VISTA: A MULTIMODAL EGOCENTRIC BENCHMARK FOR GOAL-ORIENTED PERCEPTION AND ASSISTANCE TO BLIND AND LOW-VISION USERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Blindness and low vision limit many people's ability to complete daily tasks, motivating AI assistants that provide useful guidance. Yet, existing benchmarks do not adequately test whether vision models can provide effective everyday assistance. In particular, i) existing benchmarks primarily rely on visual and textual inputs, with limited support for other modalities that are increasingly available in egocentric wearable devices such as gaze, inertial measurements, spatial trajectories, and others; and ii) existing evaluations emphasize scene understanding and question answering, with limited coverage of goal-conditioned guidance for blind and low-vision users across everyday tasks. To bridge this gap, we introduce **VISTA** (**V**isually **I**mpaired **S**cene and **T**ask **A**nnotations), a multimodal egocentric benchmark for goal-oriented assistance for blind and low-vision users. VISTA contains $\sim$1,000 egocentric recordings captured with Project Aria glasses; each recording provides synchronized sensor streams, six of which serve as model inputs, and three different types of annotation, covering ten task categories (TC), from object localization to hazard detection. We evaluate 14 representative vision-language models zero-shot on VISTA and design a lightweight multimodal model baseline that integrates all six modalities. Experimental results show that, although existing models can often describe egocentric scenes, they remain limited in generating goal-directed assistive guidance. VISTA thus provides a benchmark for evaluating multimodal models that could improve everyday accessibility.

1 INTRODUCTION

Blindness and low vision can substantially limit individuals' ability to complete daily tasks, creating an urgent need for benchmarks that evaluate assistance in everyday settings (World Health Organization, 2019). Individuals regularly encounter visual information that is difficult or impossible to access without assistance: locating objects, detecting hazards, reading text, interpreting social signals, estimating distance, and navigating unfamiliar environments. Recent advances in vision-language models (VLMs) (Radford et al., 2021; Li et al., 2023; Liu et al., 2023; Hu et al., 2024; OpenAI, 2023) suggest that they could become a core component of wearable assistive systems. However, reliable assistance requires more than generic image understanding. A useful assistant must interpret an egocentric scene from the user's perspective, infer what matters for the user's current goal (Behari et al., 2024), avoid unsafe suggestions, and generate concise guidance that can be acted on immediately.

Current benchmarks do not fully capture these requirements. While standard VQA and captioning datasets (Goyal et al., 2017; Lin et al., 2014) evaluate perception in static settings, and accessibility-oriented datasets (Gurari et al., 2018; Xiao et al., 2025) better reflect images captured by blind and low-vision users, they exhibit two critical limitations. First, existing benchmarks mainly focus on visual and textual inputs, overlooking other sensing modalities that are increasingly available in modern egocentric wearable devices, such as audio, eye tracking, inertial signals, and spatial tracking. Second, current models are typically evaluated as vision-language systems, rather than as *multimodal assistive models* capable of integrating visual observations, gaze, motion, and spatial trajectories (Chen et al., 2026). A model may correctly describe the objects in a scene while missing which object the user is attending to, or misjudge spatial relations due to a lack of inertial context.

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.

This gap motivates a benchmark whose target is not only "what is in the image?" but also "where is the user, how are they moving, and what should they do next?"

We present **VISTA**, a multimodal egocentric benchmark for evaluating vision-language models as next-step assistants for blind and low-vision users. VISTA contains ∼1,000 short wearable recordings (∼170 min) collected with Project Aria glasses (Engel et al., 2023) across ten assistive task categories, including object location, hazard detection, navigation, social cues, distance estimation, task instruction, object query, text understanding, and motion understanding. Unlike prior works, VISTA synchronizes egocentric clips combining RGB, stereo SLAM and eye-tracking cameras, dual IMU, magnetometer and barometer, together with 6-DoF trajectories, semi-dense point clouds, per-frame gaze and hand tracking and natural-language annotations for scene description, visual question answering, and goal-conditioned instructions. The goal-instruction annotations require models to reason about relevant objects, spatial relations, hazards, and social cues to generate actionable guidance.

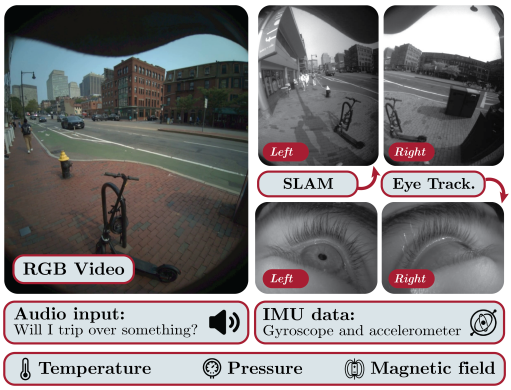

Figure 1: **Sensor modalities of the Aria glasses.** VISTA records synchronized RGB video, stereo SLAM and eye-tracking cameras, audio, IMU, barometer, and magnetometer. See Appendix A.2 (Figure 6) for an expanded visualization.

We explicitly frame VISTA as a benchmark for *next-step assistance*. Given a wearable first-person recording and a user goal, VISTA requires models to translate egocentric scene understanding into user-centered actions. We conduct zero-shot evaluations with open and proprietary vision-language models and design a lightweight multimodal baseline whose inputs are RGB video and five sensor streams (six modalities in total). Experimental results show that although existing models can often describe scenes, they remain limited in producing task-oriented, safety-aware guidance.

