gemma 4	2B	V	44.7	18.7	33.7	29.1	34.0	12.8	28.8	28.3	10.4	34.5	10.8	5.7	6.6
Group: Proprietary Models															
Gemini 3.8 Flash	–	V A	79.1	66.1	35.8	51.8	71.6	70.7	55.2	70.8	60.1	50.3	31.7	15.0	19.8
Gemini 3.7 Flash	–	V A	79.3	69.7	37.6	52.8	69.8	71.0	57.3	72.0	59.8	50.8	34.0	14.3	19.3
Gemini 3.6 Flash	–	V A	78.2	64.4	35.4	53.5	67.1	72.8	58.3	69.6	56.7	43.0	36.7	15.4	16.7
Gemini 3.5 Flash	–	V A	76.8	59.7	35.4	50.4	68.0	71.9	56.8	67.6	57.2	36.8	35.3	13.8	16.7
Gemini 3.1 Pro	–	V A	78.2	65.0	45.4	46.2	70.3	70.4	49.3	68.0	52.0	46.3	39.7	16.9	15.3
Gemini 3.1 Flash Lite	–	V A	71.2	51.0	34.8	44.7	66.2	62.8	44.7	62.1	44.3	34.7	22.4	14.7	15.2
Gemini 2.5 Pro	–	V A	72.7	55.3	40.9	43.7	64.1	60.6	40.1	63.9	46.1	35.3	23.7	15.0	16.7
Gemini 2.5 Flash Lite	–	V A	66.1	49.7	30.7	40.8	56.6	52.6	35.1	52.3	40.7	36.9	23.7	11.2	13.7
GPT 5.4	–	V	72.3	28.7	44.1	51.8	70.5	32.3	52.8	72.2	32.1	50.1	28.9	9.5	14.9
GPT 5.4 Mini	–	V	64.8	20.1	31.6	49.6	62.9	33.7	54.6	54.3	30.7	41.3	20.4	8.2	12.0
GPT 5.4 Nano	–	V	57.1	18.4	29.6	30.3	54.0	29.8	43.5	50.8	24.1	44.2	15.6	5.7	8.1
GPT 5	–	V	73.7	26.2	34.4	50.1	66.6	30.7	54.8	70.7	26.4	52.8	31.8	6.1	17.4
GPT 5 Mini	–	V	73.0	25.4	30.1	48.9	59.8	25.4	50.0	55.0	24.7	40.3	25.0	8.7	13.1
GPT 5 Nano	–	V	51.7	21.6	29.2	36.6	52.3	24.1	40.9	48.7	19.8	37.7	10.7	6.7	9.9
Opus 5	–	V	75.1	30.7	40.2	52.8	72.6	32.1	64.7	70.1	42.7	45.8	38.7	7.6	12.8
Opus 4.8	–	V	67.0	31.3	34.5	41.1	70.0	24.4	63.9	68.8	36.8	37.1	25.8	8.1	14.4
Opus 4.6	–	V	62.8	30.2	29.7	40.0	66.2	27.3	60.0	60.1	30.7	32.4	19.5	6.6	7.1
Sonnet 5	–	V	63.7	29.7	26.1	42.7	67.7	26.1	61.6	56.6	33.6	33.2	26.9	4.3	10.7
Sonnet 4.6	–	V	66.5	30.5	28.7	38.9	64.3	26.7	62.6	54.2	32.8	30.6	18.9	3.5	9.1
Haiku 4.5	–	V	45.5	20.4	23.0	27.0	44.7	28.8	47.8	43.2	26.4	35.8	18.0	3.3	5.4

pattern characterizes performance across temporal extents rather than isolating the causal effect of adding context to the same question.

E.2 Modality Dependence and Ablations

We examine whether benchmark performance reflects access to the modality that a question is intended to assess. We first compare open-source omni-modal models with and without audio on EgoMMU, then examine vision-only and audio-visual results on EgoAVU-Bench (Seth et al., 2026). Finally, we extend the controlled comparison to all three modalities using **EgoMind**. These analyses distinguish overall task performance from the contribution of the corresponding sensor input.

Table 21: **The Complete Results on EgoAVU-Bench.** We evaluate all models and their variants on segment narration (AVSN), sound-source attribution (SSA), and audio-visual hallucination (AVH). Higher scores indicate better performance, including fewer hallucinations. Aggregate results exclude questions whose referenced events fall outside the provided clips, following Section C.1.

Model	Size	Type	AVSN-R	AVSN-S	SSA	AVH				Overall (LLM Judge)
						Action	Sound	Object	Average	
Open-Source Omni Models										
Qwen3-Omni	30B	V A	14.81	1.57	1.40	3.14	2.55	2.03	2.57	2.14
Intern-Omni	9B	V A	14.70	1.82	1.47	2.37	2.41	2.27	2.35	2.07
MiniCPM-o	8B	V A	12.19	2.06	1.43	2.41	2.47	2.21	2.36	2.12
Baichuan-Omni	7B	V A	14.21	1.79	1.49	2.60	2.44	2.31	2.45	2.13
EgoAVU†	7B	V A	28.70	2.63	3.20	–	–	–	–	–
Qwen2.5-Omni	7B	V A	13.39	1.99	1.50	2.80	2.42	2.33	2.52	2.21
Qwen2.5-Omni	3B	V A	13.08	1.85	1.45	2.65	2.53	2.37	2.52	2.17
VideoLLaMA2	7B	V A	13.89	1.71	1.51	3.23	2.58	2.41	2.74	2.29
EgoGPT	7B	V A	14.30	1.52	1.37	2.68	2.39	2.37	2.48	2.07
Phi4-Multimodal	6B	V A	16.90	1.69	1.42	1.90	2.10	2.13	2.04	1.85
EgoMind†	9B	V A I	26.77	2.41	3.27	3.61	2.98	2.87	3.15	3.03
EgoMind	9B	V A I	17.53	1.73	1.61	3.27	2.71	2.58	2.85	2.38
Open-Source Vision-Only Models										
Qwen3.5	35B	V	16.21	1.71	1.59	3.11	2.50	2.21	2.61	2.22
Qwen3.5	9B	V	16.47	1.69	1.48	3.32	2.62	1.87	2.60	2.20
Qwen3.5	4B	V	16.27	1.64	1.41	3.21	2.44	1.97	2.54	2.13
Qwen3.5	2B	V	16.18	1.62	1.44	3.04	2.28	1.80	2.37	2.04
Qwen3-VL	32B	V	15.01	1.74	1.61	3.07	2.63	2.27	2.66	2.26
Qwen3-VL	8B	V	14.87	1.73	1.50	3.02	2.58	1.90	2.50	2.15
Qwen3-VL	4B	V	14.36	1.66	1.48	3.11	2.53	1.83	2.49	2.12
Qwen3-VL	2B	V	13.82	1.58	1.44	3.03	2.42	2.07	2.51	2.11
Qwen2.5-VL	72B	V	15.61	1.87	1.61	3.09	2.21	2.27	2.52	2.21
Qwen2.5-VL	7B	V	14.64	1.77	1.63	2.83	1.79	2.20	2.27	2.04
Qwen2.5-VL	3B	V	14.21	1.52	1.33	1.80	1.86	2.17	1.94	1.74
InternVL3	38B	V	16.41	1.75	1.57	3.13	2.47	2.58	2.73	2.30
InternVL3	8B	V	12.89	1.44	1.42	2.71	2.31	2.63	2.55	2.10
InternVL3	2B	V	11.91	1.44	1.48	1.91	1.95	1.97	1.94	1.75
Proprietary Models										
Gemini 3.8 Flash	–	V A	15.98	1.83	2.27	3.37	2.88	2.61	2.95	2.59
Gemini 3.7 Flash	–	V A	16.41	1.85	2.11	3.41	2.87	2.68	2.99	2.58
Gemini 3.6 Flash	–	V A	16.17	1.59	2.08	3.51	2.71	2.67	2.96	2.51
Gemini 3.5 Flash	–	V A	16.83	1.67	1.87	3.27	2.83	2.45	2.85	2.42
Gemini 3.1 Pro	–	V A	15.37	1.71	2.31	3.66	3.01	2.57	3.08	2.65
Gemini 3.1 Flash Lite	–	V A	15.21	1.77	2.01	3.28	2.67	2.26	2.74	2.40
GPT 5.4	–	V	16.44	1.61	2.20	3.51	2.61	2.31	2.81	2.45
GPT 5.4 Mini	–	V	14.23	1.64	2.13	3.47	2.68	2.41	2.85	2.47
Opus 5	–	V	15.37	2.03	2.07	3.33	2.77	2.53	2.88	2.55
Opus 4.6	–	V	15.31	1.93	1.87	3.23	2.53	2.50	2.75	2.41
Sonnet 5	–	V	16.10	2.17	1.81	3.21	2.41	2.41	2.68	2.40
Sonnet 4.6	–	V	16.37	2.11	1.77	3.04	2.27	2.27	2.53	2.29
Haiku 4.5	–	V	11.53	1.87	1.42	2.77	2.41	2.31	2.50	2.16

Controlled Audio Ablations on EgoMMU. Table 20 includes six omni-modal model variants evaluated with both video and audio and with video alone. Removing audio reduces audio multiple-choice, fact-checking, and open-ended scores for every variant, by 11.0-29.9, 11.5-29.0, and 4.6-20.5 points, respectively.

For Qwen3-Omni-30B (Qwen Team, 2025a), audio multiple-choice accuracy decreases from 59.8% to 29.9%, while visual multiple-choice accuracy remains close (63.9% versus 63.1%). Its visual fact-checking score changes from 54.2 to 52.7, and visual open-ended accuracy from 56.4% to 57.1%. Qwen2.5-Omni-7B (Xu et al., 2025b) shows the same contrast: audio multiple-choice accuracy falls from 46.6% to 19.5%, while visual accuracy changes from 56.9% to 55.4%. The concentration of losses in audio tasks supports their dependence on acoustic evidence, rather than a general degradation caused by changing the input configuration.

Visual Performance on an Audio-Visual Benchmark. The results on EgoAVU-Bench (Seth et al., 2026) in Table 21 reveal a different pattern. Several vision-only models match or exceed audio-visual models on individual metrics, including sound-source attribution. For example, vision-only Qwen3.5-9B (Qwen Team, 2026) scores 1.48 on sound-source attribution and 2.20 overall, compared with 1.40 and 2.14 for Qwen3-Omni-30B (Qwen Team, 2025a) with audio enabled.

Likewise, vision-only Qwen2.5-VL-7B (Bai et al., 2025) achieves a sound-source attribution score of 1.63, exceeding the 1.50 of Qwen2.5-Omni-7B (Xu et al., 2025b). These comparisons show that higher scores on an audio-visual benchmark do not necessarily identify models with stronger

Table 22: **The Complete Results on External Egocentric Benchmarks.** We evaluate all models and their variants on EgoSchema (Mangalam et al., 2023), EgoBlind (Xiao et al., 2025), EgoThink (Cheng et al., 2023), and EgoPolice (Gonzalez Saez-Diez et al., 2026) using the protocols in Section C. EgoBlind results are reported by question category, and EgoPolice results by temporal window length. Higher scores indicate better performance in every column.

Model	Size	Type	EgoSchema	EgoBlind						EgoThink	EgoPolice		
				inform	safety	navig	resour	tool	commu		1s	10s	60s
Group: Open-Source Omni Models													
Qwen3-Omni	30B	V A	61.0	55.6	61.0	28.6	40.4	40.0	40.5	65.6	56.5	58.5	55.2
Gemma 4	12B	V A	68.4	50.5	46.6	26.5	36.4	30.0	45.9	64.3	53.1	56.0	48.4
Gemma 4	4B	V A	62.0	42.3	33.6	15.8	29.2	22.9	18.9	55.0	46.1	48.3	40.7
Gemma 4	2B	V A	54.4	35.9	42.3	18.9	23.7	20.0	22.1	53.7	44.3	47.5	41.8
Qwen2.5-Omni	7B	V A	62.4	47.2	42.3	30.7	32.3	31.2	35.1	62.6	52.6	53.2	46.1
Qwen2.5-Omni	3B	V A	54.8	42.7	47.6	18.4	25.8	30.4	38.9	61.3	47.6	47.1	35.8
EgoGPT	7B	V A	75.4	46.4	59.0	22.4	36.4	22.9	43.2	61.4	53.8	54.3	50.4
Phi-4-Multimodal	6B	V A	59.2	44.8	40.0	23.0	31.3	28.6	38.6	58.7	51.4	46.6	39.3
EgoMind	9B	V A I	72.1	54.3	63.1	28.6	42.7	43.1	47.7	73.6	66.7	62.0	54.3
Group: Open-Source Vision-Only Models													
Qwen3.5	9B	V	67.1	52.3	61.6	27.6	40.4	45.7	45.9	73.4	59.1	53.7	43.0
Qwen3.5	4B	V	58.4	48.4	56.1	28.1	40.4	41.4	40.5	67.7	55.0	48.1	36.9
Qwen3.5	2B	V	50.4	42.2	57.4	19.9	37.4	22.9	38.5	67.1	45.1	42.5	36.8
Qwen3-VL	32B	V	64.0	51.0	58.6	30.7	42.1	37.6	44.3	66.9	56.0	53.3	40.0
Qwen3-VL	8B	V	59.4	47.7	52.1	28.6	40.4	37.1	45.9	68.7	56.6	52.1	41.8
Qwen3-VL	4B	V	57.6	46.5	54.8	25.5	39.4	35.7	39.5	65.8	50.3	50.5	40.6
Qwen3-VL	2B	V	47.8	43.9	50.2	17.9	31.3	25.7	35.1	65.3	51.9	47.5	37.3
InternVL3	8B	V	68.2	52.7	60.3	32.8	36.4	32.9	48.6	62.3	53.8	54.4	40.6
InternVL3	2B	V	51.6	51.2	56.7	28.1	29.3	27.1	40.6	59.7	41.3	44.8	37.0
Group: Proprietary Models													
Gemini 3.8 Flash	–	V A	79.0	64.9	59.7	45.9	58.6	58.6	59.5	77.1	81.6	84.2	81.8
Gemini 3.1 Pro	–	V A	76.5	60.0	61.1	47.2	48.5	62.0	51.4	75.1	82.3	83.0	79.7
Gemini 3.1 Flash Lite	–	V A	73.4	54.4	52.3	36.6	42.3	45.3	47.7	72.7	75.1	76.7	73.2
GPT 5.4	–	V	78.2	62.7	62.5	45.4	47.5	60.0	53.2	74.6	75.6	78.9	76.3
GPT 5.4 Mini	–	V	73.4	51.1	49.0	31.7	42.5	44.1	43.8	69.2	65.1	67.7	53.0
Opus 5	–	V	76.7	58.7	66.1	45.4	47.0	55.3	52.8	64.7	72.8	76.3	70.1
Opus 4.6	–	V	73.8	46.6	59.5	39.3	43.0	49.7	43.0	65.7	67.1	68.9	66.1
Sonnet 5	–	V	74.0	44.5	64.9	37.3	46.3	40.0	44.3	68.4	63.4	65.7	60.8
Sonnet 4.6	–	V	72.7	43.4	59.7	37.3	44.1	42.7	40.5	66.9	61.4	63.6	57.1
Haiku 4.5	–	V	72.1	38.4	54.4	36.7	45.5	42.4	37.0	64.3	55.4	59.1	52.0

audio understanding. Visible actions and objects may support plausible predictions about associated sounds, allowing visual inference to contribute to nominally acoustic tasks. Differences between models prevent a causal attribution, but the results highlight a limitation of interpreting these scores as isolated measures of audio perception.

Task Format Affects Modality Sensitivity. Even on EgoMMU, the effect of removing audio varies across task formats. The consistent decreases in audio multiple choice, fact checking, and open-ended answering contrast with changes of at most 3.9 points in audio temporal grounding for the six paired baselines. Since grounding scores are already low, this smaller change does not establish that audio is unnecessary; it may instead reflect limited localization ability even when audio is available. Narration decreases for every baseline by 1.4-5.2 points, consistent with references that include acoustic information alongside visual content.

Modality sensitivity therefore needs to be examined jointly with absolute performance and task requirements, rather than inferred from a single aggregate score.

Extending the Analysis to Three Modalities. Table 23 evaluates EgoMind at four sizes with full inputs and with audio, video, or IMU withheld. Each comparison keeps the checkpoint, evaluation items, and remaining inputs unchanged; removing IMU also removes its tag text. The losses are concentrated in the columns associated with the withheld stream. At 9B, removing audio reduces audio multiple-choice accuracy from 67.3% to 36.7%, while removing IMU reduces IMU multiple-choice accuracy from 66.2% to 34.7%.

Across the three larger sizes, removing IMU changes vision and audio scores by at most 2.5 points. Narration decreases under every modality ablation at every size. Together, these results support both the modality dependence of the individual task columns and the contribution of complementary inputs to the shared narration task.

Implications for Multimodal Evaluation. The presence of multiple streams in a benchmark does not by itself establish that its questions require each stream. A more informative evaluation reports modality-specific tasks alongside controlled input ablations, asking whether removing the relevant

Table 23: **Inference-Time Modality Ablations.** At each model size, we evaluate the same checkpoint with all three modalities and with Vision, Audio, or IMU withheld. The "IMU-Off" setting removes both signal tokens and tag text. Evaluation items and remaining inputs are unchanged across settings. Higher scores indicate better performance in every column.

