# WearerText: Benchmarking Wearer-Centered Scene Text Understanding in AI-Glasses Videos

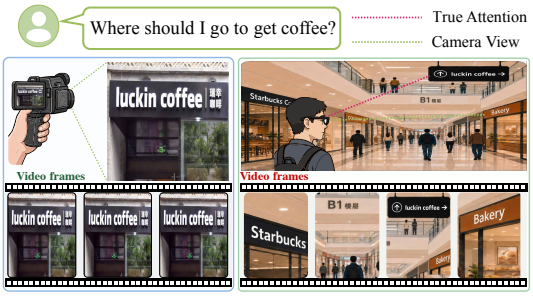
Anonymous ACL submission


## Abstract

Recent MLLMs have shown potential in scene-text-aware video understanding, but their ability in AI-glasses scenarios remains unclear. AI-glasses videos are passively captured without viewfinder feedback, so natural head motion often makes text off-center. The challenge is therefore not just reading text, but identifying the right text for the wearer's position, viewpoint trajectory, and goal. To systematically evaluate this capability, we introduce **WearerText**, a benchmark for **Wearer**-Centered Scene **Text** Understanding, comprising 1,101 real-world AI-glasses videos and 15,336 QA pairs spanning 13 tasks. The tasks are organized into a three-level diagnostic hierarchy: *L1 Text Perception*, *L2 Contextual Understanding*, and *L3 Wearer-Grounded Reasoning*, targeting unstable text reading, wearer-view alignment, and goal-oriented text reasoning, respectively. To ensure reliable construction and evaluation, we propose a multi-agent–hybrid verification pipeline for data construction, and a verified-evidence metric (VerEval) for open-ended answer evaluation. Benchmarking 20 recent MLLMs shows that current models still struggle with AI-glasses text understanding, especially on tasks requiring wearer-view alignment and goal-oriented reasoning.


## 1 Introduction

Recent multimodal large language models (MLLMs) have shown impressive progress in scene-text-aware video understanding (Feng et al., 2025; Shi et al., 2026; Zhou et al., 2025), yet their effectiveness on wearable assistive devices such as AI glasses remains largely underexplored. Closing this gap is critical for everyday applications such as shopping, transportation, and travel navigation, especially for improving accessibility for people with visual impairments (Waisberg et al., 2024; Xiao et al., 2025; Karakostas et al., 2024).




(a) Videos with Viewfinder Feedback (b) AI-glasses Videos without Viewfinder Feedback

Figure 1: Viewfinder feedback changes the nature of video text understanding. (a) With viewfinder feedback, the recorder can adjust the frame and keep target text stable. (b) AI glasses passively capture the wearer's camera view without such feedback; the camera view only approximates attention, so relevant text can be off-center, fragmented, or spatially ambiguous.

AI-glasses videos differ from videos recorded with a viewfinder, where the operator can actively frame the shot to keep relevant text stable and centered. In contrast, AI-glasses videos are passively captured from the wearer's first-person perspective, with no viewfinder feedback to guide composition. As illustrated in Figure 1, natural head movements can make text off-center and unstable. *The challenge is therefore not merely reading text, but reading the right text for the wearer's position, viewpoint trajectory, and goal.*

This wearer-centered setting demands three capabilities that existing text-aware video benchmarks rarely isolate as distinct evaluation axes (Table 1): *Unstable text perception* requires reading text that may be off-center, fragmented, or only briefly visible under natural head motion. *Wearer-view alignment* requires grounding the relevant text within the surrounding scene context, distinguishing it from competing signs according to the wearer's position and viewpoint. *Goal-oriented text reasoning* requires selecting, linking, and reasoning over fine-grained textual cues across frames according to the

Table 1: Comparison of WearerText with existing text-aware video benchmarks. **Video Source** indicates how the video data was obtained: web-sourced videos have unconstrained original capture conditions, while device-captured videos (dashcam, head-mounted camera, AI glasses) are passively recorded from a fixed first-person viewpoint without viewfinder framing. **Targeted Challenges**: ✓ indicates the benchmark *explicitly evaluates* this challenge as a first-class axis (dedicated task split or metric); ✗ indicates the phenomenon may occur in the data but is not isolated for evaluation. **Construction**: *Manual* = human-only annotation; *Hybrid* = LLM-assisted with human verification; **MAH-V** = our multi-agent–hybrid verification with factual and logical consistency checks. **Evaluation**: *String* = exact/F1 string match; *LLM-judge* = single-model semantic judgment; **VerEval** = our verified-evidence metric scoring factual grounding *and* reasoning consistency.