To summarize, our contributions are:

1. We introduce **VISTA**, a multimodal egocentric benchmark for goal-oriented assistance for blind and low-vision users.
2. We evaluate state-of-the-art VLMs zero-shot on VISTA and show that current models make frequent errors on assistive tasks, particularly on goal-conditioned guidance.
3. We design a multimodal baseline that fuses these sensor streams with RGB; combined with task adaptation, it improves answer quality on every backbone, most strongly on guidance.

2 Related Work

Egocentric video understanding. First-person datasets such as EPIC-Kitchens (Damen et al., 2018; 2022), Charades-Ego (Sigurdsson et al., 2018), Ego4D (Grauman et al., 2022), and Ego-Exo4D (Grauman et al., 2024) support activity recognition and long video understanding. Egocentric QA benchmarks test task understanding (Jia et al., 2022), long-term reasoning (Mangalam et al., 2023), episodic memory (Bärmann & Waibel, 2022), instructional help (Wong et al., 2022), and scene-text reading (Zhou et al., 2025). These benchmarks target general users and mostly use RGB video, even though devices such as the Project Aria glasses (Engel et al., 2023) also record audio, IMU, gaze, and trajectory data. VISTA combines these sensor streams with annotations written for blind and low-vision users.

Visual assistance for blind and low-vision users. VizWiz (Gurari et al., 2018) contains questions that blind users asked about their own photos, but it is limited to single images. VIEW-QA (Song et al., 2024) and EgoBlind (Xiao et al., 2025) extend this setting to egocentric video, while WalkVLM (Yuan et al., 2025) and GuideDog (Kim et al., 2026) focus on walking and navigation. Assistive apps and prototypes based on OCR, SLAM, or depth sensing solve specific tasks but are hard to compare (Okolo et al., 2024; Mukhiddinov & Cho, 2021). Most of these works



Figure 2: **Examples of the ten task categories.** Each strip shows three clips, with the RGB frame on top and the SLAM and eye-tracking frames below. Each category has about 100 clips (997 in total); definitions are in Table 4.

use RGB input with QA or navigation targets only. VISTA adds synchronized wearable sensors and provides scene descriptions, question–answer pairs, and goal-conditioned instructions across ten assistive task categories.

Multimodal models and sensor fusion. Open VLMs such as LLaVA-Video (Zhang et al., 2025b), InternVL3 (Zhu et al., 2025), Qwen3-VL (Bai et al., 2025), and Gemma 3 (Gemma Team et al., 2025) perform well on images and video but only accept visual and text inputs. For other modalities, attention bottlenecks (Nagrani et al., 2021) exchange information between streams through a few shared tokens, PRIMUS (Das et al., 2025) provides a pretrained IMU encoder, and LoRA (Hu et al., 2022) adapts large models with few trainable parameters. Few works apply these tools to assistive VLMs with wearable sensors; we combine them in a sensor-augmented baseline for VISTA (Sec. 4).

Evaluating open-ended assistance. Reference-based metrics such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), BERTScore (Zhang et al., 2020), and Sentence-BERT similarity (Reimers & Gurevych, 2019) measure overlap with references but do not check whether a response matches the scene, completes the task, or is useful to a blind user. LLM-as-a-judge methods (Zheng et al., 2023) score open-ended answers more flexibly and agree well with human ratings. VISTA uses a vision-language judge that sees the video frames and scores evidence grounding, answer correctness, task completion, and accessible communication.

3 VISTA Benchmark

3.1 Overview

VISTA contains egocentric recordings captured with first-generation Project Aria glasses (Engel et al., 2023). The recordings are organized into ten assistive task categories with approximately balanced coverage (Figure 2). Each recording lasts between roughly 10 seconds and one minute and synchronizes multiple modalities. Participants wore the glasses while moving through indoor and outdoor environments and simulating practical assistive situations. The recordings include kitchens, corridors, streets, transit-like settings, parks, and other everyday locations with variation in lighting,

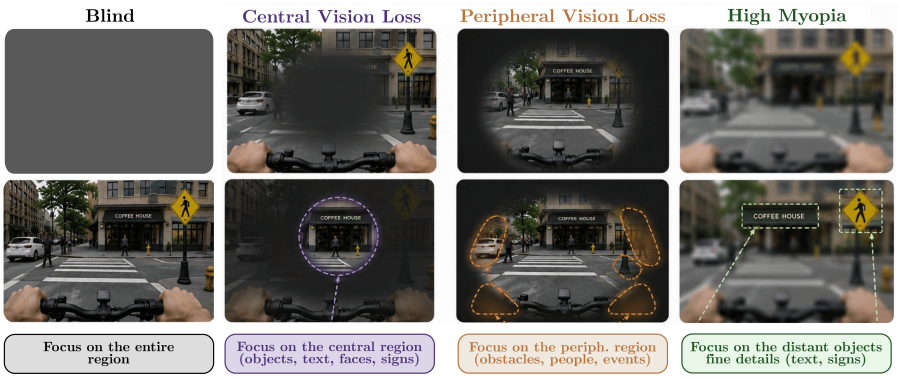

Figure 3: **Simulated vision conditions (conceptual illustration).** Top: The same scene under blindness, central vision loss, peripheral vision loss, and high myopia. Bottom: Regions that may be most useful to describe for each condition . The regions are illustrative and are not evaluated in this work.

motion, pedestrian activity, and ambient noise. This diversity in capture conditions and sensing modalities is important, as the specific information a user needs from a scene depends heavily on the nature of their visual impairment. Figure 3 is a conceptual illustration of how different vision conditions may change which parts of a scene are most useful to describe.