Model	Type	MCQ			Narr	Fact Check			Open QA			Temporal		
		V	A	I		V	A	I	V	A	I	V	A	I
Model Size: 9B														
EgoMind	V A I	70.3	67.3	66.2	52.1	61.6	62.0	66.7	61.9	52.6	62.2	34.5	16.7	20.3
Audio-Off	V I	67.9	36.7	67.4	45.3	58.0	44.3	69.3	62.8	28.3	63.0	28.7	9.5	19.6
Vision-Off	A I	31.2	65.7	64.3	40.9	33.9	53.8	67.0	41.0	54.3	61.1	10.2	15.1	18.3
IMU-Off	V A	69.7	66.8	34.7	46.1	62.0	63.1	44.5	59.8	55.1	43.8	33.7	17.1	5.7
Model Size: 4B														
EgoMind	V A I	61.5	55.1	47.2	44.7	54.7	53.8	57.1	56.2	44.8	52.8	29.5	14.0	18.5
Audio-Off	V I	59.4	25.4	45.1	34.7	54.1	32.7	59.3	55.0	25.3	53.1	27.6	5.0	19.3
Vision-Off	A I	28.7	52.3	44.0	30.2	28.7	50.3	54.1	30.9	40.3	47.1	8.1	12.0	16.7
IMU-Off	V A	62.3	54.7	33.1	37.0	55.1	52.9	34.3	55.3	42.9	37.8	28.8	13.8	5.3
Model Size: 2B														
EgoMind	V A I	56.7	48.9	43.8	39.3	47.3	50.7	49.2	42.3	38.0	43.5	25.2	10.5	14.7
Audio-Off	V I	54.3	22.0	44.0	33.6	45.8	24.3	50.1	44.1	26.1	44.4	26.1	5.4	17.3
Vision-Off	A I	29.3	46.1	40.3	30.3	25.3	45.1	44.7	27.1	33.4	45.3	8.7	9.3	12.3
IMU-Off	V A	55.3	47.7	34.9	36.0	45.7	49.8	30.1	43.3	36.1	34.9	27.7	10.5	5.7
Model Size: 1B														
EgoMind	V A I	50.5	39.4	39.0	36.1	41.9	46.1	44.7	37.2	35.7	39.4	18.1	10.3	12.5
Audio-Off	V I	51.1	23.7	37.1	28.3	40.2	23.3	46.3	38.3	24.0	38.7	16.3	4.1	12.8
Vision-Off	A I	22.1	20.1	33.3	25.1	20.3	39.7	44.1	26.7	30.3	37.1	6.3	8.7	9.3
IMU-Off	V A	49.8	38.7	30.1	24.7	38.8	40.3	31.3	38.2	33.7	30.0	16.5	11.3	4.0

evidence selectively reduces performance. EgoMMU exhibits this pattern across several model families and, with EgoMind, across all three modalities.

Nevertheless, residual performance after removal shows that the tasks are not perfectly isolated from shared context or learned priors. We therefore interpret the ablations as evidence of modality dependence at the task level, rather than proof that every item is answerable exclusively from its designated stream.

E.3 Encoder Pretraining

We analyze the pretraining stages in Section D.3 through their optimization dynamics, held-out captioning performance, and sensitivity to the input signal. We then examine whether pretraining improves subsequent instruction tuning and downstream task performance.

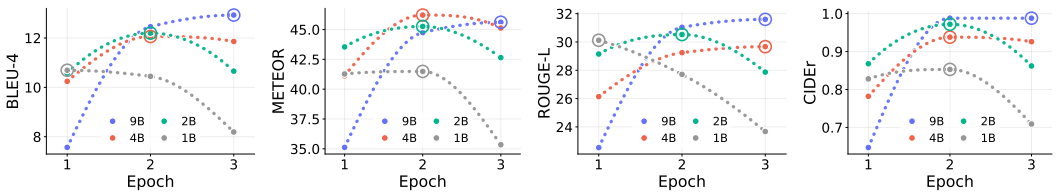
E.3.1 Audio Alignment

Figure 3 reports the Phase A training loss defined in Eq. 17 across the four model sizes. The curves follow similar trajectories, with 83-92% of the total loss reduction occurring during the first epoch. After that epoch, larger backbones achieve lower loss at every logged step. Most of the optimization progress therefore occurs early, while the persistent separation between model sizes indicates that the frozen backbone affects how well the trained audio pathway predicts the reference captions.

Training loss alone does not establish whether the pathway uses clip-specific acoustic evidence: predictable caption patterns can also reduce the generation loss. We therefore complement these curves with the input-substitution probes in Section E.3.4 and the downstream initialization ablation in Section E.3.5.

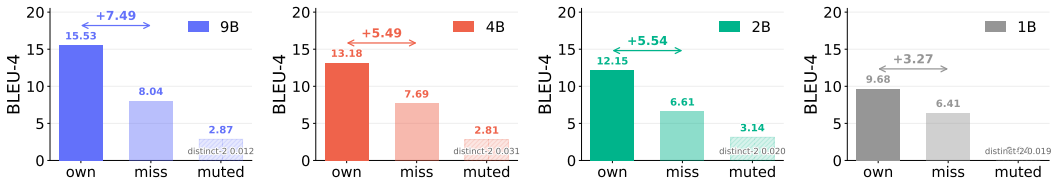
E.3.2 IMU Contrastive Pretraining

The contrastive step in Eq. 20 trains a shared IMU encoder, which initializes the subsequent alignment stage at all four backbone sizes, as described in Section D.3.2. Figure 4 shows that the contrastive loss decreases from 3.509 to 0.357 over 50 epochs, while top-1 retrieval accuracy reaches 86.5% in both directions.



(a) BLEU-4 (b) METEOR (c) ROUGE-L (d) CIDEr

Figure 8: **Held-Out Captioning during IMU Alignment.** We report four captioning metrics after each training epoch across the four model sizes. Markers denote measured scores, dotted lines connect epochs, and rings indicate each curve's maximum. CIDEr is plotted on a separate scale.



(a) EgoMind-9B (b) EgoMind-4B (c) EgoMind-2B (d) EgoMind-1B

Figure 9: **Signal Specificity of the IMU Pathway.** We compare captioning performance using the clip's own IMU signal (*own*), another clip's signal (*miss*), and a zero-valued signal (*muted*). The "own"-"miss" gap measures input specificity. The muted condition is reported separately because it produces repetitive outputs.

Within-Batch Retrieval. Retrieval is evaluated within batches of 64 paired examples, giving a top-1 chance level of 1.56%. After recall at 5 first reaches 1.000 at epoch 25, its mean over subsequent logged batches is 0.990, and the median rank remains 1. Mean reciprocal rank continues to improve, reaching 0.88. Together, these measurements indicate that the encoder learns to distinguish the paired caption from other candidates within the batch. They do not establish retrieval performance over the full dataset, where the candidate set is larger and may contain more similar descriptions. The two retrieval directions differ by up to 2.7 percentage points during training and coincide at the final epoch; this equality is an empirical observation rather than a requirement of the symmetric objective.

Embedding Alignment and Uniformity. We further examine the embedding geometry using alignment and uniformity (Wang & Isola, 2020). Alignment measures the mean squared distance between paired normalized embeddings, equivalent to $2 - 2\cos$ for each pair. Uniformity measures the logarithm of the mean pairwise Gaussian potential, with lower values indicating a more dispersed embedding distribution under this measure.

Alignment initially improves from 2.049 to 0.470 at epoch 5.5, then increases to 0.872 at the final epoch, corresponding to a mean matched-pair cosine similarity of approximately 0.56. Uniformity decreases from -0.596 to -3.416 throughout training. Thus, improved retrieval does not require matched embeddings to become progressively closer in absolute distance: later training improves their relative ranking while the embedding distribution becomes more dispersed. The learned similarity scale $1/\kappa$ also increases from approximately 14.3 to 29.9, so the contrastive loss reflects changes in both embedding geometry and similarity scaling.

E.3.3 IMU Alignment

Figure 10 reports the caption-alignment loss in Eq. 21. Across model sizes, 89-92% of the total loss reduction occurs during the first epoch, and larger backbones achieve lower final losses. These trajectories show that the contrastively initialized encoder can be adapted to caption generation early in training. However, the curves alone do not isolate the contribution of contrastive initialization, since that would require a matched comparison with caption alignment from a randomly initialized encoder.

Table 24: **Captioning Probes with Pretrained and Random Initialization.** We compare held-out captioning scores after audio and IMU alignment with those of randomly initialized trainable pathways of the same architecture. Higher scores indicate better performance in every metric.

Configuration		Phase A (Pretraining: Audio Alignment)				Phase I (Pretraining: IMU Alignment)				Instruction Tuning				
		BLEU	METEOR	ROUGE-L	CIDEr	BLEU	METEOR	ROUGE-L	CIDEr	MCQ	Narr	FC	OQA	TG
9B	Rand. Init.	0.0	1.6	0.8	0.024	0.0	2.3	2.4	0.013	40.1	30.7	42.3	39.5	11.4
	Ours	3.7	28.3	21.2	0.552	12.9	45.6	31.6	0.988	67.9	52.1	63.4	58.9	23.8
4B	Rand. Init.	0.0	0.7	0.7	0.019	0.0	1.6	2.0	0.011	36.7	27.8	42.0	31.3	7.8
	Ours	3.2	26.4	19.3	0.526	11.9	45.1	29.7	0.926	54.6	44.7	55.2	51.3	20.7
2B	Rand. Init.	0.0	0.4	0.6	0.012	0.0	1.3	1.6	0.010	30.5	26.7	28.1	29.7	6.0
	Ours	2.8	22.7	17.4	0.453	10.7	42.7	27.9	0.862	49.8	39.3	49.1	41.3	16.8
1B	Rand. Init.	0.0	0.1	0.3	0.007	0.0	0.2	0.4	0.010	22.3	24.8	23.1	20.9	5.2
	Ours	2.4	22.9	16.2	0.427	8.2	35.3	23.7	0.709	43.0	36.1	44.2	37.4	13.6

Held-Out Captioning. Figure 8 shows that continued optimization does not consistently improve held-out caption quality. At 9B, the final epoch achieves the highest score on three of the four metrics. At 4B and 2B, the second epoch performs best, while the 1B scores peak during the first or second epoch and decrease on all four metrics by the third. For example, 1B BLEU-4 changes from 10.69 to 10.45 and then 8.19.

Across the 16 model-size and metric combinations, the final epoch ranks highest in five and lowest in seven. This divergence between training loss and held-out scores suggests that later optimization may increasingly fit the training captions without improving generalization, particularly at smaller sizes. The probes contain only 32-64 clips, however, so they provide limited evidence for selecting size-specific stopping points. We use the final epoch consistently across sizes rather than selecting checkpoints separately on these small probes.

E.3.4 Signal Specificity

We test whether generated captions depend on the corresponding sensor input using the specificity criterion in Eq. 18. Each pathway is evaluated on 32 held-out clips under three conditions: the clip's own signal (*own*), a signal from another clip (*miss*), and a zero-valued signal (*muted*). Model weights, prompts, and references remain fixed. The own-miss difference measures the change in reference agreement when the input no longer corresponds to the target clip.

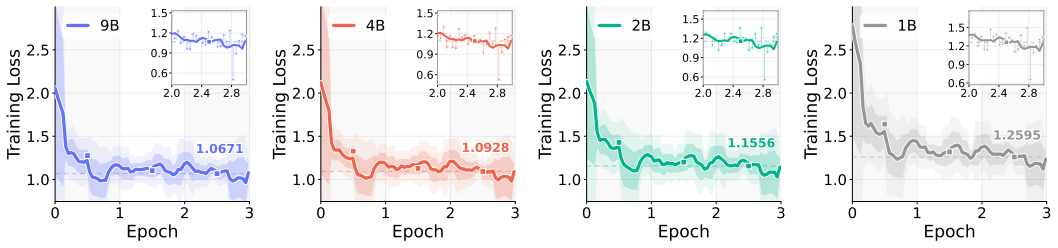
Dependence on the Recorded Signal. Figure 9 shows positive IMU specificity at every model size. The BLEU-4 gaps between own and miss inputs are 7.49, 5.49, 5.54, and 3.27 from 9B to 1B, while own-input scores range from 15.53 at 9B to 9.68 at 1B. The positive gaps indicate that correct signal-caption pairing contributes to generation quality, although specificity does not increase strictly with model size.

For audio, the corresponding BLEU-4 gap at 9B is 1.11. This smaller gap indicates weaker separation under this captioning probe, but it does not directly measure the relative downstream value of audio and IMU. The modalities differ in their caption vocabulary, reference variability, and compatibility with lexical-overlap metrics. The controlled task evaluations in Section E.2 provide complementary evidence of how each input affects task performance.

Zero-Signal Behavior. The muted condition produces repetitive outputs in the evaluated runs. Distinct-2 decreases from approximately 0.43 with own inputs to 0.012-0.031, repeated 4-grams reach 58% of the output, and generated lengths depart from reference lengths. A zero-valued input therefore changes both the available information and generation behavior. It is also different from omitting the modality block, as done in the inference-time ablations. We use mismatched recorded signals for the specificity criterion because they preserve a realistic sensor input while breaking its correspondence with the reference, and report the muted condition separately.

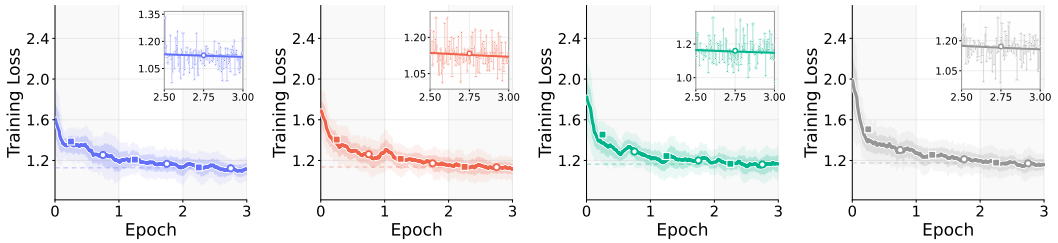
E.3.5 Initialization Ablation

Table 24 compares pretrained and randomly initialized sensor pathways of the same architecture at each model size. Before instruction tuning, randomly initialized pathways score 0.0 BLEU on the captioning probes, whereas pretrained pathways achieve 2.4-3.7 for audio and 8.2-12.9 for IMU.



(a) EgoMind-9B (b) EgoMind-4B (c) EgoMind-2B (d) EgoMind-1B

Figure 10: **Effect of IMU Caption Alignment.** Training loss across four model sizes after shared contrastive pretraining. The IMU encoder and backbone-specific projection are optimized through a frozen language model using Eq. 21.



(a) Both Pretrain (b) Audio Pretrain only (c) IMU Pretrain only (d) No Pretrain

Figure 11: **Effect of Pretraining on Instruction Tuning.** Training loss for the 9B model under four combinations of pretrained and randomly initialized audio and IMU pathways. All settings use the same instruction-tuning data, update budget, and trainable components.

More importantly, after instruction tuning, pretrained initialization yields higher scores on all five task families at every size. At 9B, multiple-choice performance is 67.9 versus 40.1, and temporal grounding is 23.8 versus 11.4. These comparisons show that the benefit of pretraining persists beyond the captioning stage under the evaluated instruction-tuning schedule.

Alignment Before Instruction Tuning. Figure 10 shows the caption-alignment stage that produces the pretrained IMU pathways used in this ablation. All four runs start from the same contrastively pretrained IMU encoder, then jointly optimize the encoder and a projection matched to the language backbone. Larger backbones reach lower final caption losses despite the shared encoder initialization. The differences therefore reflect subsequent adaptation together with the language backbone, rather than different initial contrastive representations. Table 24 evaluates the combined benefit of audio alignment and the two-stage IMU pretraining procedure under the same instruction-tuning schedule.

Separate Contributions of Audio and IMU Pretraining. We further evaluate a 2×2 initialization study at 9B, with each pathway initialized either from pretrained weights or randomly. All four settings use the same instruction-tuning data, number of updates, and trainable components. Figure 11 reports their training trajectories. The loss gap between fully pretrained and fully random initialization is 0.084, 0.046, and 0.048 at the three epoch boundaries. Instruction tuning reduces the initial difference but does not eliminate it, indicating that random initialization does not recover the same training loss within the allocated schedule. These loss comparisons characterize optimization; downstream task improvements are assessed separately in Table 24.

Role of IMU Tag Text. All four settings retain the deterministic IMU tags. Consequently, the random-IMU setting still receives motion information through text, and the comparison measures the additional benefit of initializing the learned IMU pathway with pretrained weights in the presence of those tags. It does not measure the total contribution of IMU information or establish a lower bound on pretraining benefits without tags. Separating these effects would require an additional comparison that varies tag availability alongside encoder initialization.

E.4 Qualitative Examples

Figures 12–24 present examples from all thirteen EgoMMU task-modality combinations. Figures 25–27 and Figure 28 extend the analysis to EgoAVU-Bench and EgoBlind. Each figure pairs representative media with questions or claims, reference answers, and model predictions. These selected cases illustrate specific behaviors and complement the aggregate results in Section E.1 and controlled modality ablations in Section E.2.