| Benchmark | Video Source | Targeted Challenges (Evaluation Axes) | | | Scale | | Construction | Evaluation |
|---|---|---|---|---|---|---|---|---|
| | | Unstable Text (Viewpoint) | Wearer-View Alignment | Goal-Oriented Text Evidence | # QA | Videos Number | | |
| *Web-sourced video (capture conditions unconstrained)* | | | | | | | | |
| M4ViTeVQA (Zhao et al., 2022) | Web video | ✗ | ✗ | ✗ | 2,103 | 7,620 | Manual | String |
| VideoOCR (Fei et al., 2025) | Web video | ✗ | ✗ | ✗ | 2,961 | 1,028 | Manual | String |
| MME-VideoOCR (Shi et al., 2026) | Web video | ✗ | ✗ | ✗ | 2,000 | 1,464 | Manual | String |
| VidText (Yang et al., 2025b) | Web video | ✗ | ✗ | ✗ | 2,857 | 939 | Manual | String |
| *Real-world device capture (passive, no viewfinder feedback)* | | | | | | | | |
| RoadTextVQA (Tom et al., 2023) | Dashcam (driving) | ✗ | ✗ | ✓ | 1,052 | 3,222 | Manual | String |
| EgoTextVQA (Zhou et al., 2025) | Head-mounted (driving, home) | ✓ | ✗ | ✓ | 7,064 | 1,507 | Manual | LLM-judge |
| **WearerText (Ours)** | **AI glasses (shopping, tourism, transit)** | ✓ | ✓ | ✓ | **15,336** | **1,101** | **MAH-V** | **VerEval** |

wearer's current goal and viewpoint trajectory. We therefore design our benchmark as a three-level diagnostic hierarchy that maps directly onto these requirements.

Concretely, we introduce **WearerText**, a benchmark for *Wearer-Centered Scene Text Understanding* in real-world AI-glasses videos (also referred to as *AI-glasses text understanding*), consisting of 1,101 videos and 15,336 QA pairs spanning 13 tasks across three levels: *L1 Text Perception* targets unstable text recognition and localization; *L2 Contextual Understanding* evaluates wearer-view alignment of relevant text in the surrounding scene; and *L3 Wearer-Grounded Reasoning* tests reasoning over viewpoint trajectory and temporal text evidence. For QA construction we develop MAH-V, a multi-agent–hybrid verification pipeline; for open-ended evaluation we propose VerEval, a verified-evidence metric covering factual grounding and reasoning consistency.

Benchmarking 20 recent MLLMs reveals that AI-glasses text understanding is far from solved. While models handle basic perception reasonably well, performance degrades sharply on tasks that demand goal-oriented text evidence reasoning, wearer-view alignment, and viewpoint-trajectory-aware reasoning. This setting is not a simple extension of Video OCR (Shi et al., 2026), as it requires reliable wearer-trajectory-aware reasoning. Our contributions are as follows:

- We formulate AI-glasses text understanding as a *wearer-centered* problem that couples unstable text reading, wearer-view alignment, and goal-oriented text reasoning.
- We construct WearerText, a benchmark of 1,101 real-world AI-glasses videos and 15,336 QA pairs across 13 tasks in a three-level diagnostic hierarchy.
- We propose MAH-V, a closed-loop Generate → Verify → Refine pipeline that improves data generation through multi-agent hybrid checks and human review, achieving $\kappa = 0.89$ inter-annotator agreement.
- We benchmark 20 MLLMs and find that even the best models score only 46.2 on L3.3 and 38.9 on L3.6, revealing large gaps in wearer-centered AI-glasses video understanding.

## 2 Related Work

**Egocentric Video Understanding.** Egocentric vision has received growing attention due to its promise for real-world AI assistance. Large-scale efforts such as HD-EPIC (Perrett et al., 2025), EmbodiedBench (Yang et al., 2025a), Ego4D (Grauman et al., 2022) and Project Aria (Engel et al., 2023) provide long-form first-person videos with rich annotations for activity understanding and human–object interaction. These corpora are valuable for generic egocentric video understanding, but they do not specifically target scene-text understanding, wearer-view alignment, or goal-oriented

Table 2: Hierarchy of Wearer-Centered Scene Text Understanding Tasks and Examples in WearerText.