3.2 Data Collection Setup

The Aria glasses record multiple time-aligned sensor streams. The sensor suite includes a forward-facing egocentric RGB camera; two grayscale fisheye SLAM cameras (left and right) used for visual odometry and spatial localization; an inward-facing eye-tracking camera; two IMUs (left and right) measuring 3-axis acceleration and angular velocity; a barometer; and a magnetometer measuring the 3-axis magnetic field. The zero-shot baselines use RGB frames and task text. Our multimodal baseline adds five sensor streams derived from these recordings (IMU, magnetometer/barometer, eye gaze, hand tracking, and 6-DoF trajectory; Section 4); audio and the raw SLAM and eye-tracking images are recorded but are not model inputs.

3.3 Data Annotation Pipeline

Figure 4 illustrates the end-to-end annotation pipeline, from clip review to structured label writing and quality control.

Annotation schema. VISTA annotations are designed around assistive use rather than generic visual labeling. Each clip contains three annotation types:

i) Scene descriptions summarize the environment, relevant objects, spatial layout, events, and possible risks.

(i) Scene description

"You are standing beside a large tree in a grassy park. Thick roots extend across the ground directly in front of you, creating a tripping hazard."

ii) Question-answer pairs represent questions a blind or low-vision user might ask, paired with answers grounded in the clip.

(ii) Question–answer pair

Question: "How could I get a fork?" **Answer:** "The utensil dispenser is directly ahead on the counter. [...] press down on a handle in one of the two left slots to release a fork."

iii) Goal-instruction pairs state a high-level assistive goal and a concrete instruction for accomplishing it.

(iii) Goal–instruction pair

Goal: "Bypass the landscaping obstacle to reach the street." **Instruction:** "Take 2 steps to the right, and follow the paved path around the bush, avoiding the dirt bed."

The goal-instruction layer is the main distinction from conventional captioning or VQA. For example, instead of only describing that a staircase is visible, an annotation may define the goal "go down the stairs safely" and the instruction "move forward slowly, keep to the handrail on the right, and step down carefully." This structure enables evaluation of guidance quality, specificity, and safety.

Annotation protocol. Eight trained annotators labeled the clips in a customized Label Studio (Tkachenko et al., 2020) interface, following written guidelines and a demonstration of the workflow. For each clip, annotators watched the full video, wrote a clip-level scene description of relevant objects, layout, events, and risks, and then wrote question–answer pairs and one or more goal–instruction pairs that a blind or low-vision user might need. They could pause, rewind, or revisit frames to verify object locations, visible text, motion, and safety-relevant details. Task labels follow the test-case subset (TC01–TC10), and all natural-language fields are written manually.

A reviewer from the blind and low-vision (BLV) community assessed 27 clips (34 annotations) for clarity, usefulness, and appropriateness. The feedback pointed to overly long responses, unclear spatial descriptions, missing hazard information, and instructions that relied on the user's vision. We added this feedback to the guidelines, and three annotators then revised every clip in a second full pass.

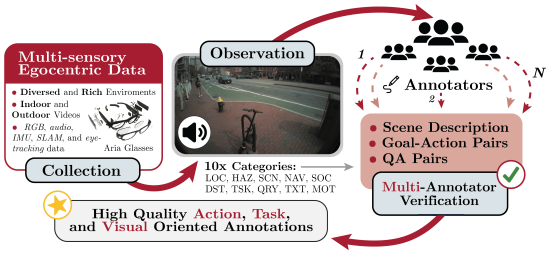

Figure 4: **Annotation pipeline.** Recorded clips are annotated by multiple annotators with scene descriptions, goal–action pairs, and QA pairs, and then verified.

To measure consistency, 30 clips were labeled by multiple annotators. Their annotations reach a BERTScore (Zhang et al., 2020) of 0.897 for question–answer pairs and 0.877 for scene descriptions (SBERT (Reimers & Gurevych, 2019): 0.634 and 0.666), indicating similar content with varied wording. We keep all overlapping annotations as multiple references rather than merging them.

Accessibility and ethics. All participants provided consent before data collection. Recordings in public spaces were restricted to nonsensitive environments. Face and license-plate blurring (Raina et al., 2023) are applied before release to reduce bystander privacy risk. Since the target domain involves assistance for disabled users and safety-relevant guidance, VISTA should be used with care: benchmark scores should not be interpreted as sufficient evidence for deployment without human-centered evaluation, safety testing, and participation from blind and low-vision users.

Annotation Statistics. The current annotation release comprises 1,063 scene descriptions, 1,191 question-answer pairs, and 2,067 goal-instruction pairs across 997 clips. On average, each clip includes 1.07 scene descriptions, 1.19 question-answer pairs, and 2.07 goal-instruction pairs. Table 6 summarizes these statistics globally.

Across the full dataset, the ten task categories are approximately balanced, containing 97–102 clips, with minor variation attributable to the collection and filtering process. Annotation density is likewise consistent across subsets, ranging from 1.00 to 1.12 scene descriptions per clip, 1.09 to 1.29 question-answer pairs per clip, and 1.94 to 2.19 goal-instruction pairs per clip. This provides approximately balanced coverage across the ten assistive task categories, reducing skew toward any single task type.