Visual Details and Wearer Motion. Figures 12, 16, and 19 examine object transfers, action order, and the resolution of a dropped handle. Several incorrect responses recognize the relevant objects but reverse event order or confuse a resting hand with a grip. For example, recognizing the cyclist and pedestrian does not establish whom the cyclist passes first. Figures 14, 18, and 21 further distinguish working in place, remaining idle, and pausing during an activity. In the street example, some models infer uninterrupted movement from continuous frame timestamps, although the reference identifies an initial pause. Together, these cases show why recognizing the surrounding activity is insufficient to recover the interaction or motion state required by a question.

Acoustic Evidence and Sound Sources. Figures 13, 17, and 20 examine acoustic texture, musical onset, and distinct voices. Visible participants do not establish how many voices are audible, and a visible spray bottle does not establish a spraying sound. Responses sometimes substitute visual expectations for the requested acoustic observation. Acknowledging unavailable audio also differs from confidently answering a speaker-count question using the number of people visible in the scene. Figure 26 extends this analysis to sound-source attribution: our model associates cutting sounds with branch cutting, while another response attributes the sounds to walking through vegetation. These examples emphasize both identifying an audible event and relating it to the corresponding action or object.

Narration and Event Coverage. In Figure 15, **EgoMind** combines the music stand, guitar notes, and steady head position, preserving information distributed across modalities. The displayed baseline instead focuses on looking at sheet music and omits the musical performance. Other examples distinguish cleaning a wooden box from organizing its contents and identify opening a dishwasher beyond entering the kitchen. Figure 25 examines shorter audio-visual segments. The ironing prediction captures the central action, while the workshop prediction describes the setting but omits the reference's grinding event. These contrasting cases illustrate that relevant scene details can enrich a description without ensuring coverage of its main event.

Faithfulness and Question Relevance. Figure 27 tests whether responses address the requested action and reject unsupported claims. the method rejects the proposed stirring action, attributes changing surroundings to camera movement, and describes touchscreen interaction, whereas another response discusses footsteps. Correct central answers can nevertheless contain unsupported details. In Figure 17, **EgoMind** correctly rejects an interruption but adds a continuous background announcement not established by the displayed reference. Similarly, an IMU answer in Figure 21 captures the predominantly stationary state but assigns movement to an interval labeled stationary. These cases separate answer agreement from the faithfulness of supporting explanations.

Temporal Localization. Figures 22–24 distinguish identifying an event's neighborhood from estimating its boundaries. For pump-hose removal, our model predicts 9–13 s against the reference 8–12 s. Its strongest-motion prediction similarly shifts both boundaries by one second, while other examples match the reference intervals. The background-pedestrian example provides such a case, with the prediction recovering both the annotated onset and offset. These results illustrate variation in boundary precision even when the relevant event is identified, motivating interval-overlap evaluation alongside semantic correctness.

Assistive Question Answering. Figure 28 focuses on the user's immediate information needs. The robotic-guide response includes checking for obstructions and requesting movement again, while the price response matches the reference of 13.90 yuan. Such questions require relevant guidance or precise local details, rather than a general scene description. These external examples also broaden the evaluation context: EgoAVU-Bench (Seth et al., 2026) supplies video and audio, whereas EgoBlind (Xiao et al., 2025) supplies vision only. Their predictions therefore illustrate performance under the available inputs in each benchmark.

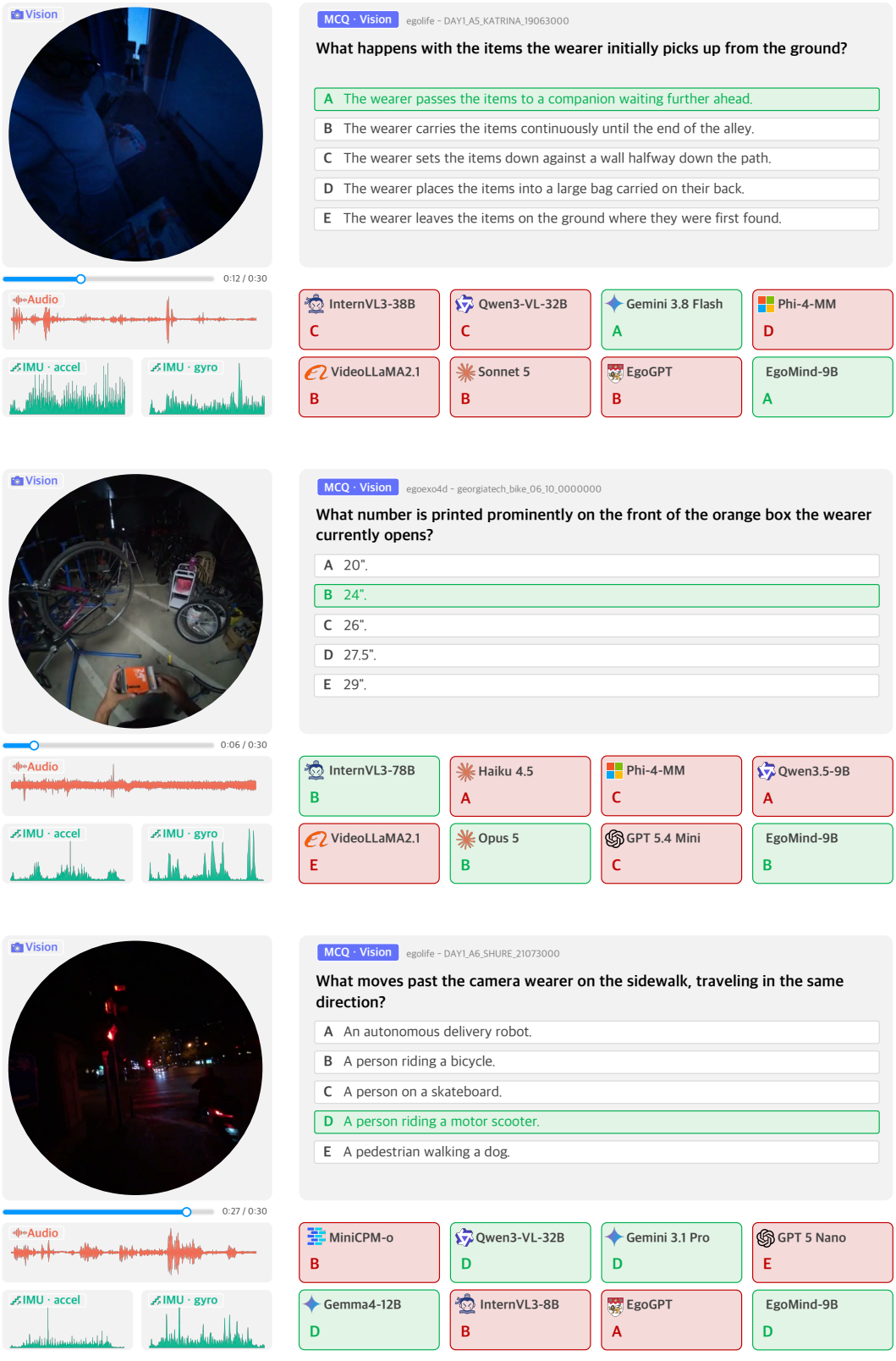


Figure 12: **Qualitative examples and result** of state-of-the-art egocentric scene understanding models for performing on the MCQ-V task from the proposed EgoMMU benchmark.

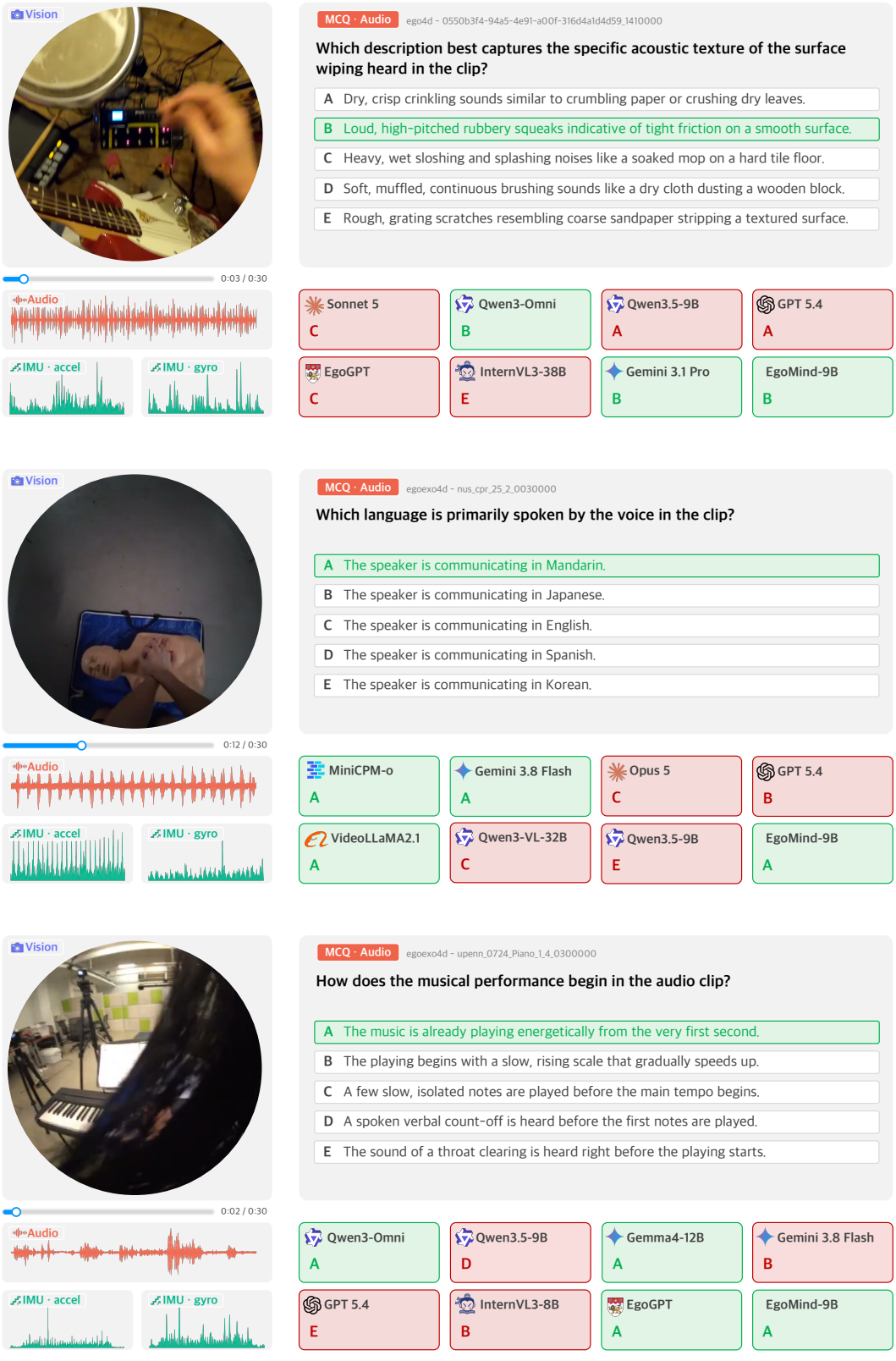


Figure 13: **Qualitative examples and result** of state-of-the-art egocentric scene understanding models for performing on the MCQ-A task from the proposed EgoMMU benchmark.

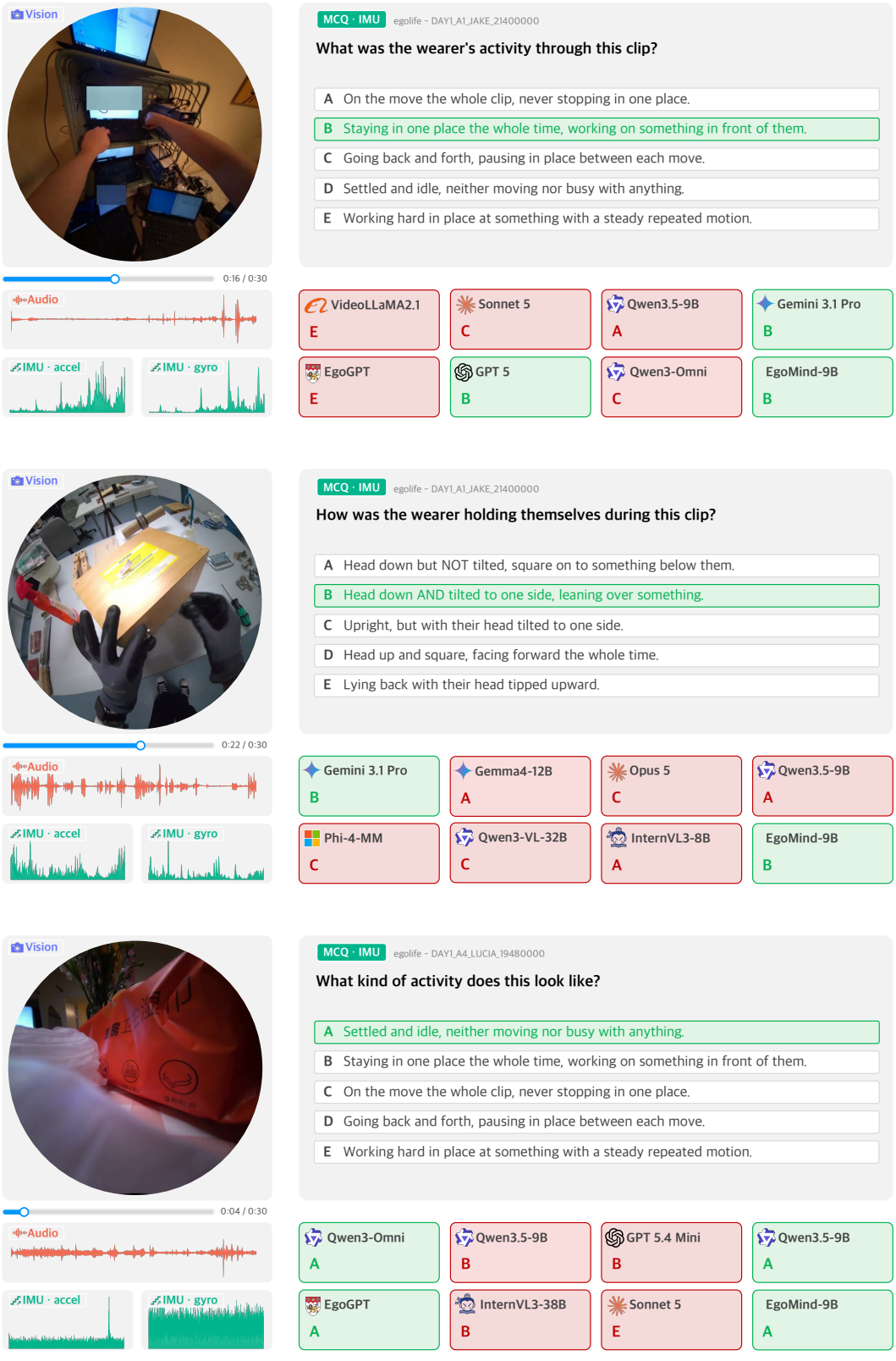


Figure 14: **Qualitative examples and result** of state-of-the-art egocentric scene understanding models for performing on the `MCQ-I` task from the proposed EgoMMU benchmark.



Figure 15: **Qualitative examples and result** of state-of-the-art egocentric scene understanding models for performing on the `Narration` task from the proposed EgoMMU benchmark.

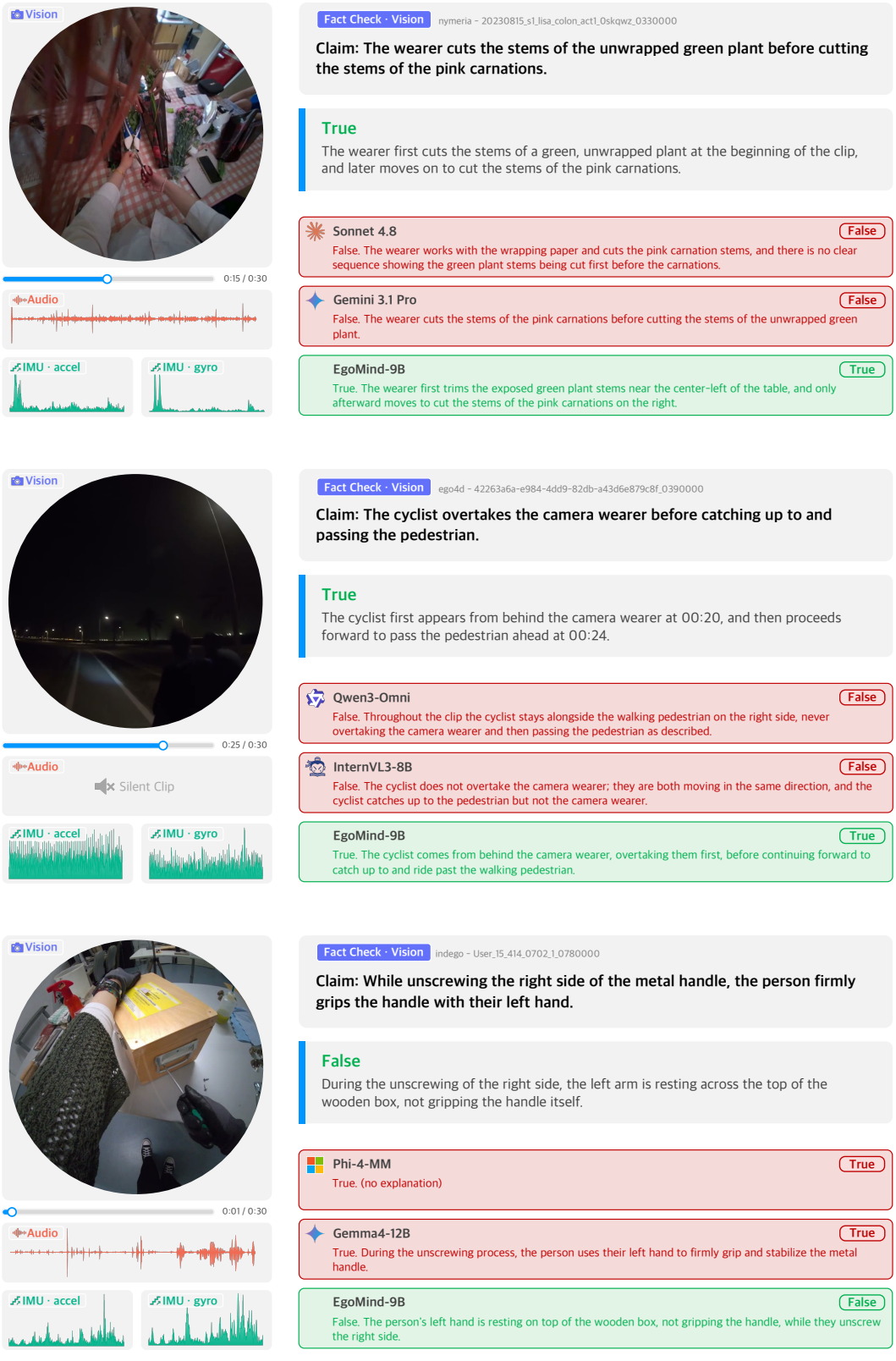


Figure 16: **Qualitative examples and result** of state-of-the-art egocentric scene understanding models for performing on the FC-V task from the proposed EgoMMU benchmark.

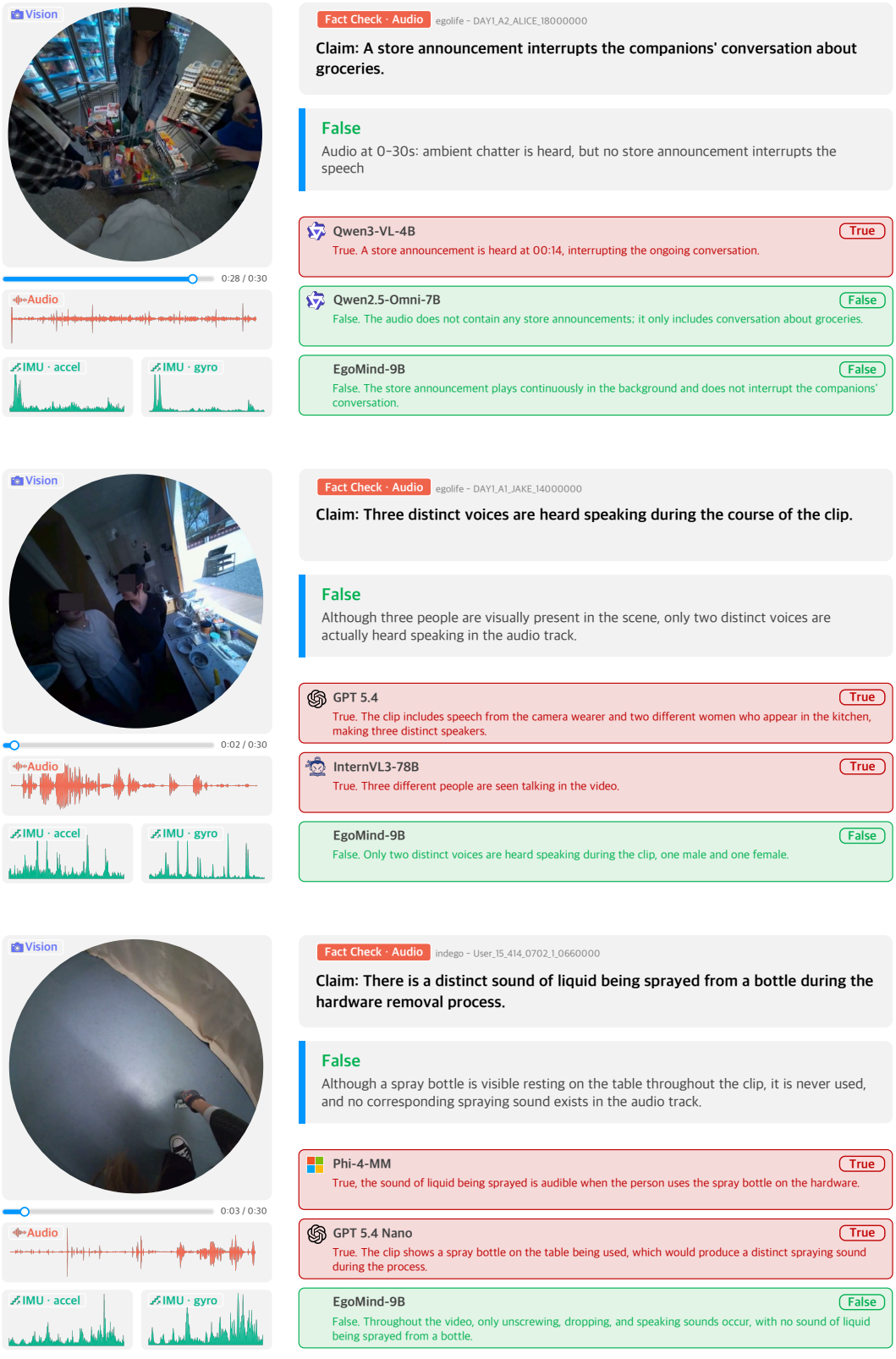


Figure 17: **Qualitative examples and result** of state-of-the-art egocentric scene understanding models for performing on the FC-A task from the proposed EgoMMU benchmark.

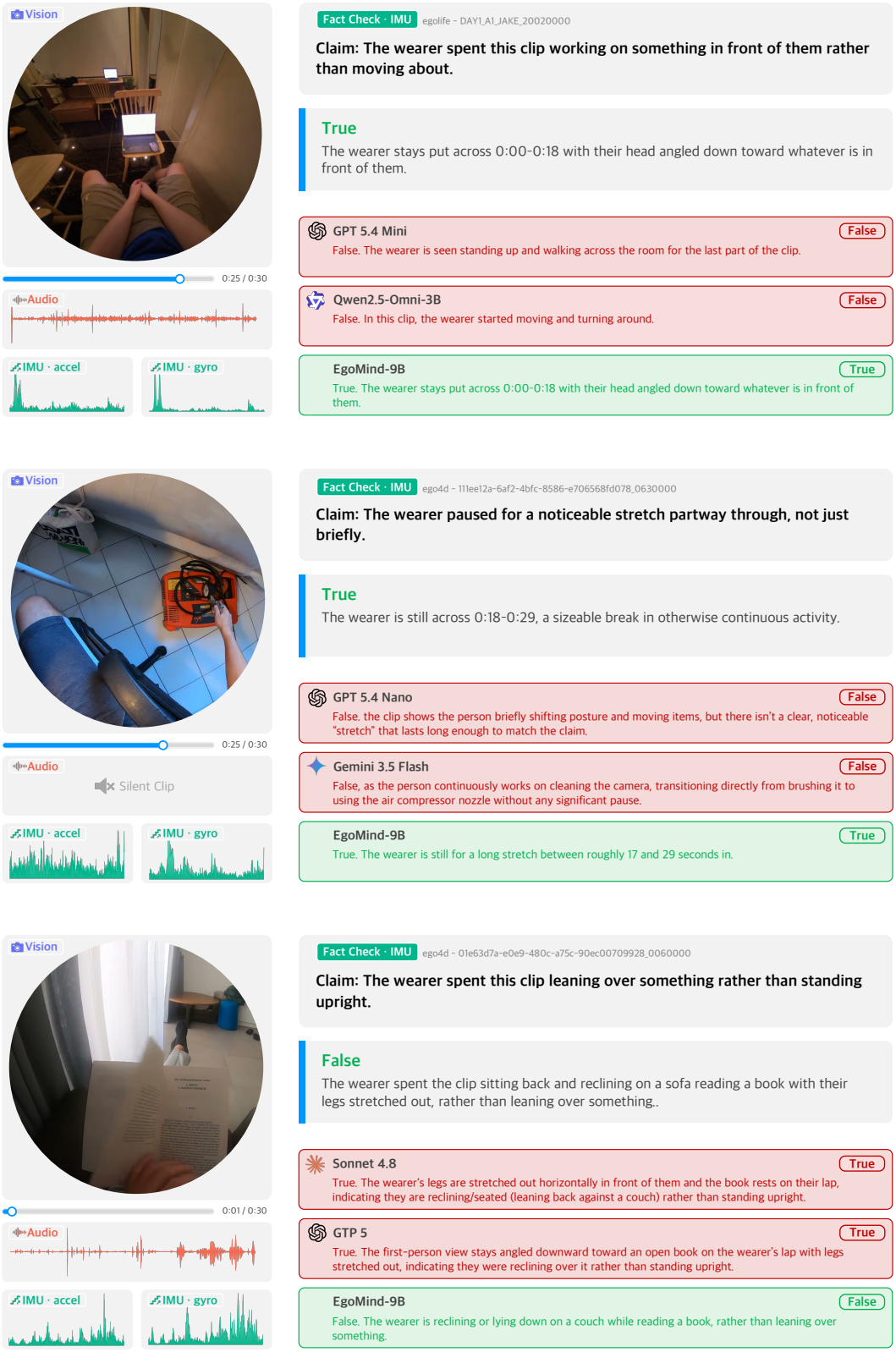


Figure 18: **Qualitative examples and result** of state-of-the-art egocentric scene understanding models for performing on the FC-I task from the proposed EgoMMU benchmark.



Open QA · Vision egolife - DAY1_A2_ALICE_13160000

What activity were the wearer and the two people closest to them engaged in at the table?

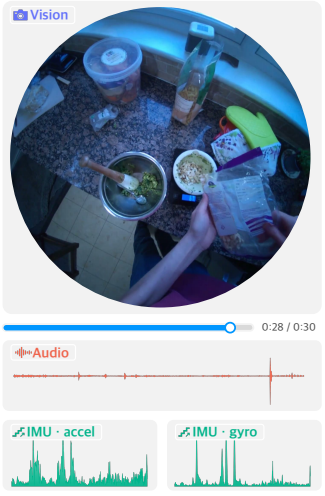
Reference
The wearer and two other people were engaged in assembling a wooden puzzle or craft on a round wooden base.

InternVL3-8B
They were decorating a small round object, possibly a cookie or a piece of cake.

Opus 4.6
They were working on a wood-burning craft project, drawing or tracing designs on a round wooden coaster.

Gemini 3.8 Flash
They were assembling a wooden puzzle together.

EgoMind-9B
Assembling/putting together a wooden jigsaw puzzle on the table.



Open QA · Vision hdepic - P05-20240424-171038_00630

What items did the wearer pour into the small white bowl on the scale, and how did they process them afterwards?

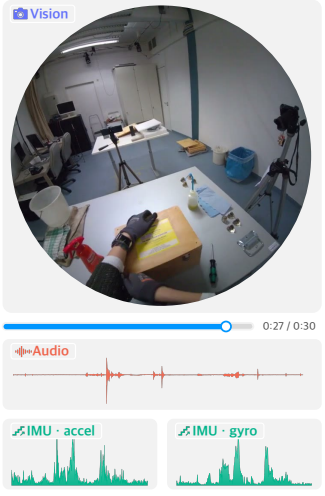
Reference
The wearer poured small nuts or seeds from a plastic bag into the white bowl on the scale, and then used a mortar and pestle to crush them.

Haiku 4.5
The wearer poured cereal into the small white bowl, and afterwards they ate it with a spoon from the bowl.

Qwen2.5-VL-72B
The wearer poured rice from a plastic bag into the small white bowl and then added it to the pot on the stove.

GPT 5.4 Nano
They poured chopped green herbs into the small white bowl and then measured, transferred, and boiled them.

EgoMind-9B
The wearer poured nuts into the small white bowl and then crushed them with a pestle.



Open QA · Vision indego - User_15_414_0702_1_0660000

What incident occurred with one of the metal pieces the wearer was handling, and how did they resolve it?

Reference
One of the metal pieces, a handle, fell from the table to the floor. The wearer then bent down and picked it up.

GPT 5.4 Mini
A metal piece bent while being handled; they fixed it by repairing it before continuing the assembly.

Qwen2.5-Omni-3B
Metal piece fell off table. How did the person in the video handle the metal piece that fell off the table?

Phi-4-MM
Injury.

EgoMind-9B
The wearer dropped a metal handle onto the floor and resolved the issue by picking it back up.

Figure 19: **Qualitative examples and result** of state-of-the-art egocentric scene understanding models for performing on the OQA-V task from the proposed EgoMMU benchmark.

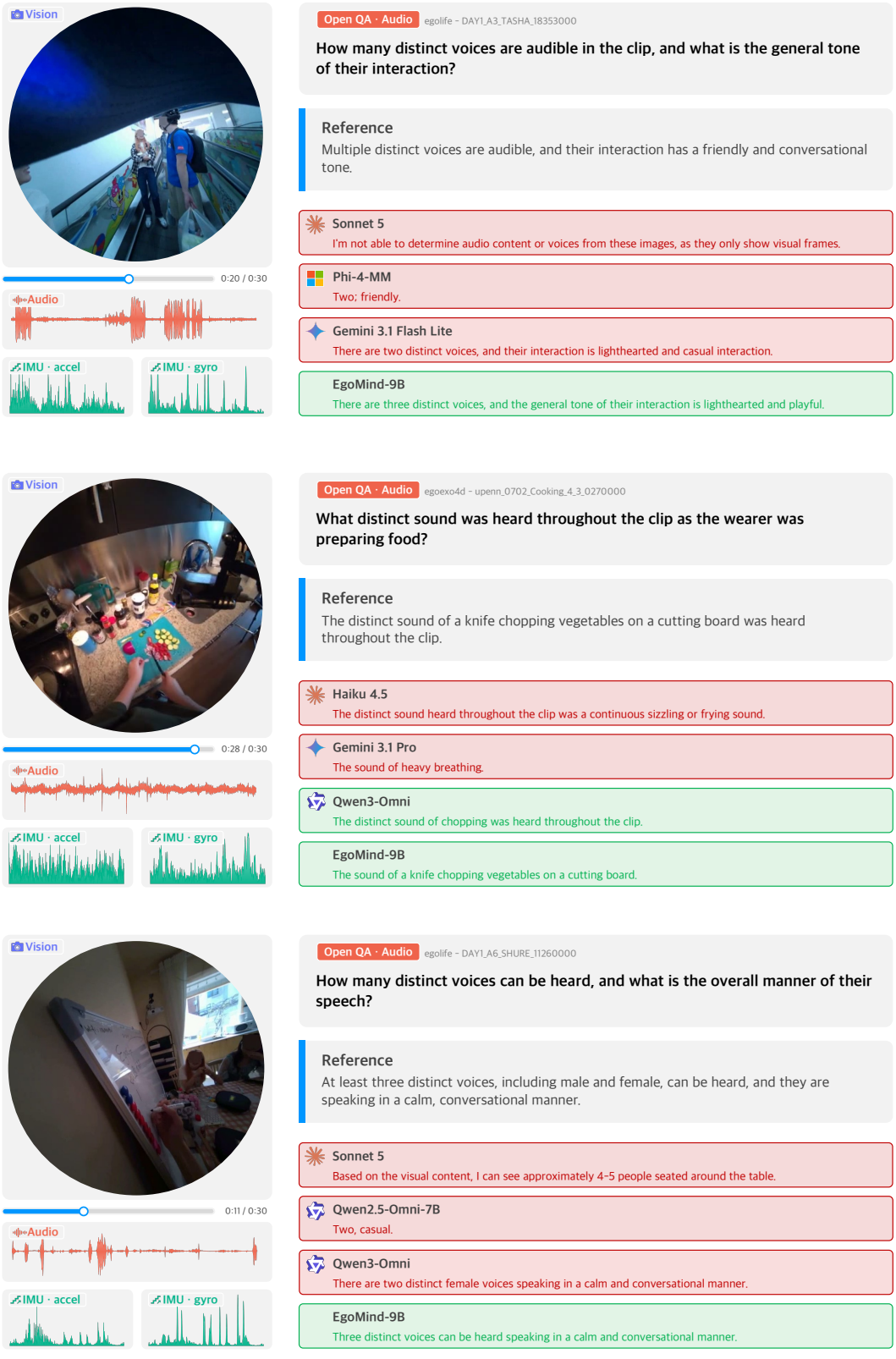


Figure 20: **Qualitative examples and result** of state-of-the-art egocentric scene understanding models for performing on the OQA-A task from the proposed EgoMMU benchmark.



Figure 21: **Qualitative examples and result** of state-of-the-art egocentric scene understanding models for performing on the `OQA-I` task from the proposed EgoMMU benchmark.