3.4 Dataset Comparison

Table 1 compares VISTA with representative egocentric VQA datasets. Existing general-purpose datasets, such as EgoTaskQA, VidEgoThink, EgoSchema, and EgoMemoria, emphasize broad reasoning over egocentric videos (Jia et al., 2022; Cheng et al., 2024; Mangalam et al., 2023; Ye et al., 2025), but they are not designed around blind and low-vision users. Assistive datasets such as QAEgo4D, AssistQ, and EgoTextVQA move closer to task-oriented assistance (Bärmann & Waibel,

Table 1: **Comparison with egocentric and assistive VQA datasets.** BLV: blind and low-vision users; Multi-Ans.: multiple reference answers; Ques.-Cat.: question-category labels; OE/MC: open-ended/multiple-choice QA. ✓/✗: supported/not supported; –: not reported. Human eval.: human assessment reported; for VISTA, a BLV community reviewer assessed 34 annotations from 27 clips (model outputs were not human-evaluated). VISTA (highlighted) is the only dataset among those compared that supports scene description, QA, and goal-driven instruction together.

Type	Dataset	# Videos	# Queries	Avg. len. (s)	Setting		Modalities					Annotation		QA format		Supported tasks			Human eval.
					BLV orient.	Smart glasses	RGB	Audio	Gaze	IMU	SLAM/ depth	Multi-Ans.	Ques.-Cat.	OE	MC	Scene desc.	QA	Goal instr.	
General purpose	EgoTaskQA (Jia et al., 2022)	2K	40K	25.0	✗	✗	✓	–	✗	✗	✗	✗	✓	✓	✗	✗	✓	✗	–
	VidEgoThink (Cheng et al., 2024)	217	600	23.4	✗	✗	✓	–	✗	✗	✗	✗	✓	✓	✗	✗	✓	✗	–
	EgoSchema (Mangalam et al., 2023)	5K	5K	180.0	✗	✗	✓	–	✗	✗	✗	✗	✗	✗	✓	✗	✓	✗	–
	EgoMemoria (Ye et al., 2025)	629	7K	858.5	✗	✗	✓	–	✗	✗	✗	✗	✗	✗	✓	✗	✓	✗	–
Assistive	QAEgo4D (Bärmann & Waibel, 2022)	1.3K	14.5K	495.1	✓	✗	✓	–	✗	✗	✗	✗	✗	✓	✗	✗	✓	✗	–
	AssistQ (Wong et al., 2022)	100	531	115.0	✓	✗	✓	–	✗	✗	✗	✗	✓	✗	✓	✗	✓	✓	–
	EgoTextVQA (Zhou et al., 2025)	1.5K	7.0K	101.7	✓	✗	✓	–	✗	✗	✗	✗	✓	✓	✗	✗	✓	✗	–
BLV	VIEW-QA (Song et al., 2024)	1.0K	4.1K	34.4	✓	✗	✓	–	✗	✗	✗	✗	✓	✓	✗	✗	✓	✗	–
	EgoBlind (Xiao et al., 2025)	1.3K	5.3K	40.0	✓	✗	✓	–	✗	✗	✗	✓	✓	✓	✗	✗	✓	✗	✓
	VISTA (Ours)	**1.0K**	**1.2K**	**10.5**	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓	✓	✓

2022; Wong et al., 2022; Zhou et al., 2025), yet they mainly focus on questions and answers and do not provide the richer supervision needed for grounded guidance. Datasets centered on blind and low-vision users, including VizWiz, VIEW-QA, and EgoBlind, are more aligned with accessibility scenarios (Gurari et al., 2018; Xiao et al., 2025), but they are typically image-based or limited to QA-style annotations rather than multi-level assistive understanding.

VISTA fills this gap by combining assistive scenarios designed around blind and low-vision users (recorded by sighted participants who simulated them), egocentric video, open-ended QA, scene descriptions, and goal-instruction pairs in a single benchmark. This design makes VISTA more suitable for evaluating whether a model can not only answer questions about an egocentric scene, but also summarize the environment and produce actionable, safety-aware guidance for BLV users.

4 Multimodal Baseline

VISTA supports both evaluation of current vision-language models and development of assistive models that use synchronized wearable sensor streams. The benchmark results show that RGB-only models still struggle with tasks that require spatial reasoning, action understanding, and assistive guidance. We therefore augment visual and textual inputs with VISTA's sensor streams. Our baseline does so through sensor fusion and LoRA adaptation of a pretrained VLM.

Figure 5 illustrates the multimodal architecture, which augments a pretrained VLM (InternVL3, Qwen3-VL, or Gemma 3) with five sensor streams that, together with RGB video, form six modalities: IMU, magnetometer/barometer, eye gaze, hand tracking, and 6DoF trajectory. Eight RGB frames are processed at full 1408×1408 coverage by the backbone's frozen vision encoder, while the task query is encoded using the language model's text embeddings. IMU windows are embedded by a frozen PRIMUS encoder (Das et al., 2025); the remaining streams are aligned to a 64-bin temporal grid, normalized with training-split statistics, and paired with validity masks. For each stream, a temporal adapter projects the observations to a shared 256-dimensional space and uses eight learned queries to produce eight tokens. Question-conditioned reliability gates weight the streams, and two bottleneck-exchange layers (Nagrani et al., 2021) combine them into eight shared tokens, which replace dedicated sensor placeholder tokens in the prompt. Only the language LoRA adapters and the sensor-fusion network are trained (evaluation in Sec. 5.3).

Sensor inputs. The Aria glasses (Engel et al., 2023) provide several signals that complement RGB observations. RGB frames capture visible objects, text, people, and environmental hazards. Eye gaze provides a cue to the user's current visual attention, and hand tracking describes how the user interacts with nearby objects. IMU and trajectory signals describe device motion and user movement, while the magnetometer and barometer provide heading and changes in elevation. Because these streams differ in sampling rate, dimensionality, and availability, missing observations are represented by explicit masks rather than treated as valid measurements. Combining these signals allows the model to condition its response on more than the visible scene.

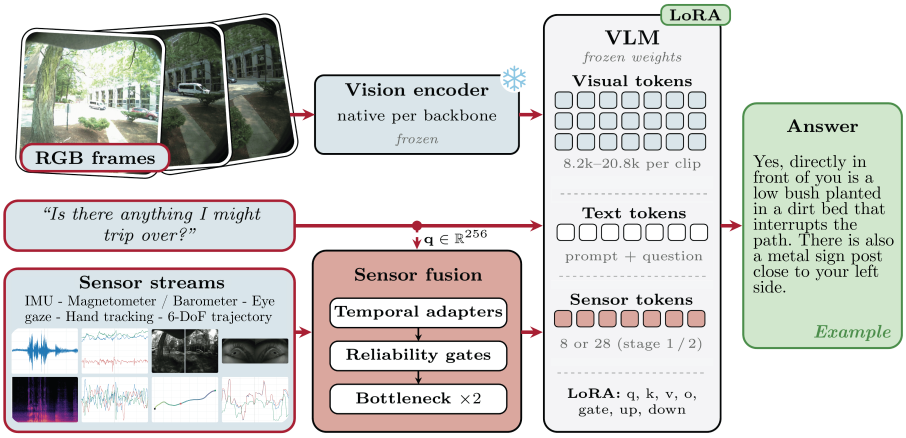

Figure 5: **Multimodal baseline.** RGB frames are encoded by the frozen vision encoder of the VLM. The five sensor streams used as inputs (IMU, magnetometer/barometer, eye gaze, hand tracking, 6DoF trajectory) pass through temporal adapters, question-conditioned reliability gates, and bottleneck exchange, giving 8 (stage 1) or 28 (stage 2) sensor tokens that are inserted into the prompt. Only the LoRA adapters and the sensor-fusion network are trained.

5 Experiments

5.1 Evaluation Setting

VISTA evaluates three response forms: scene description, visual question answering, and goal-conditioned guidance. We test InternVL3 (Zhu et al., 2025), Qwen3-VL (Bai et al., 2025), Gemma 3 (Gemma Team et al., 2025), and VideoLLaMA3 (Zhang et al., 2025a) across model scales, together with Gemini 3.1 Flash-Lite (Google, 2026) and GPT-Realtime Mini (OpenAI, 2025) as proprietary zero-shot baselines. All experiments share the VISTA test split, task prompts, eight uniformly sampled frames, and a maximum output length of 4,096 tokens. As an external reference, we evaluate ten of these models (eight open-source and both proprietary) on EgoBlind (setup in Appendix B, results in Appendix C.2).

The evaluation protocol is shared by the RGB zero-shot baselines and the multimodal model experiments. GPT-6 Luna (OpenAI, 2026) scores each prediction against the input frames and reference annotations on evidence grounding (0–1.75), answer correctness (0–1.75), task completion (0–1), and accessible communication (0–0.5), yielding a total from 0 to 5. Predictions containing potentially harmful instructions are flagged, with the final score capped at 1.95. The multimodal experiments change only the available sensor inputs or adaptation configuration; prompts, test data, judge, and rubric remain fixed. Appendix B gives the judge prompt, inference settings, and reproducibility details.

5.2 Zero-shot Results

Zero-shot Analysis. Table 2 shows that larger models generally improve on VISTA, although gains are not monotonic across every family. Across models, guidance remains harder than scene description and question answering; for example, InternVL3-14B scores 3.81 on scene description but 2.91 on guidance. On EgoBlind, evaluated with the same judge (Table 12), model rankings are broadly similar, but VISTA separates models much more: across the eight models evaluated on both benchmarks, scores span 3.11–3.58 on EgoBlind and 1.76–3.31 on VISTA. VISTA additionally evaluates three response forms and provides synchronized RGB, audio, gaze, inertial, and spatial streams.

Failure Case Analysis. Four error types recur across all models: body-relative orientation, objects outside the current view, dynamic motion, and metric distance. We count an item as a universal failure when all zero-shot models score below 3/5 on it. Even including GPT-Realtime Mini and Gemini 3.1 Flash-Lite, 4.3% of VISTA items meet this criterion, and these failures fall mostly on goal-instruction items (7.7%) rather than question-answer items (2.4%), suggesting that the bottle-

Table 2: **Zero-shot results on VISTA.** Judge scores (0–5). EG: evidence grounding (0–1.75); AC: answer correctness (0–1.75); TC: task completion (0–1); ACom: accessible communication (0–0.5); Scene, Q&A, Guidance: mean score per annotation type. **Bold**/underline: best/second-best per column within each model group; –: size not disclosed.