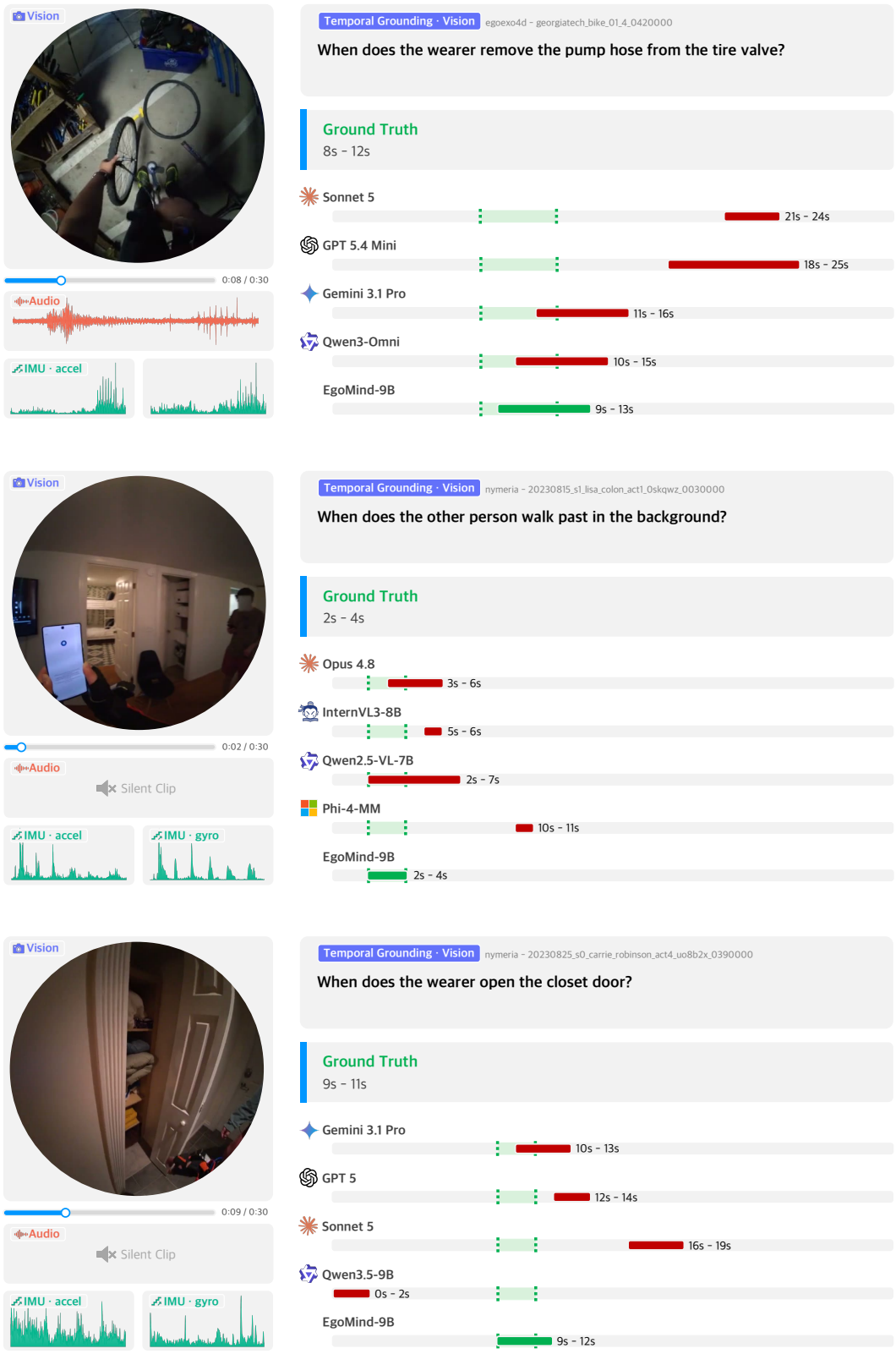


Figure 22: **Qualitative examples and result** of state-of-the-art egocentric scene understanding models for performing on the TG-V task from the proposed EgoMMU benchmark.

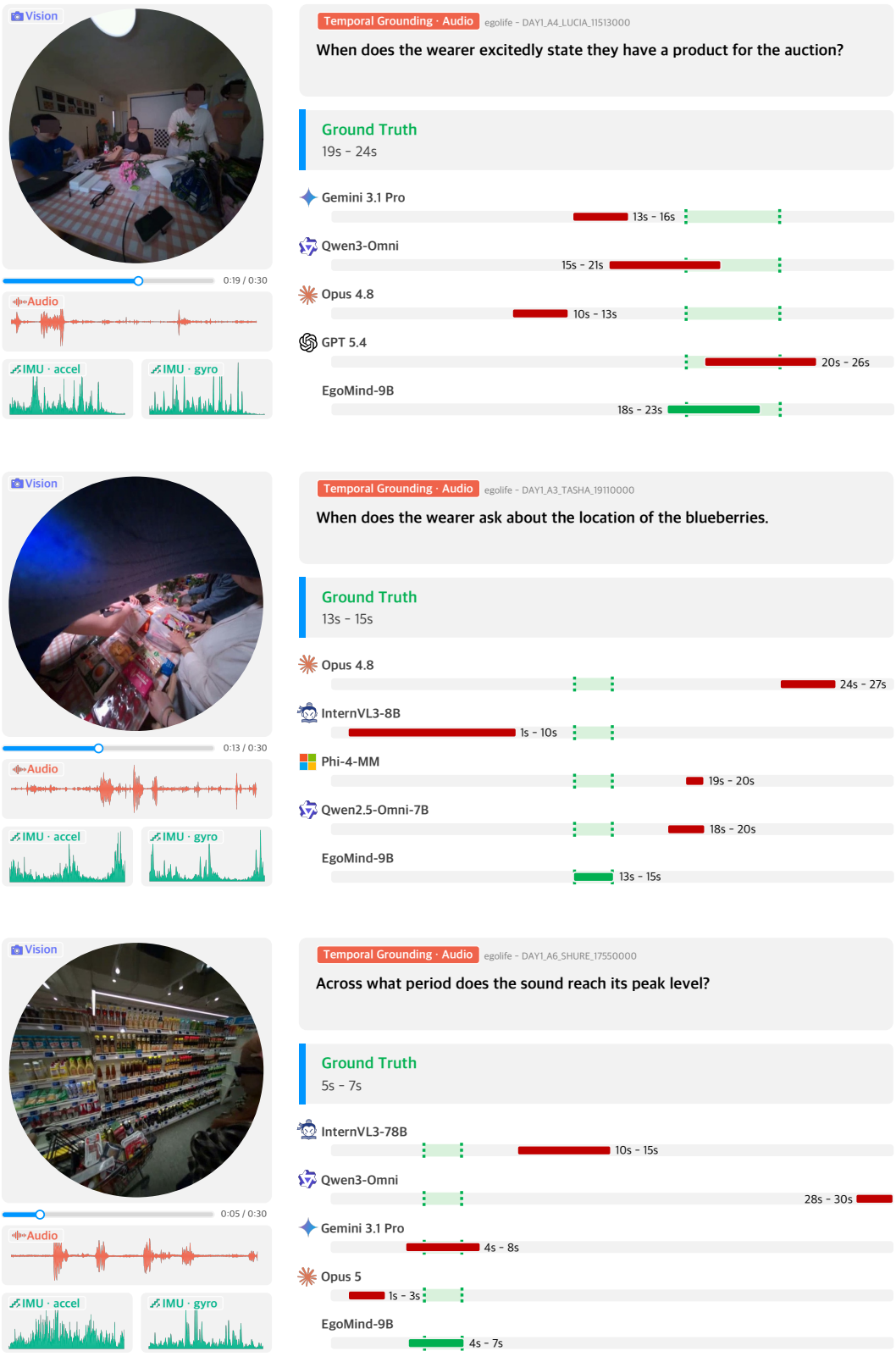


Figure 23: **Qualitative examples and result** of state-of-the-art egocentric scene understanding models for performing on the TG-A task from the proposed EgoMMU benchmark.

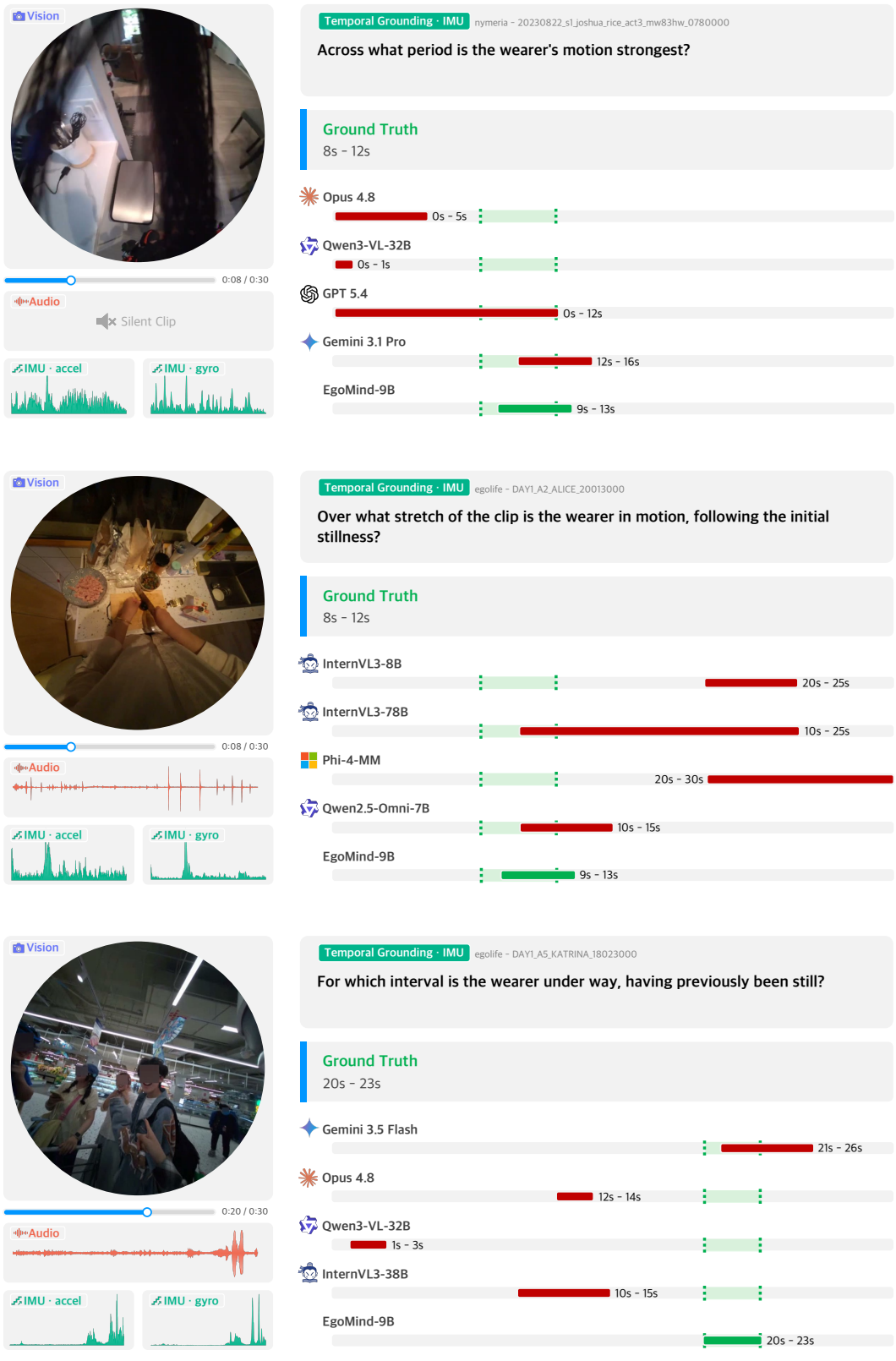
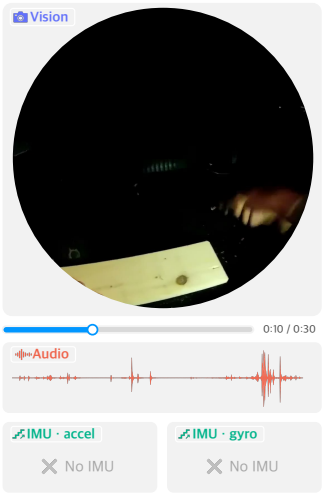


Figure 24: **Qualitative examples and result** of state-of-the-art egocentric scene understanding models for performing on the TG-I task from the proposed EgoMMU benchmark.



AVSN · Audio egoavu-bench - ffd0a9b6-bb01-44e0-b6db-41d6a979d648_00120000_00219937_qa30_00060000

Between 10 seconds and 20 seconds, what information can be gathered from the video's sights and sounds?

Reference

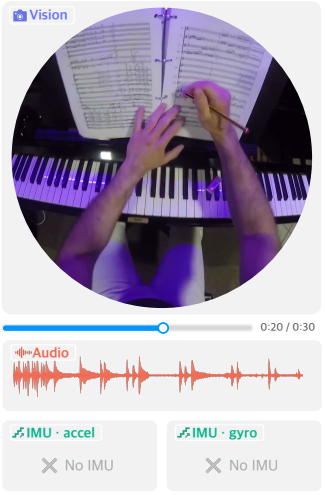
A person used a grinder on a piece of metal, creating a shower of sparks that flew everywhere, in a typical workshop scene with various tools and a mess scattered about, and a welding machine, wires, and a car were also present.

Opus 4.8

I'm unable to determine the exact timing of specific moments in the video, as I can only see a series of still frames without timestamps or audio.

EgoMind-9B

The person walks through a garage and approaches a workbench. Visible on the workbench and surrounding walls are various tools, . . . Heavy breathing and the rustling and footstep sounds of the person walking across the workshop floor can be heard . . .



AVSN · Audio egoavu-bench - fbe0a401-9bf3-4725-bddb-b2c44472cda0_00120000_00454687_qa30_00290000

What are the key sounds between 20 seconds and 30 seconds, and what actions or objects might be responsible?

Reference

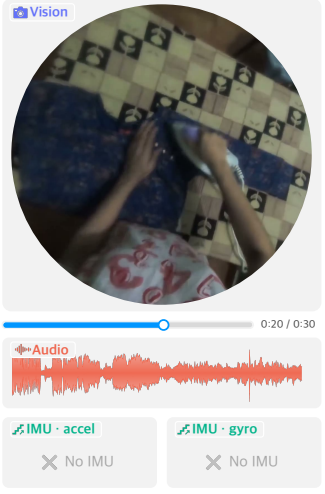
A person played the piano, creating a rhythmic pattern of short, sharp notes, with their hands moving across the keys. The piano, which had a cool blue light around it, was accompanied by music sheets on its stand, and a phone was placed on a stand beside it. The scene was captured from an above angle, providing a clear view of the person's hands and the piano keys.

Qwen3.5-9B

The scene is set in a room with a piano, where a person is playing the piano. The piano has a black and white key layout, and there are sheet music books . . . The person's hands are visible as they play the piano. The room has a purple light, and . . . The audio consists of the sound of the piano being played . . .

EgoMind-9B

A person sat at a piano, playing a piece of music. The person's hands moved across the keys, and the sound of the piano filled the room . . . The person's fingers moved quickly and precisely, and the music was played with skill and precision . . . The person's hands were visible, and the piano was in focus, while the background . . .



AVSN · Audio egoavu-bench - f6512a71-ba0d-4282-97f7-749dedd9a002_00120000_00448625_qa30_00250000

Between 20 seconds and 30 seconds, what are the primary visual elements and sounds present in the scene?

Reference

A person wearing a white dress with red patterns ironed a purple shirt on a table covered with a checkered tablecloth featuring black and white squares. They used an iron with a white cord and a blue light, carefully moving it back and forth over the shirt while . . . A blue bowl was visible on the table, but it was not interacted with during the ironing process.

InternVL3-8B

The primary visual elements are the person, the iron, the blue shirt, and the checkered tablecloth. The sounds are likely the sound of the iron moving over the fabric and the person's movements.

EgoMind-9B

The person was wearing a white dress with red patterns and was holding the iron in their right hand. The person was also holding the shirt with their left hand . . . The person was moving the iron back and forth on the shirt, and the sound of the iron could be heard. The person was also looking at the shirt while ironing it.

Figure 25: **Qualitative examples and result** of state-of-the-art egocentric scene understanding models for performing on the AVSN task from EgoAVU-Bench (Seth et al., 2026).



SSA · Audio egoavu-bench - 2c66293f-b808-4494-bb6d-ea439888a920_00060000_00400062_qa30_00060000

During 10 seconds and 20 seconds, what are the most prominent sounds, and what might be causing them?

Reference

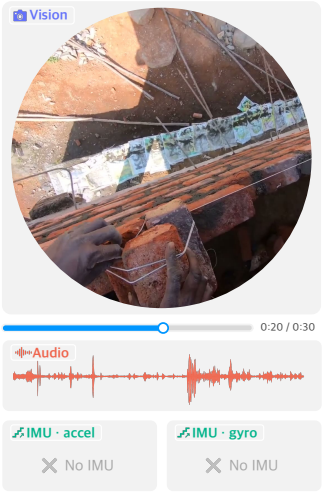
During the segment, three distinct sounds were heard. First, a tapping or thumping sound occurred when the person dropped the cutting board on the counter. Next, a chopping sound was produced when the person chopped onions on the chopping board. Lastly, a cutting sound was caused by the person cutting the head of the eggplant with the knife.

Qwen3-Omni

The most prominent sounds are a series of metallic clinks and scrapes, . . . , followed by a loud, sustained electronic buzz. The clinking sounds are likely caused by . . . The loud, sustained buzz is the sound of the electric stove's heating element being turned on, which is a common sound for . . .

EgoMind-9B

The most prominent sounds are, first, sizzling caused by the liquid egg mixture cooking in the hot rectangular frying pan . . . Second, tapping or scraping sounds that caused by the wooden chopsticks moving and stirring the cooking egg against the surface of the pan . . .



SSA · Audio egoavu-bench - 2efc9b47-0a23-4247-89a6-db46dae6c998_00120000_00457750_qa30_00070000

What are the key sounds between 20 seconds and 30 seconds, and what actions or objects might be responsible?

Reference

During the segment, three distinct sounds were heard. The person stirring cement with a trowel produced a scraping noise. Next, an impact sound occurred when the person applied cement to the wall bricks. Lastly, a whooshing sound was generated as the person spread the cement on the bricks with a spoon.

Opus 4.8

Based on this construction scene showing masonry work at height, here are the key sounds likely occurring: grinding sounds, from the trowel spreading mortar . . . ; knocking of bricks, from bricks being picked up . . . ; metallic clanging, from the rebar/steel rods visible extending outward, . . .

EgoMind-9B

During the segment, a distinct clinking sound was heard as the person . . . clean the surface of the brick. The clinking sound was accompanied by a brief moment of silence, followed by a soft thud and a slight clatter . . . The person then used a tool to clean the surface . . . , and the clinking sound continued as they worked . . .



SSA · Audio egoavu-bench - 3bb64bc3-c2ca-488e-a322-26137c7e4733_00240000_00393000_qa30_00050000

During 20 seconds and 30 seconds, what sounds are audible, and what might be causing them?

Reference

During the segment, one distinct sound was heard. The sound was a cutting or impact sound, which occurred when the person cut a branch with the pruning shears.

Gemini 3.1 Pro

The video features sounds of rustling and snapping. The rustling is likely caused by the person walking on dry grass and twigs.

EgoMind-9B

During the segment, two distinct sounds were heard. First, a rustling sound occurred when the person pulled the branches. Next, a cutting sound was produced when the person cut the branches.

Figure 26: **Qualitative examples and result** of state-of-the-art egocentric scene understanding models for performing on the SSA task from EgoAVU-Bench (Seth et al., 2026).



AVH · Audio egoavu-bench - f36a1cef-a868-4cc7-8424-a166d9752c1f_00060000_00120125_qa30_00030125

Did the person stir the food in a bowl using a wooden spoon between 10 and 30 seconds?

Reference

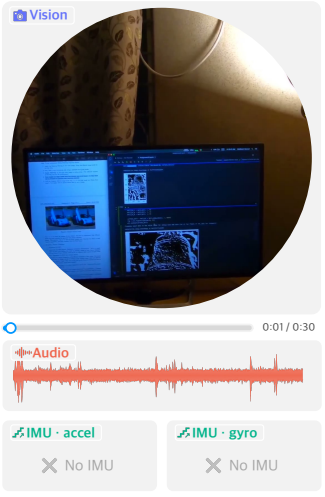
No, there is no visible stirring action with a wooden spoon — the person uses other utensils to cut and arrange ingredients, but stirring is not observed.

EgoGPT

Yes. (no explanation)

EgoMind-9B

No, the person did not stir the food in a bowl using a wooden spoon between 10 and 30 seconds. Instead, they used a knife to cut the food.



AVH · Audio egoavu-bench - e922e85b-06f2-40aa-aa08-181cd8006e0a_00120000_00180125_qa30_00030125

What action does the person perform that changes the visible surroundings between 0 and 30 seconds?

Reference

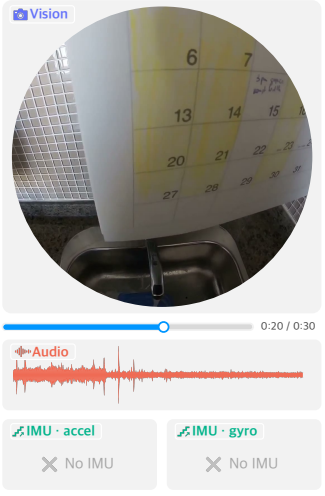
The viewpoint gradually expands, suggesting the person is slowly panning or adjusting their head position to reveal more of the room, including the light's reflection and hanging fabric.

InternVL3-8B

The person moves the laptop.

EgoMind-9B

The person changes the visible surroundings by moving the camera down and to the right, revealing the desk, keyboard, mouse, and laptop screen, before returning the camera to its original position.



AVH · Audio egoavu-bench - d9b4c3ac-7b2c-4454-a84f-c216a9129909_00060000_00120125_qa30_00030125

What action is the person performing with their device between 20 and 30 seconds?

Reference

The person is holding the device, looking at its screen, tapping it repeatedly, and then showing the screen to someone, likely sharing an image or message.

Gemini 3.7 Flash

The repetitive tapping sound heard throughout the video is caused by the person's footsteps (hard-soled shoes, heels, or slippers) tapping against the hard floor as they walk around the room.

EgoMind-9B

The person is interacting with a tablet device, holding it with both hands and using their right hand to touch the screen.

Figure 27: **Qualitative examples and result** of state-of-the-art egocentric scene understanding models for performing on the AVH task from EgoAVU-Bench (Seth et al., 2026).

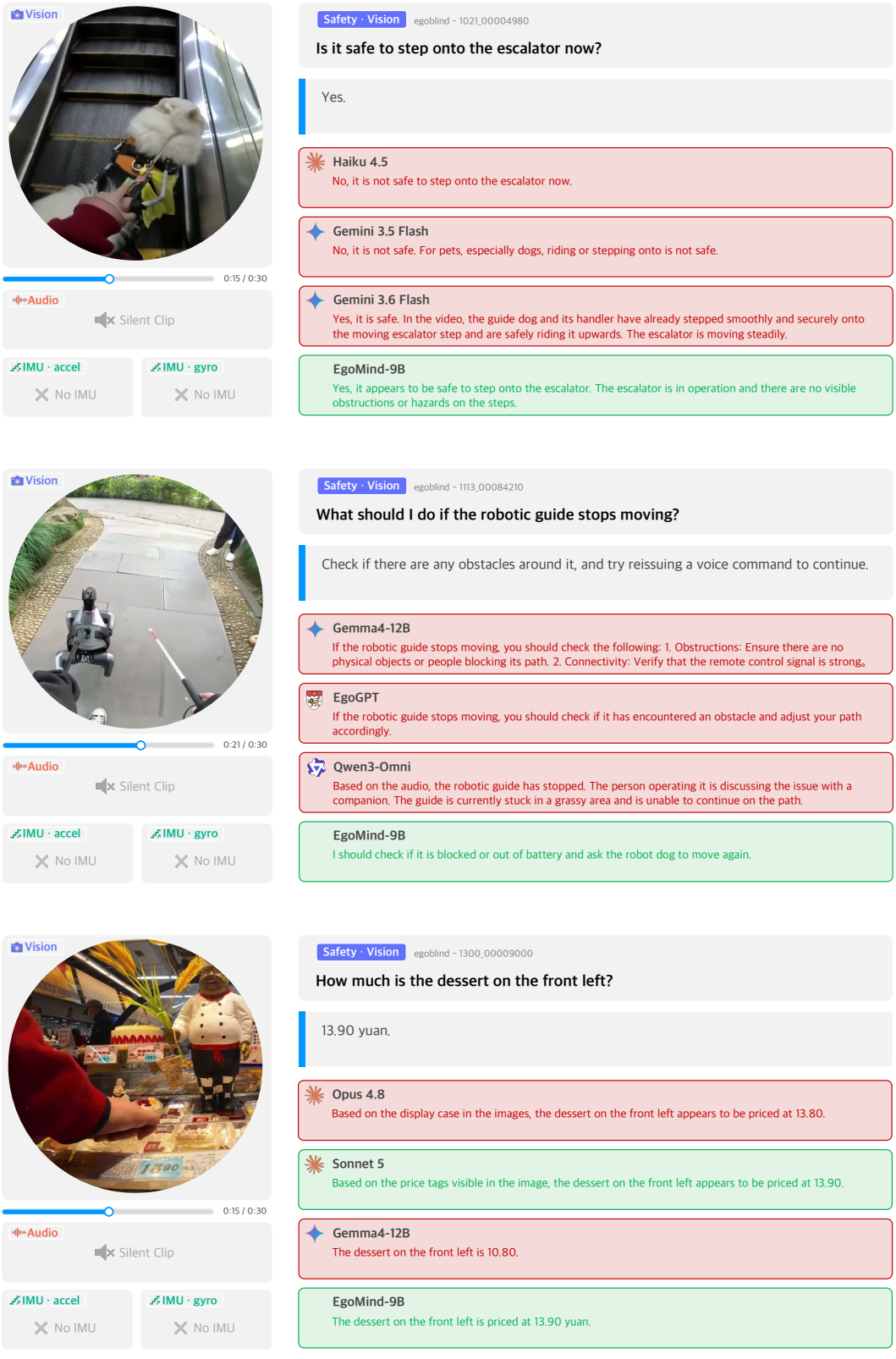


Figure 28: **Qualitative examples and result** of state-of-the-art egocentric scene understanding models for performing on the `Safety` task from EgoBlind (Xiao et al., 2025).

F PROMPTS

A prompt is part of the definition of whatever it produced, so the ones this work depends on are reproduced in full rather than described. They fall into four families, following the order in which the pipeline uses them: prompts that caption each modality on its own from its raw signal, prompts that generate training labels from raw recordings, prompts that author benchmark items from those labels together with the annotations a source ships, and prompts that grade a model's answer against a reference. Braced names are placeholders substituted per item.

F.1 SINGLE-MODALITY CAPTIONING

Before any label or benchmark item is authored, each modality is captioned on its own from its raw signal: video from the frames, audio from the waveform, and IMU from a rule-based feature pass over the two head-mounted sensors that this prompt turns into a caption. The audio and IMU captions are consumed as structured text by the fusion step in the next subsection; the video caption is not – the fusion prompt sees the raw video directly, so the prompt below instead serves the vision-only training data on its own.

F.1.1 VISION

System instruction.

```
You are an expert at describing egocentric (first-person) video from a head-mounted camera. Produce a dense free-form description and structured annotations in JSON format.

Rules for the caption:
- Start directly with actions; never reference "the video" or "the clip."
- Write in third person ("The wearer...", "The person...").
- Include temporal markers ("Initially...", "Around 10 seconds in...").
- Five to eight sentences, with rich temporal structure.
- Describe only what is visually observable -- no sounds, no speculation.

Rules for the structured fields:
- Timestamps are approximate integers in seconds; a tolerance of a few seconds is acceptable.
- Only report what you confidently observe; omit a field entirely if nothing qualifies.
- Describe an object's location relative to the SCENE ("on the table," "in the wearer's right hand"), never relative to the camera view ("left," "right," "center"). If an object moves, give it separate entries for each location and time range.
- Track the left and right hand separately, and only when clearly visible.
- Describe the wearer's own physical movement (head turns, walking, leaning) as ego-motion.
- Assess overall visual clarity and note any dark, blurry, or occluded periods.

Output valid JSON only. No markdown code fences, no explanation.
```

User prompt.

```
Watch this {duration}-second egocentric video clip carefully. Produce a JSON response:

{
  "caption": "dense description, 5-8 sentences, third person, temporal, starting directly with actions",
  "objects": [
    {"label": "short name", "time_start": 0, "time_end": 30, "location": "scene-relative location"}
  ],
  "actions": [
    {"label": "verb phrase", "time_start": 0, "time_end": 10}
  ],
  "hands": [
    {"hand": "left or right", "state": "what the hand is doing", "time_start": 0, "time_end": 10}
  ],
  "ego_motion": [
    {"type": "stationary/turns_left/turns_right/looks_up/looks_down/walks_forward/leans_forward", "time_start": 0, "time_end": 10}
  ],
  "scene": {
```

```
    "setting": "indoor/outdoor and type",
    "lighting": "bright/dim/dark",
    "notable_features": ["feature 1", "feature 2"]
  },
  "visual_quality": {
    "overall": "clear/dim/dark/blurry/partially_occluded",
    "issues": ["any quality issue with its timing, e.g. 'dark from 0-5s'"]
  }
}

Requirements:
- Objects: focus on distinctive or interacted-with items; a repositioned object gets a new entry rather than an updated one.
- Hands: omit entirely if no hand is in frame.
- Use an empty array where nothing qualifies rather than guessing.

Respond only with valid JSON.
```

F.1.2 AUDIO

System instruction.

```
You are an expert at describing audio from egocentric (first-person) recordings captured by a head-mounted wearable device. Produce a dense free-form description and structured annotations in JSON format.

Rules for the caption:
- Start directly with what you hear; never begin with a preamble like "The audio features..."
- Include temporal markers ("Initially...", "Around 5 seconds in...").
- Four to six sentences.
- Describe only what you hear -- do not speculate about visual content.

Rules for the structured fields:
- Timestamps are approximate integers in seconds; a tolerance of a few seconds is acceptable.
- Only report what you confidently hear; use an empty array where nothing qualifies.
- For speech, give a short excerpt of the key content rather than a full transcript, with a speaker label, language, and tone.
- Classify each non-speech sound as action (a physical interaction), ambient (a persistent background sound), or music.
- Non-verbal vocalizations (laughter, breathing, coughing, sighing) are reported separately from speech.

Output valid JSON only. No markdown code fences, no explanation.
```

User prompt.

```
Listen to this {duration}-second audio clip from a wearable device carefully. Produce a JSON response:

{
  "caption": "dense description, 4-6 sentences, temporal, starting directly with what is heard",
  "speech": [
    {"text": "key content excerpt", "speaker": "speaker_1 or wearer", "language": "zh/en/ja/etc", "tone": "casual/excited/instructing/questioning/whispering/shouting", "time_start": 0, "time_end": 5}
  ],
  "sounds": [
    {"label": "short description", "type": "action or ambient or music", "time_start": 0, "time_end": 30}
  ],
  "non_verbal": [
    {"label": "laughter/breathing/coughing/sighing/humming/grunting", "speaker": "speaker_1 or wearer", "time_start": 0, "time_end": 5}
  ],
  "audio_scene": {
    "environment": "indoor/outdoor and type",
    "noise_level": "quiet/moderate/noisy",
    "n_speakers": 0,
```

```
    "primary_language": "zh/en/ja/none"
  }
}

Requirements:
- If there is no speech at all, use an empty speech array and set n_speakers to 0 and primary_language to "none."

Respond only with valid JSON.
```

F.1.3 IMU

System instruction.

```
You describe physical motion from wearable IMU sensor data mounted on the head, using structured motion analysis from both the left and right sensors. Produce a qualitative caption and structured annotations in JSON format.

Rules for the caption:
- Write in first person ("I was walking...", "I remained still...").
- Two to four sentences, purely qualitative -- no raw numbers (degrees, Hz, m/s^2).
- Include temporal markers for transitions.
- Do not speculate about objects, people, scenes, or activities.
- Keep it brief if the wearer is mostly stationary.

Rules for the structured fields:
- Use standard categories only: locomotion (stationary/light_motion/walking/running/rhythmic_motion/vehicle), intensity (still/light/moderate/vigorous), head pitch (looking_up/looking_ahead/looking_down), head dynamics (still_head/slow_scanning/active_scanning/quick_turns).
- Report significant state changes as transitions, copied directly from the sensor data.
- Compare the left and right sensor patterns for bilateral asymmetry.
- Timestamps come from the sensor data directly and are objective, to about a second.

Output valid JSON only. No markdown code fences, no explanation.
```

User prompt.

```
Two head-mounted IMUs (accelerometer and gyroscope) recorded this {duration}-second clip:

=== LEFT IMU ANALYSIS ===

{tags_text}

=== RIGHT IMU ANALYSIS ===

{right_tags_text}

Produce a JSON response:

{
  "caption": "qualitative description, 2-4 sentences, first person, no raw numbers",
  "motion_segments": [
    {"locomotion": "stationary/light_motion/walking/running/rhythmic_motion/vehicle", "intensity": "still/light/moderate/vigorous", "time_start": 0, "time_end": 10}
  ],
  "head_orientation": [
    {"pitch": "looking_up/looking_ahead/looking_down", "dynamics": "still_head/slow_scanning/active_scanning/quick_turns", "time_start": 0, "time_end": 30}
  ],
  "transitions": [
    {"type": "locomotion/intensity/posture", "from": "previous state", "to": "new state", "time": 8}
  ],
  "bilateral": {
    "symmetry": "symmetric/asymmetric/left_dominant/right_dominant/single_sensor",
    "left_description": "brief description of the left sensor's pattern",
    "right_description": "brief description of the right sensor's pattern, or that no right sensor exists"
```

```
  },
  "summary": {
    "dominant_locomotion": "category",
    "dominant_intensity": "category",
    "dominant_head_pitch": "category",
    "dominant_head_dynamics": "category",
    "movement_quality": "smooth/irregular/rhythmic/abrupt"
  }
}

Requirements:
- motion_segments: merge short phases into coherent segments rather than reporting every short window.
- head_orientation: report changes only -- one entry is enough if orientation never changes.
- If the right sensor's input is exactly "(not available)," the device has only one IMU (as on some headsets): set symmetry to "single_sensor" and right_description accordingly, rather than calling it asymmetric.

Respond only with valid JSON.
```

F.2 Multi-Modality Captioning

The audio and IMU captions above are fused with the raw video into a single narration used as a training label. One rule shapes it: because the audio pathway this caption trains is a sound-event encoder rather than a speech recognizer, the caption may state that speech occurred and in what tone, but never what was said, since a target stating speech content teaches the model to hallucinate words it cannot hear. Grounding and anonymization are otherwise as below, and source-native annotations are released with the data rather than reproduced here. A companion prompt, not shown, appends instruction-tuning items to the same call for the multi-task pool under the same rule and schema.