Model	Size	Score	Rubric components				Response type		
			EG	AC	TC	ACom	Scene	Q&A	Guidance
Open-source models (frames + text)									
InternVL3	1B	2.39	0.82	0.79	0.47	0.33	3.49	3.09	1.42
	2B	2.75	0.96	0.91	0.54	0.35	3.77	3.19	1.96
	8B	3.16	1.09	1.07	0.64	0.38	3.72	3.48	2.69
	9B	3.13	1.08	1.06	0.63	0.38	3.63	3.32	2.76
	14B	**3.33**	**1.15**	**1.13**	0.68	**0.40**	**3.81**	3.62	**2.91**
VideoLLaMA3	7B	1.76	0.61	0.57	0.32	0.27	2.48	2.43	1.00
Qwen3-VL	2B	2.65	0.91	0.89	0.54	0.33	3.48	3.33	1.82
	4B	3.22	1.11	1.10	0.68	0.38	3.78	3.69	2.65
	8B	3.31	**1.15**	**1.13**	**0.69**	0.39	3.76	3.69	2.85
Gemma 3	4B	2.73	0.92	0.91	0.58	0.34	3.19	3.49	2.04
	12B	3.07	1.04	1.05	0.66	0.36	3.41	3.76	2.48
	27B	3.06	1.03	1.05	0.66	0.35	3.32	**3.81**	2.48
Proprietary models (frames + text)									
GPT-Realtime Mini	–	3.42	1.16	1.17	0.71	0.40	3.90	3.86	2.91
Gemini 3.1 Flash-Lite	–	**3.75**	**1.27**	**1.30**	**0.80**	**0.42**	**3.92**	**4.20**	**3.39**

neck lies in generating grounded, goal-directed guidance rather than in perceiving the scene. Many errors are unsafe rather than simply unhelpful: the judge flags 10.2% of open-source answers as containing an error that could cause harm if followed, reaching 21.9% on hazard detection and 17.6% on navigation. GPT-Realtime Mini (3.42 overall, 9.5% critical) and Gemini 3.1 Flash-Lite reduce this without removing it: Gemini 3.1 Flash-Lite is the highest-scoring model overall (3.75) yet still commits a critical error on 7.4% of items, and on 11.6% of navigation items. Among the twelve open-source models, higher overall scores do not consistently correspond to lower judge-flagged critical-error rates: InternVL3-14B scores highest (3.33) yet errs critically more often (9.4%) than InternVL3-9B (8.0%) and Qwen3-VL-2B (8.6%), which have lower overall scores. Appendix D reports failure statistics and each model's response and judge score for four representative items.

5.3 Multimodal Enhancement

We evaluate whether VISTA's synchronized sensor streams can improve answer quality using the multimodal architecture introduced in Figure 5. For each backbone, the language LoRA adapters (Hu et al., 2022) and the sensor-fusion network are trained on the VISTA training split with a per-example mean answer NLL, so that long reference answers do not dominate the loss. Training has two stages: stage 1 injects the eight bottleneck tokens, and stage 2 adds four direct readout tokens per stream (28 sensor tokens in total), beginning with a sensor-only warmup before joint adaptation. Each stream is dropped independently with probability 0.15 during training. Answers are generated with greedy decoding and scored on the test split with the same judge and rubric as the zero-shot evaluation; the pretrained rows correspond to the zero-shot results in Table 2.

Main results. Table 3 shows that VISTA improves the overall score of every backbone. The largest gains are for InternVL3-2B (2.75 to 3.54, **+28.7%**) and Qwen3-VL-4B (3.22 to 3.95, **+22.7%**), followed by Qwen3-VL-8B (**+18.7%**) and InternVL3-8B (**+17.7%**); for these four models, all rubric components and all three response forms improve. With VISTA, Qwen3-VL-4B obtains the highest score in our evaluation, above both the best zero-shot open-source model, InternVL3-14B (3.33), and the best proprietary model (3.75) in Table 2. Guidance improves the most, from 1.96 to 3.28 for InternVL3-2B and from 2.65 to 3.69 for Qwen3-VL-4B. In InternVL3 and Qwen3-VL, the combined adaptation and sensor-fusion procedure produces larger score gains for the smaller of

Table 3: **VISTA versus pretrained backbones.** Columns as in Table 2; VISTA rows are shaded, with the gain over the pretrained model in green. **Bold**/underline: best/second-best per column.

Model	Size	Method	Score	Rubric components				Response type		
				EG	AC	TC	ACom	Scene	Q&A	Guidance
InternVL3	2B	Pretrained	2.75	0.96	0.91	0.54	0.35	3.77	3.19	1.96
		VISTA (ours)	**3.54** (+28.7%)	1.24	1.20	0.68	0.42	3.82	3.78	3.28
	8B	Pretrained	3.16	1.09	1.07	0.64	0.38	3.72	3.48	2.69
		VISTA (ours)	**3.72** (+17.7%)	1.29	1.27	0.73	**0.44**	4.03	4.07	3.39
Qwen3-VL	4B	Pretrained	3.22	1.11	1.10	0.68	0.38	3.78	3.69	2.65
		VISTA (ours)	**3.95** (+22.7%)	**1.39**	**1.36**	**0.76**	**0.44**	**4.25**	**4.18**	3.69
	8B	Pretrained	3.31	1.15	1.13	0.69	0.39	3.76	3.69	2.85
		VISTA (ours)	**3.93** (+18.7%)	**1.39**	1.35	**0.76**	**0.44**	4.19	4.04	**3.74**
Gemma 3	12B	Pretrained	3.07	1.04	1.05	0.66	0.36	3.41	3.76	2.48
		VISTA (ours)	**3.24** (+5.5%)	1.08	1.10	0.64	0.43	3.00	3.30	3.32

the two evaluated backbones, and after adaptation Qwen3-VL-4B matches Qwen3-VL-8B (3.95 vs. 3.93).

Task-level trade-offs. Gemma 3-12B shows a smaller gain, from 3.07 to 3.24 (+5.5%). Its guidance score improves from 2.48 to 3.32 and accessible communication from 0.36 to 0.43, whereas scene description and question answering drop from 3.41 to 3.00 and from 3.76 to 3.30. The adapted Gemma 3 model therefore shifts toward guidance at the cost of descriptive and question-answering quality, unlike the InternVL3 and Qwen3-VL models, which improve on all three response forms. Guidance scores improve for every backbone.

6 Discussion and Future Work

Sensor fusion. The adapted models gain most on guidance, where the answer depends on the user's position and motion, and smaller backbones benefit more than larger ones. These gains combine sensor fusion with task adaptation, and the reliability gates, learned only through the answer loss, do not measure each stream's causal importance; isolating each sensor's contribution is future work.

Limitations. VISTA is small compared with general video datasets and was collected in one geographic region by sighted participants simulating assistive scenarios. Evaluation relies on a single pass of an automatic judge, which cannot fully measure real-world usefulness or safety. Our model also processes clips offline, with trajectories estimated from the full recording, so it does not yet support real-time assistance.

Future work. We plan to collect data with blind and low-vision participants in more locations, extend the model to audio, depth, and streaming inputs, and evaluate usefulness and safety with human raters before making deployment claims. Promising extensions that our baseline does not implement include a compact environment state tracking nearby objects, walkable regions, and user motion, and risk-aware generation that prioritizes safety-relevant information.

7 Conclusion

VISTA introduces a multimodal egocentric benchmark for evaluating assistance to low-vision and blind users across ten task categories. Our evaluation reveals a persistent gap between scene understanding and assistance: current VLMs frequently fail to translate visual observations into grounded, actionable guidance. A baseline combining synchronized sensor inputs with task adaptation improves guidance scores across all evaluated backbones, demonstrating the promise of multimodal approaches for assistive tasks. These findings motivate models that connect perception of the user's surroundings with the actions needed to achieve their goals.

AI USE STATEMENT

In this work, we used generative AI tools for synthetic data generation and method implementation. Specifically, we used GPT-4o mini TTS to generate synthetic audio from text, and we used generative AI tools to assist in implementing the code for data generation and the pipeline. We have not used generative AI tools to help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, design or provide feedback on research methodology or experiments, assist with translation, clean or reformat datasets, support qualitative or thematic data analysis, or interpret results. Additionally, we used generative AI tools to edit the paper for readability, to format figure and table captions, and to revise the sensor-input panel of one figure. We have reviewed all AI-assisted work: the authors verified and tested all AI-assisted code before use, manually checked each generated audio file for quality and agreement with its source text, and reviewed every AI-suggested edit to the text and figure. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Leonard Bärmann and Alex Waibel. Where did I leave my keys? – episodic-memory-based question answering on egocentric videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 1560–1568, 2022. doi: 10.1109/CVPRW56347.2022.00162.

Nikhil Behari, Edwin Zhang, Yunfan Zhao, Aparna Taneja, Dheeraj Nagaraj, and Milind Tambe. A decision-language model (DLM) for dynamic restless multi-armed bandit tasks in public health. In *Advances in Neural Information Processing Systems*, volume 37, pp. 3964–4002, 2024.

Haoyu Chen, Kaichen Zhou, Hang Hua, Kaile Zhang, Jingwen Qian, Wufei Ma, Haonan Chen, Chunjiang Liu, Yizhou Zhao, Xiaoyuan Wang, et al. Memobench: Benchmarking world modeling in dynamically changing environments. *arXiv preprint arXiv:2606.27537*, 2026.

Sijie Cheng, Kechen Fang, Yangyang Yu, Sicheng Zhou, Bohao Li, Ye Tian, Tingguang Li, Lei Han, and Yang Liu. Videgothink: Assessing egocentric video understanding capabilities for embodied ai. *arXiv preprint arXiv:2410.11623*, 2024.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 720–736, 2018.

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

Arnav M. Das, Chi Ian Tang, Fahim Kawsar, and Mohammad Malekzadeh. PRIMUS: Pretraining IMU encoders with multimodal self-supervision. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2025. doi: 10.1109/ICASSP49660.2025.10888874.

Jakob Engel, Kiran Somasundaram, Michael Goesele, Albert Sun, Alexander Gamino, Andrew Turner, Arjang Talattof, Arnie Yuan, Bilal Souti, Brighid Meredith, et al. Project aria: A new tool for egocentric multi-modal ai research. *arXiv preprint arXiv:2308.13561*, 2023.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Google. Gemini 3.1 Flash-Lite. https://ai.google.dev/gemini-api/docs/models/gemini-3.1-flash-lite, 2026. Accessed: 2026-09-23.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6904–6913, 2017.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18995–19012, 2022.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19383–19400, 2024.

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

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.