F.2.1 System Instruction

```
=== SOUND-EVENT AUDIO RULE ===

This caption is a training target for a model whose audio encoder is a sound-event encoder, not a speech recognizer -- it cannot transcribe words. A target that states what was said is unlearnable from that pathway and only teaches the model to hallucinate speech it cannot hear. Describe audio the way the encoder can actually represent it: as sound.

Describe: the presence and type of a sound ("voices talking", "a single voice", "laughter", "a drill", "running water", "music playing", "footsteps"); its acoustic and paralinguistic properties (how many voices, near or far, loud or quiet, calm or excited or tense, indoor or outdoor, overlapping or sequential); and the identity of a spoken language as a sound property ("someone speaking Mandarin") -- the language is audible even though the content is not.

Never state: what anyone said, verbatim or paraphrased; the topic of a conversation; a reporting verb with its content ("explained that...", "asked whether..."); or any fact knowable only by understanding the words rather than hearing the sound.

Apply the muted-transcript test to every audio phrase: could you still write it if you heard the clip with the words bleeped out -- voices, tone, who, how many, the language, but not the words? If it survives, keep it; if it needs the words, delete it or replace it with the sound itself. ("Two people talk in a relaxed tone over background music" survives; "a companion explains that the playlist needs more upbeat songs" does not.)

A source-native ground-truth annotation, where one exists, may be used to know that speech occurred and in what language or tone, and to ground actions and events -- never to supply what was said.

=== GROUNDING, ANONYMIZATION, AND DIARY SUITABILITY ===

Every sentence must be derivable from this clip's video, audio, and IMU alone, with no fact that depends on an annotation but is not itself observable in the clip; anchor to observable events rather than absolute timestamps.

```

```
Never emit a proper name or an annotation-only person label. Replace each person with a stable, clip-grounded descriptor built from what is visible, reused consistently for that person; the wearer is always first person ("I").

Write the narrative so its first sentence stands alone as a clean, diary-style summary of the wearer's action and location, since it also serves as a memory-system entry; populate the structured observations richly, and note any scene or location change explicitly.

=== TASK ===

You are producing a rich, accurate first-person egocentric scene description by integrating video (which you SEE), audio (which you HEAR, muxed into the same file), and structured IMU motion data.

Produce only structured observations and a natural first-person narrative -- no questions, multiple choice, fact claims, or other benchmark items; this is a narration target only.

Core rules:
- Video is the primary visual reference; ground every visual claim in what is actually seen. Where a source-native annotation exists, use it to keep events and actions accurate, but never copy more than a handful of consecutive words from it, and never copy a name.
- Audio: describe only sound events and acoustic or paralinguistic properties, per the sound-event rule above -- never transcribe or summarize speech content.
- IMU: trust it for physical state -- presence, cadence, a change in activity level over time, head orientation. A single head-mounted sensor measures head motion only: no left/right or hand/arm laterality, no spatial localization, no subjective intensity language.
- Write the narrative in first person ("I walked...", "I heard voices nearby...").
- Output valid JSON matching the schema exactly.
```

F.2.2 User Prompt

```
I'm showing you a {duration}-second egocentric video clip from a head-mounted camera. The file has audio muxed in -- listen to it natively, but describe it only as sound events, per the sound-event rule. Structured annotations follow, to ground actions, language, and tone only -- never speech content or names.

=== AUDIO ANNOTATION (language/tone/sound presence only) ===

Speech utterances (with tone): {speech_json}
Sound events: {sounds_json}
Non-verbal vocalizations: {non_verbal_json}
Audio scene: {audio_scene_json}

=== IMU MOTION DATA ===

Motion segments: {motion_segments_json}
Head orientation: {head_orientation_json}
Transitions: {transitions_json}
Bilateral (left vs. right): {bilateral_json}
Sensor stats: {sensor_stats_json}

=== NATURAL-LANGUAGE SUMMARIES ===

Audio: {audio_caption}
IMU: {imu_caption}

{ground_truth}

Produce a JSON response matching this schema:

{
  "observations": {
    "scene": "setting and key objects; people by grounded descriptor, never by name",
    "actions": ["action 1 with timing", "action 2 with timing"],
    "audio_events": ["sound-event phrases only, e.g. 'two voices talking, relaxed tone', 'background music', 'clattering dishes' -- no speech content or topic"],
    "speech_sound": {"voices_present": true, "n_speakers": 0, "language": "language(s) heard, or null", "tone": "relaxed/excited/tense/instructing/questioning, or null"},
    "sounds": ["sound 1", "sound 2"],
```

```
    "motion_state": "locomotion, activity level over time, and head orientation -- no intensity language, no laterality"
  },
  "narrative": "first-person, 3-5 sentences, integrating video actions, sound-event audio, and IMU motion; first sentence is a diary-style action-and-location summary; grounded descriptors, never names"
}

Apply the muted-transcript test to every audio mention in the output: no speech content, no quoted speech, no conversation topic. No proper names anywhere, and no long verbatim span copied from any annotation.

Respond only with valid JSON. No markdown code fences.
```

F.3 Benchmark Authoring

Every benchmark item is produced by one prompt call per clip, one call per egocentric source. The call fuses the raw clip with its audio track intact, the audio and IMU analyses from the captioning prompts above, and, where the source ships one, a human-verified annotation native to that source. The prompt below is the design common to every source; the annotations themselves are source-specific and released with the benchmark data. The one source without an audio track omits the audio section and every audio-dependent item, rather than substituting an empty one.

F.3.1 System Instruction

```
=== GOVERNING CONTRACT (overrides anything below that conflicts) ===

Every item must satisfy four grounding requirements:
G1. The answer or claim must be derivable from this clip's video, audio, and IMU alone -- no fact that depends on an annotation but is not itself seen, heard, or measured in the clip.
G2. A human watching only this clip, with no external labels or names, must be able to answer it; otherwise drop the item.
G3. Ecological validity: phrase every question as something a real wearer would plausibly ask an assistant, not as trivia about a hyper-specific detail no one would ask.
G4. No absolute-timestamp trivia in question text ("between 12 and 14 seconds"); anchor to an observable event instead. Integer-second windows remain valid as temporal-grounding answers.
Anonymization: never emit a proper name or an annotation-only person label. The wearer is always first person ("I"); everyone else is referred to by a stable, clip-grounded descriptor built from what is visible (clothing, role, position) and reused consistently for that person across the clip.
Temporal-light: emit at most one temporal-grounding item per clip (zero if none is clean); do not pad.

=== MODALITY-QUALITY CONTRACT (overrides anything below that conflicts) ===

Canonical modality. Every item carries two fields: `modality` (exactly one of video / audio / imu, or null for a genuinely multi-modal item) and `modality_set` (the list of modalities actually needed to answer). These are authoritative.

Answer-evidence, not anchor. `modality` names what is needed to KNOW the answer, not what is needed to LOCATE the moment. Ask: is this modality required to answer, or only to find when it happens? Only the former counts. A "video" item whose answer secretly needs audio (e.g. "did the wearer sigh after finishing?") is a labeling error -- retag it or move it to a multi-modal open-QA item.

Binary balance. Across a clip's true/false and yes/no items, keep the label split near 50/50 -- never let either be guessable from the distribution.

IMU realism. A single head-mounted IMU measures head motion only. No IMU item may ask about left/right or hand/arm/leg laterality, spatial location ("at the sink"), or visual conditions ("in the dark"). IMU question and claim text may not use subjective intensity language ("vigorous," "gentle," "rapid," and similar) -- these have no objective IMU criterion. The only allowed framings are presence (moving vs. still), cadence or rhythm, a qualitative change in activity level over time, head orientation, and a count of motion segments or transitions.

```

```
No timestamp anchors. No closed-form question text (multiple choice, fact claim, or the temporal-grounding QUESTION itself) may contain explicit seconds or the word "timestamp." The temporal-grounding ANSWER is the one place integer start/end seconds belong.

Answer normalization. Binary answers are exactly "True"/"False" or "Yes"/"No," with no trailing punctuation or extra words.

Fact-claim dimension purity. Each fact claim is tied to exactly one dimension -- visual, audio, motion, or IMU-specific -- and its `modality` field must match. The claim must be verifiable from that one modality alone: a visual claim covers only what is seen, an audio claim only what is heard, a motion claim only IMU-derived motion. A claim may assert only what the modality directly shows, never the purpose, cause, or result behind it (a whistle blowing is audio; that it signals the start of a climb is not).

No subjective, meta, or identity questions. Ban predicates with no objective criterion ("carefully," "neatly," "properly"). Ban meta-phrasing that addresses the model rather than the content ("Can you determine...," "without seeing the video"). Ban named-identity or world-knowledge questions answerable only from outside the clip.

=== TASK ===

You are producing rich, accurate egocentric scene descriptions and evaluation items from multiple information sources at once. You SEE a video clip directly and HEAR its audio track, muxed into the same file. You also receive structured audio annotations (speech with tone, discrete sound events, non-verbal vocalizations), structured IMU motion data (including a left-vs-right bilateral comparison where the sensor supports it), and natural-language summaries of both.

Integrate all of this into structured observations, a first-person narrative, and a set of benchmark evaluation items.

Core rules:
- The video is the primary visual reference; ground every visual claim in what is actually seen.
- The muxed audio track is the primary auditory reference; use it for paralinguistic features (intonation, certainty, urgency), ambient-sound quality, and anything only audio can disambiguate.
- The structured audio annotations carry ground-truth verbatim text, per-utterance timestamps, and speaker turns.
- The IMU data is objective and precisely timed; trust it for physical state, and use the bilateral field for left/right asymmetry where the sensor is bilateral.
- Cross-validate: align audio events, visual actions, and IMU transitions that share a timestamp.
- Where sources disagree: video is authoritative for scene and objects, structured audio for verbatim speech and speaker attribution, raw audio for paralinguistic and ambient cues, IMU for physical motion.
- Write the narrative in first person ("I walked...", "I heard...").
- Be temporally specific using the timestamps carried in the structured inputs.
- Output valid JSON matching the schema exactly.

Where this source ships a native ground-truth annotation, treat it as authoritative for the facts it directly covers, and never let an identifying detail from it -- a proper name, a subject code -- leak into a question, option, claim, or the narrative; replace it with a grounded descriptor as above. Where no such annotation exists, or a fact falls outside what it covers, ground the claim on the observed video, audio, and IMU alone.

Hardening rules (apply to every clip):
- At least one wrong option in every multiple-choice item must reference a real entity from this clip -- a real timestamp, a real spoken word, or a real observed object, action, or sound.
- True fact claims should prefer details that are visible, audible, or measurable in the clip but not stated in any annotation; false fact claims should prefer plausible-but-wrong details that contradict the observed video, audio, or IMU.
- A cross-modal question must reference timestamps from two or more modalities.
- No multiple-choice option may contain a run of five or more consecutive words copied verbatim from any annotation.

Timestamp formatting: every user-facing timestamp in question text is a whole second ("at 14s," never "at 13.6s"); all internal time fields are integers.

Hard constraints: no span of thirty or more consecutive tokens copied verbatim from any annotation into the narrative or an item; every fact claim must be supported by at least one of video, raw audio, an audio annotation, IMU data, or this source's ground-truth annotation, as available.

```

```
Benchmark rules:
- Multiple choice (video, audio, IMU): five options, exactly one correct, position chosen independently at random for each item, spread across at least three different indices within the clip.
- IMU multiple choice: one to two items when the motion data shows three or more distinct segments or a bilateral asymmetry; categorical answers only, no numeric values.
- Audio multiple choice: roughly two-fifths of clips draw the item from a harder reasoning mode rather than plain transcription recall -- for example resolving which of several speakers said something, reconciling audio against video, or reading paralinguistic cues -- with wrong options that are not eliminable by transcription alone.
- Temporal grounding: each event spans at most fourteen seconds; vary the phrasing across items.
- Open-ended QA: four items -- one each requiring video+audio, video+IMU, and audio+IMU, plus one IMU-anchored item where IMU is the primary evidence and video or audio only secondary context.
- Fact claims: eleven items total -- three true/false pairs across the visual, audio, and motion dimensions; one true/false pair of IMU-specific claims; one open interaction question requiring two or more modalities; and one Yes-correct/No-correct pair, each naming the modality that would have prevented the error. Every claim carries an explanation and a self-assessed explanation-quality score.

Explanation-quality self-assessment, scored 1.0 / 0.5 / 0.0: 1.0 if the explanation cites a timestamp, names specific evidence (a verbatim utterance, a named object, a specific IMU segment), and states the modality of that evidence; 0.5 for two of those three; 0.0 for a generic or evidence-free explanation. Low-quality items are kept and flagged rather than dropped, so they can be filtered downstream.

Modality field rules: `requires` and `modality_that_prevents_error` are exactly one of video / audio / imu; a compound value (video+audio, etc.) appears only in an open-QA item's `requires`.
```

F.3.2 User Prompt

```
I'm showing you a {duration}-second egocentric video clip from a head-mounted camera. The file has audio muxed in -- listen to it natively for paralinguistic features and anything only audio can disambiguate. You also have structured annotations from separate audio and IMU analyses.

=== AUDIO ANNOTATION ===

Speech utterances (with tone):
{speech_json}

Sound events:
{sounds_json}

Non-verbal vocalizations:
{non_verbal_json}

Audio scene: {audio_scene_json}

=== IMU MOTION DATA ===

Motion segments:
{motion_segments_json}

Head orientation:
{head_orientation_json}

Transitions:
{transitions_json}

Bilateral (left vs. right):
{bilateral_json}

Sensor stats:
{sensor_stats_json}

=== AUDIO SUMMARY (natural language) ===

```

```
{audio_caption}

=== IMU SUMMARY (natural language) ===

{imu_caption}

Based on what you SEE in the video, HEAR in the muxed audio, the structured audio/IMU data above, and any ground-truth annotation given, produce a JSON response matching this schema:

{
  "observations": {
    "scene": "setting and key objects; describe people by grounded descriptor, never by name",
    "actions": ["action 1 with timing", "action 2 with timing"],
    "speech": {"language": "...", "speakers": 0, "topic": "..."},
    "sounds": ["sound 1", "sound 2"],
    "motion_state": "locomotion, intensity, head orientation",
    "temporal_phases": [
      {"time": "0-10s", "description": "..."},
      {"time": "10-20s", "description": "..."},
      {"time": "20-30s", "description": "..."}
    ]
  },
  "narrative": "first-person, 3-5 sentences, integrating video/audio/motion and any ground-truth annotation; no names",
  "benchmark": {
    "mcq": [
      {"question": "video question, no timestamp anchors", "options": ["A","B","C","D","E"], "answer": <0-4>, "modality": "video", "modality_set": ["video"]},
      {"question": "audio question", "options": ["A","B","C","D","E"], "answer": <0-4>, "modality": "audio", "modality_set": ["audio"], "question_type_audio": "..."},
      {"question": "IMU question, categorical only", "options": ["A","B","C","D","E"], "answer": <0-4>, "modality": "imu", "modality_set": ["imu"], "question_type_imu": "..."}
    ],
    "open_qa": [
      {"question": "requires video+audio", "answer": "...", "modality_set": ["video","audio"]},
      {"question": "requires video+imu", "answer": "...", "modality_set": ["video","imu"]},
      {"question": "requires audio+imu", "answer": "...", "modality_set": ["audio","imu"]},
      {"question": "IMU-anchored, IMU is the primary evidence", "answer": "...", "modality_set": ["video","imu"]}
    ],
    "fact_claims": [
      {"claim": "true visual", "label": true, "dimension": "visual", "modality": "video", "explanation": "...", "explanation_quality_self_assessment": 1.0},
      {"claim": "false visual", "label": false, "dimension": "visual", "modality": "video", "explanation": "...", "explanation_quality_self_assessment": 1.0},
      {"claim": "true audio", "label": true, "dimension": "audio", "modality": "audio", "explanation": "...", "explanation_quality_self_assessment": 1.0},
      {"claim": "false audio", "label": false, "dimension": "audio", "modality": "audio", "explanation": "...", "explanation_quality_self_assessment": 1.0},
      {"claim": "true motion", "label": true, "dimension": "motion", "modality": "imu", "explanation": "...", "explanation_quality_self_assessment": 1.0},
      {"claim": "false motion", "label": false, "dimension": "motion", "modality": "imu", "explanation": "...", "explanation_quality_self_assessment": 1.0},
      {"claim": "true IMU-specific, no timestamp anchors in the claim text", "label": true, "dimension": "imu_specific", "modality": "imu", "explanation": "...", "explanation_quality_self_assessment": 1.0},
      {"claim": "false IMU-specific", "label": false, "dimension": "imu_specific", "modality": "imu", "explanation": "...", "explanation_quality_self_assessment": 1.0},
      {"question": "interaction question, timestamps from 2+ modalities", "answer": "...", "dimension": "interaction"},
      {"question": "yes/no, correct=Yes", "answer": "Yes", "dimension": "hallucination", "modality_that_prevents_error": "video|audio|imu", "explanation": "...", "explanation_quality_self_assessment": 1.0},
      {"question": "yes/no, correct=No", "answer": "No", "dimension": "hallucination", "modality_that_prevents_error": "video|audio|imu", "explanation": "...", "explanation_quality_self_assessment": 1.0}
    ],
    "temporal_grounding": [
      {"question": "varied phrasing", "answer_start": 8, "answer_end": 12, "source": "video|audio|imu"}
    ]
  }
```

```
}

Respond only with valid JSON. No markdown code fences.
```

F.4 GRADING

Four of the benchmarks we report involve a language-model grader: our own, and three of the external suites. Each is graded under the prompt below, held fixed across every model, and the grader sees only the question or claim, the reference, and the prediction, never the media or the identity of the model that produced the answer. EgoSchema, EgoPolice and EgoLife-QA are multiple choice and scored in closed form, so no grader is involved and no prompt applies.

Our own benchmark uses three graders, one per open-ended task family, and each is built around a single question the grader must answer rather than a checklist, because a multi-step rubric collapses in practice into whichever step is easiest to answer.

F.4.1 EGOMMU

Grading for Fact Claim.

```
You are grading the REASONING in an egocentric fact-check.

The model's True/False verdict is scored separately and deterministically. Your job is ONLY to grade how well the model's reason is grounded in the clip.

=== WHY THE REASON IS GRADED AT ALL ===

The claim is True/False, so a model that understands nothing is still right half the time. The reason is the only evidence that a correct verdict was earned rather than guessed. Grade it on that basis.

=== WHAT COUNTS AS EVIDENCE ===

Evidence is something OBSERVED in the clip that bears on this specific claim: a named object, a described action, a sound, a moment or ordering. These are NOT evidence:

- Restating the claim, or asserting the verdict again in other words
- Generic plausibility ("this seems unlikely", "that is typical")
- Reasoning that would hold for ANY clip
- Describing the scene in a way that never touches what the claim asserts

Many claims are compound, joining two conditions with and / before / after / while / both ("Footsteps are heard as the wearer moves across the floor AFTER mixing the salad"). For those, a reason that settles one condition and ignores the other is PARTIAL, not complete. This is the single most common reason for a mid-range score, so check for it explicitly.

=== SCALE ===

5 = Settles the claim. Cites specific observed evidence covering EVERY condition the claim asserts, and that evidence is consistent with the ground-truth verdict.
4 = Settles the main point but is thin: the evidence is real and specific, yet vague about when or where it happens, or one minor qualifier of the claim is left unaddressed.
3 = PARTIAL. Cites real observed evidence for SOME of what the claim asserts but demonstrably not all of it -- most often one half of a compound claim, or the object but not the action, or the action but not the stated ordering or timing.
2 = Restates the claim or the verdict, or describes the clip only in terms so general that they neither support nor refute the claim. No specific observation is offered.
1 = Generic plausibility, off-topic, or the cited evidence CONTRADICTS the verdict the model gave (for example asserting True while describing the thing not happening).

A verdict with NO reason at all scores 1. If the answer is bare ("True.", "False.") or offers nothing beyond the verdict, there is no reasoning to grade and no evidence the clip was understood. Do not award a middle score for the absence of a claim to evaluate -- absence is the weakest case, not an average one.

```

```
Judge the reason as written. Do not credit a reason for being long, confident, or well phrased; do not penalise it for being brief if it names the right evidence. If the reason is cut off mid-sentence, grade only what is present.

Claim: {claim}

Ground-truth verdict: {label}

Model's answer (verdict + reason): {prediction}

Reply with ONLY the single digit 1, 2, 3, 4, or 5. No explanation.
```

Grading for Narration.

```
You are grading a description of a 30-second first-person (egocentric) clip against a reference description written by an annotator who saw, heard, and had motion data for that clip.

Answer ONE question only, and answer it about CONTENT.

=== THE QUESTION ===

How much of what the reference reports does the model description also report?

Count only what the reference actually says. Do not reward the description for detail the reference does not contain, and do not penalise it for omitting things the reference never mentions.

=== WHAT COUNTS AS A MATCH ===

The reference reports a set of facts about the clip. A fact is MATCHED when the description conveys the same thing, in any wording. Facts come in three kinds and all three count equally:

- Visible: the place, the objects, what the wearer or others are doing
- Audible: speech, voices, music, machine noise, footsteps, breathing, or explicit quiet
- Movement: the wearer walking, turning, pausing, crouching, scanning, standing still

A description that gets the room and the objects right but misses a conversation the reference reports has matched some facts and missed others. That is the thing you are counting.

=== WHAT DOES NOT COUNT ===

- A sound or movement the description asserts that the reference does NOT report is a fabrication. It does not add credit, and if there are several it should pull the score down: an invented soundscape is not evidence of listening.
- Restating the question, or generic scene-setting that could apply to any clip.
- Length, fluency, confidence, and preambles like "Based on the video clip". Ignore them entirely.
- If the description is cut off mid-sentence, grade what is there. Do not penalise the cut itself.

=== SCALE ===

Judge the SHARE of the reference's facts that the description matches:

4 = Nearly all of them. Almost nothing in the reference is missing, and nothing important is invented.
3 = Clearly more than half, with some real omissions or one or two fabrications.
2 = Roughly half. A substantial part of the reference is missing or contradicted.
1 = A small part. The description touches the clip but most of the reference is absent.
0 = Essentially none. A different scene, an empty answer, or almost entirely invented.

Most descriptions are not perfect and not worthless. If you find yourself about to give every description the same score, you are grading style rather than content; re-read the reference and count what is actually matched.

Reference description:
{reference}

Model description:
{prediction}

Reply with ONLY the single digit 0, 1, 2, 3, or 4. No explanation.
```

Grading for Open-Ended QA.

```
You are grading an egocentric video-QA answer. The prediction is CORRECT if
it conveys the meaning of the reference answer; minor wording differences
are fine.

[included only when the question spans more than one modality]
This question spans MORE THAN ONE modality: what is SEEN and what is HEARD
and how the wearer MOVED. The reference answer resolves all of them. Mark
INCORRECT if the prediction addresses only one side and leaves the other
unanswered or wrong -- answering half of a multi-modal question is not a
correct answer.

Mark INCORRECT if the prediction hedges across several possibilities so that
one of them happens to match, restates the question without answering it, or
answers a different question than the one asked. A prediction that lists
alternatives is only correct if the reference itself is that broad.

Question: {question}

Reference answer: {answer}

[included only when the item declares alternatives]
Also acceptable (equivalent answers): {aliases}

Prediction: {prediction}

Reply with ONLY 1 (correct) or 0 (incorrect). No explanation.
```

F.4.2 EGOAVU-BENCH

The source suite's own rubric, ported unchanged, which is why the five-point scale and the structured output are its conventions rather than ours. The integer rating is the reported value; the accompanying reason is retained but not scored.

```
Objective:
Act as an impartial grader to evaluate a PREDICTED_ANSWER against a GROUNDING_ANSWER with respect
    ↪ to a QUESTION.

Input:
- QUESTION: {question}
- GROUNDING_ANSWER: {grounding}
- PREDICTED_ANSWER: {prediction}

Instructions for Grading:
1. Compare the PREDICTED_ANSWER to the GROUNDING_ANSWER with respect
   to the QUESTION.
2. Assign Rating (1-5, integer only):
   - 5: Fully correct, complete, and faithful to the grounding.
   - 4: Mostly correct; minor omissions or inaccuracies.
   - 3: Partially correct; misses important details.
   - 2: Largely incorrect; substantial errors.
   - 1: Incorrect or irrelevant.
3. Briefly explain the rating.

Output Format:
Return valid JSON only:
{
  "rating": <int between 1 and 5>,
  "reason": "<brief explanation>"
}
```

F.4.3 EGOTHINK

A binary judgement against the single reference the source provides, tolerating paraphrase and additional detail but not hedging. The source grades the same shape of judgement with a different

model, so figures produced under this prompt are comparable to one another but not to the published ones.

```
You are grading an answer to a question about a first-person photograph. The prediction is CORRECT if it conveys the meaning of the reference answer; minor wording differences, extra detail, and paraphrase are fine. It is INCORRECT if it contradicts the reference, answers a different question, or hedges without committing to an answer.

Question: {q}

Reference answer: {ref}

Prediction: {pred}

Reply with ONLY 1 (correct) or 0 (incorrect). No explanation.
```

F.4.4 EGOBLIND

This suite carries between one and four accepted answers per question, because the useful response to an assistive question is frequently not unique. Which of them a judgement may match against changes what is being measured, so two prompts exist and the choice between them is explicit.

Matching any accepted answer. The default. A prediction is correct if it conveys the meaning of any one of the references, which is the appropriate form when several answers genuinely serve the asker.

```
You are grading an egocentric visual-QA answer for a blind user. The prediction is CORRECT if it conveys the meaning of ANY ONE of the acceptable reference answers (match-to-any); minor wording differences are fine.

Question: {q}

Acceptable reference answers: {refs}

Prediction: {pred}

Reply with ONLY 1 (correct - matches at least one reference) or 0 (incorrect). No explanation.
```

Matching one designated answer. Used where the system under test was trained against a single designated reference. Admitting any alternative there would credit an answer the system never saw, measuring the breadth of the reference set rather than the system.

```
You are grading an egocentric visual-QA answer for a blind user. The prediction is CORRECT if it conveys the meaning of the reference answer; minor wording differences are fine.

Question: {q}

Reference answer: {refs}

Prediction: {pred}

Reply with ONLY 1 (correct - matches the reference) or 0 (incorrect). No explanation.
```

G Public Resources Used

G.1 Datasets

We acknowledge the use of the following public data sources, during the course of this work:

- EgoLife[1] ..... MIT License
- Ego-Exo4D[2] ..... Ego-Exo4D Dataset License
- Ego4D[3] ..... Ego4D Dataset License
- HoloAssist[4] ..... CDLAv2 License
- Aria Everyday Activities[5] ..... Aria Everyday Activities Dataset License
- Nymeria[6] ..... CC BY NC 4.0 License
- EgoAVU-Bench[7] ..... FAIR Noncommercial Research License
- EgoSchema[8] ..... Ego4D Dataset License
- EgoThink[9] ..... Apache License 2.0
- EgoBlind[10] ..... CC BY NC 4.0 License
- EgoPolice[11] ..... MIT License

G.2 Models & Checkpoints

We acknowledge the use of the following public models and their checkpoints, during the course of this work:

- Qwen3.5[12] ..... Apache License 2.0
- Qwen3-VL[13] ..... Apache License 2.0
- Qwen2.5-VL[14] ..... Apache License 2.0
- Qwen3-Omni[15] ..... Apache License 2.0
- Qwen2.5-Omni[16] ..... Apache License 2.0
- InternVL3[17] ..... Apache License 2.0
- Intern-Omni[18] ..... MIT License
- Gemma 4[19] ..... Apache License 2.0
- MiniCPM-o[20] ..... Apache License 2.0
- Baichuan-Omni-1.5[21] ..... Apache License 2.0

[1] https://egolife-ai.github.io.
[2] https://ego-exo4d-data.org.
[3] https://ego4d-data.org.
[4] https://holoassist.github.io.
[5] https://www.projectaria.com/datasets/aea.
[6] https://www.projectaria.com/datasets/nymeria.
[7] https://huggingface.co/datasets/facebook/EgoAVU_data.
[8] https://egoschema.github.io.
[9] https://github.com/AdaCheng/EgoThink.
[10] https://github.com/doc-doc/EgoBlind.
[11] https://github.com/princetonvisualai/egopolice.
[12] https://huggingface.co/collections/Qwen/qwen35.
[13] https://huggingface.co/collections/Qwen/qwen3-vl.
[14] https://huggingface.co/collections/Qwen/qwen25-vl.
[15] https://huggingface.co/collections/Qwen/qwen3-omni.
[16] https://huggingface.co/collections/Qwen/qwen25-omni.
[17] https://huggingface.co/collections/OpenGVLab/internvl3.
[18] https://huggingface.co/OpenGVLab/InternOmni.
[19] https://huggingface.co/collections/google/gemma-4.
[20] https://huggingface.co/openbmb/MiniCPM-o-2_6.
[21] https://huggingface.co/baichuan-inc/Baichuan-Omni-1d5.

- VideoLLaMA2.1[22] . . . . . Apache License 2.0
- EgoGPT[23] . . . . . S-Lab License
- Phi-4-Multimodal[24] . . . . . MIT License
- BEATs[25] . . . . . MIT License
- Whisper[26] . . . . . MIT License

G.3 Proprietary Model Services

We use proprietary model services through their APIs for three purposes:

1. Evaluating proprietary models as baselines on the benchmarks considered in this work.
2. Generating per-modality captions and benchmark items, as described in Appendix A.
3. Grading model responses according to the evaluation protocols in Section C.

We acknowledge the following service providers. Access to their models is governed by the respective terms of service:

- OpenAI[27] . . . . . Provider Terms of Service
  GPT-4o, GPT 4.1, GPT 5, GPT 5 Mini, GPT 5 Nano, GPT 5.4, GPT 5.4 Mini, GPT 5.4 Nano
- Google[28] . . . . . Provider Terms of Service
  Gemini 1.5 Pro, Gemini 2.5 Pro, Gemini 2.5 Flash Lite, Gemini 3.1 Pro, Gemini 3.1 Flash Lite, Gemini 3.5 Flash, Gemini 3.6 Flash, Gemini 3.7 Flash, Gemini 3.8 Flash
- Anthropic[29] . . . . . Provider Terms of Service
  Claude Haiku 4.5, Claude Sonnet 4.6, Claude Sonnet 5, Claude Opus 4.6, Claude Opus 4.8, Claude Opus 5

[22] https://huggingface.co/DAMO-NLP-SG/VideoLLaMA2.1-7B-AV.
[23] https://github.com/EgoLife-ai/EgoLife.
[24] https://huggingface.co/microsoft/Phi-4-multimodal-instruct.
[25] https://github.com/microsoft/unilm/tree/master/beats.
[26] https://github.com/openai/whisper.
[27] https://platform.openai.com.
[28] https://ai.google.dev/gemini-api.
[29] https://www.anthropic.com/api.

H LIMITATIONS & BROADER IMPACT

H.1 LIMITATIONS

Data and Benchmark Coverage. EgoMM combines several egocentric datasets, but its coverage remains bounded by their participants, activities, environments, and capture devices. Fixed 30 s windows can separate events from their broader context and limit questions requiring longer temporal dependencies. Prompt-based annotation and model-panel filtering also influence the vocabulary and difficulty of retained items. Human verification and shortcut checks reduce errors and unintended cues, but do not establish generalization to unseen users or guarantee that every question requires its designated modality.

Sensor Representations and Attribution. The audio pathway targets sound events and paralinguistic cues, leaving detailed speech understanding outside its intended scope. Head-mounted IMUs measure head motion and provide only indirect evidence about broader activity. Their text descriptors summarize predefined signal properties whose interpretation can depend on device placement and calibration. Since IMU withholding removes both learned tokens and tag text, the reported ablations measure their combined contribution. Separate evaluations of these inputs would clarify the contribution of learned representations.

Evaluation Reliability. Automated grading measures agreement with references and can inherit their omissions or ambiguities. The graders do not inspect the media, so their judgments cannot independently verify every detail in a response. Modality ablations provide complementary evidence, although residual performance can reflect shared context and learned priors. Temporal localization remains challenging, while the small captioning probes provide limited evidence about generalization. Further human assessment would help characterize these sources of uncertainty.

Long-Horizon Memory. The memory experiments cover one participant and use video and audio, leaving the contribution of IMU to long-horizon retrieval untested. Retrieval operates over precomputed summaries with coarse temporal restrictions; records and summaries are not strictly truncated at the question time and may contain later observations. A strictly causal evaluation is therefore needed to assess online assistance. Text consolidation can also omit useful details or propagate captioning errors, while retaining all clip records makes storage grow with recording duration.

Capture and Deployment Efficiency. The capture experiments use uniform frame sampling with full-clip audio. They do not evaluate sensor-triggered acquisition or establish reductions in device power, heat, or end-to-end latency. Deployment requires measuring sensor processing, memory construction, and retrieval costs, alongside robustness to synchronization errors and degraded inputs.

H.2 BROADER IMPACT

Potential Benefits. EgoMM can support research on wearable assistants that describe surrounding activity, recover past observations, and answer questions using complementary sensory evidence. These capabilities may benefit personal memory, accessibility, and everyday task support. Modality-specific evaluation can also help reveal whether predictions depend on relevant observations, providing a basis for assessing the reliability of assistive systems.

Privacy and Misuse. Egocentric recordings and derived memories may reveal sensitive information about wearers and bystanders, including their routines, locations, and interactions. Removing names from generated annotations does not anonymize the underlying media or prevent identity inference from context. Persistent memory could also enable intrusive surveillance or behavioral profiling. Responsible deployment should address consent, access control, retention limits, and user control over recording and deletion. Derived resources should preserve the access and redistribution restrictions of their source datasets.

Reliability and User Agency. Incorrect descriptions or retrieved memories may mislead users, particularly when fluent answers obscure missing evidence. Benchmark performance alone does not establish suitability for safety-sensitive assistance, and aggregate scores may conceal differences across users and environments. Future systems should communicate uncertainty, retain links to supporting observations, and allow users to inspect and correct stored information. Evaluation with intended users is needed to assess usefulness, uneven performance, and risks of overreliance.