Wenbo Hu, Yifan Xu, Yi Li, Weiyue Li, Zeyuan Chen, and Zhuowen Tu. Bliva: A simple multimodal llm for better handling of text-rich visual questions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 2256–2264, 2024.

Baoxiong Jia, Ting Lei, Song-Chun Zhu, and Siyuan Huang. EgoTaskQA: Understanding human tasks in egocentric videos. In *Advances in Neural Information Processing Systems*, volume 35, pp. 3343–3360, 2022. doi: 10.52202/068431-0242. Datasets and Benchmarks Track.

Junhyeok Kim, Jaewoo Park, Junhee Park, Sangeyl Lee, Jiwan Chung, Jisung Kim, Ji Hoon Joung, and Youngjae Yu. GuideDog: A real-world egocentric multimodal dataset for blind and low-vision accessibility-aware guidance. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5545–5574. Association for Computational Linguistics, July 2026. doi: 10.18653/v1/2026.acl-long.251. URL https://aclanthology.org/2026.acl-long.251/.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, pp. 19730–19742, 2023.

Weiyue Li, Minda Zhao, Weixuan Dong, Jiahui Cai, Yuze Wei, Michael Pocress, Yi Li, Wanyan Yuan, Xiaoyue Wang, Ruoyu Hou, Kaiyuan Lou, Wenqi Zeng, Yutong Yang, Yilun Du, and Mengyu Wang. Grading scale impact on LLM-as-a-judge: Human–LLM alignment is highest on 0–5 grading scale. *npj Artificial Intelligence*, 2026. doi: 10.1038/s44387-026-00153-8.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, 2004.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 740–755, 2014.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 34892–34916, 2023. doi: 10.52202/075280-1516.

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 (NeurIPS)*, volume 36, pp. 46212–46244, 2023.

Mukhriddin Mukhiddinov and Jinsoo Cho. Smart glass system using deep learning for the blind and visually impaired. *Electronics*, 10(22):2756, 2021.

Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion. In *Advances in Neural Information Processing Systems*, volume 34, pp. 14200–14213, 2021.

Gabriel Iluebe Okolo, Turke Althobaiti, and Naeem Ramzan. Assistive systems for visually impaired persons: Challenges and opportunities for navigation assistance. *Sensors*, 24(11):3572, 2024. doi: 10.3390/s24113572.

OpenAI. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

OpenAI. GPT-Realtime Mini. https://developers.openai.com/api/docs/models/gpt-realtime-mini, 2025. Accessed: 2026-09-23.

OpenAI. Introducing GPT-6 Sol and Luna. https://openai.com/index/introducing-gpt-6-sol-and-luna/, 2026. System card: https://deploymentsafety.openai.com/gpt-6-astra. Accessed September 2026.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 311–318, 2002.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pp. 8748–8763, 2021.

Nikhil Raina, Guruprasad Somasundaram, Kang Zheng, Sagar Miglani, Steve Saarinen, Jeff Meissner, Mark Schwesinger, Luis Pesqueira, Ishita Prasad, Edward Miller, Prince Gupta, Mingfei Yan, Richard Newcombe, Carl Ren, and Omkar M. Parkhi. EgoBlur: Responsible innovation in Aria. *arXiv preprint arXiv:2308.13093*, 2023.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1410.

Gunnar A. Sigurdsson, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, and Karteek Alahari. Charades-ego: A large-scale dataset of paired third and first person videos. *arXiv preprint arXiv:1804.09626*, 2018.

Inpyo Song, Minjun Joo, Joonhyung Kwon, and Jangwon Lee. Video question answering for people with visual impairments using an egocentric 360-degree camera. *arXiv preprint arXiv:2405.19794*, 2024. Presented at the CVPR 2024 EgoVis Workshop.

Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. Label Studio: Data labeling software. https://github.com/HumanSignal/label-studio, 2020. Open source software.

Benita Wong, Joya Chen, You Wu, Stan Weixian Lei, Dongxing Mao, Difei Gao, and Mike Zheng Shou. AssistQ: Affordance-centric question-driven task completion for egocentric assistant. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 485–501, 2022.

World Health Organization. World report on vision. https://www.who.int/publications/i/item/9789241516570, 2019.

Junbin Xiao, Nanxin Huang, Hao Qiu, Zhulin Tao, Xun Yang, Richang Hong, Meng Wang, and Angela Yao. EgoBlind: Towards egocentric visual assistance for the blind. In *Advances in Neural Information Processing Systems*, volume 38, pp. 29772–29806, 2025. doi: 10.52202/085713-0886. Datasets and Benchmarks Track.

Hanrong Ye, Haotian Zhang, Erik Daxberger, Lin Chen, Zongyu Lin, Yanghao Li, Bowen Zhang, Haoxuan You, Dan Xu, Zhe Gan, Jiasen Lu, and Yinfei Yang. MMEgo: Towards building egocentric multimodal LLMs for video QA. In *International Conference on Learning Representations*, 2025.

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

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025a.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations (ICLR)*, 2020.

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.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, pp. 46595–46623, 2023. doi: 10.52202/075280-2020. Datasets and Benchmarks Track.

Sheng Zhou, Junbin Xiao, Qingyun Li, Yicong Li, Xun Yang, Dan Guo, Meng Wang, Tat-Seng Chua, and Angela Yao. EgoTextVQA: Towards egocentric scene-text aware video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3363–3373, June 2025.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, 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.