| Task ID | Task Name | Task Definition | Examples |
|---|---|---|---|
| **Level 1: Text Perception.** *Can the model* SEE *and* LOCALIZE *unstable visible text in dynamic AI-glasses videos?* | | | |
| L1.1 | Dynamic Text Recognition (DTR) | Transcribe scene text from unstable egocentric views with head motion and viewpoint changes. | "I was walking fast and missed the phone number on that red sign. Can you tell me what it was?" |
| L1.2 | Dynamic Text Localization (DTL) | Locate queried text regions in dynamic frames and describe where they appear in the observed scene. | "Where is the floor and unit number located on the billboard?" |
| **Level 2: Contextual Understanding.** *Can the model* GROUND *relevant text in the wearer's surrounding context?* | | | |
| L2.1 | Query-Guided Text Extraction (QTE) | Select and extract relevant information from dense environmental text according to the user's query. | "How many years has this 'Dianmao' restaurant been specializing in Peking Duck?" |
| L2.2 | Confirmation (CF) | Verify the user's prior belief or memory using current visual-textual evidence. | "The sign says 'Floor 4'. I'm in the mall now; this shop is on the 4th floor, right?" |
| L2.3 | Cross-Lingual Understanding (CLU) | Interpret abbreviations, symbols, pinyin, or multilingual identifiers within the physical context. | "What does 'DIANMAO' at the top of the sign represent in this context?" |
| L2.4 | Situated Semantic Interpretation (SSI) | Interpret the semantic meaning of slogans, signs, or brand descriptions beyond literal recognition. | "What does the slogan 'Unmatched in Beijing, Renowned Everywhere' imply about this brand?" |
| L2.5 | Wearer-Centered Situational Inference (WSI) | Infer scene-level attributes or venue suitability by grounding text in the wearer's current situation. | "Based on the style and the '20-year' claim, is this a good place to treat friends to authentic local food?" |
| **Level 3: Wearer-Grounded Reasoning.** *Can the model* REASON *and* ASSIST *according to viewpoint trajectories and user goals?* | | | |
| L3.1 | History-Conditioned Intent Inference (HII) | Infer the wearer's current intent by combining previous scene-text observations with the current view. | "I'm tempted by the duck but have a meeting soon. Are there any 'fast-eating' sets mentioned?" |
| L3.2 | Action-Trajectory Intent Prediction (AIP) | Interpret the wearer's view trajectory to predict latent goals or possible next actions. | "I looked at the coffee shop sign and then the arrow sign. Am I trying to find the coffee shop?" |
| L3.3 | Structured Text Synthesis (STS) | Integrate structured textual sources, such as menus, or floor boards, while preserving their layout logic. | "Translate the entire list of stores on this directory into English for me." |
| L3.4 | Spatio-Temporal Navigation Reasoning (STN) | Provide action guidance by integrating current location, target text, and movement cues over time. | "I see the 'Floor 4' ad now. What is the fastest way for me to get to the actual store from here?" |
| L3.5 | Goal-Oriented Comparison (GOC) | Compare text-described options across the video and select the one that matches the wearer's goal. | "Considering both 'Professionalism' and 'Price', is this shop a better choice than a fast-food outlet?" |
| L3.6 | Viewpoint Trajectory Description (VTD) | Describe how the wearer's recorded viewpoint evolves over time. | "Describe how my viewpoint changed." |

text evidence reasoning, which are essential for wearer-centered guidance.

**Text-Centric Video Understanding.** Text in videos carries explicit and structured semantics, and several benchmarks have been proposed to evaluate text-aware video understanding (Zhao et al., 2022; Wu et al., 2021; Reddy et al., 2020; Zhou et al., 2025; Wu et al., 2023). Road-TextVQA (Reddy et al., 2020) focuses on driving scenes with vehicle-mounted viewpoints, while M4-ViTeVQA (Zhao et al., 2022) collects diverse YouTube videos to study general video text understanding. More recent benchmarks such as MME-VideoOCR (Shi et al., 2026) and VidText (Yang et al., 2025b) further evaluate OCR-related abilities in videos. EgoTextVQA (Zhou et al., 2025) moves closer to egocentric assistance by using head-mounted videos from driving and indoor activities. However, these benchmarks do not systematically isolate the key challenges of AI-glasses text understanding caused by natural head motion, wearer-view alignment, and goal-oriented text evidence reasoning. In addition, most existing datasets rely on manual construction and standard string matching or LLM-judge evaluation, which may limit scalability and fine-grained diagnosis.

Table 1 compares WearerText with existing text-aware and egocentric video benchmarks. Unlike prior datasets, WearerText is built from real-world AI-glasses videos and is designed around a three-level diagnostic hierarchy that directly targets unstable text reading, wearer-view alignment, and goal-oriented text evidence reasoning.

## 3 WearerText

In this section, we present the design rationale and construction pipeline for WearerText, which supports scalable construction and verifiable annotation for AI-glasses text understanding.

### 3.1 Hierarchical Task Design

To systematically evaluate AI-glasses text understanding, we organize WearerText into a three-level hierarchy with 13 tasks (Table 2). The hierarchy

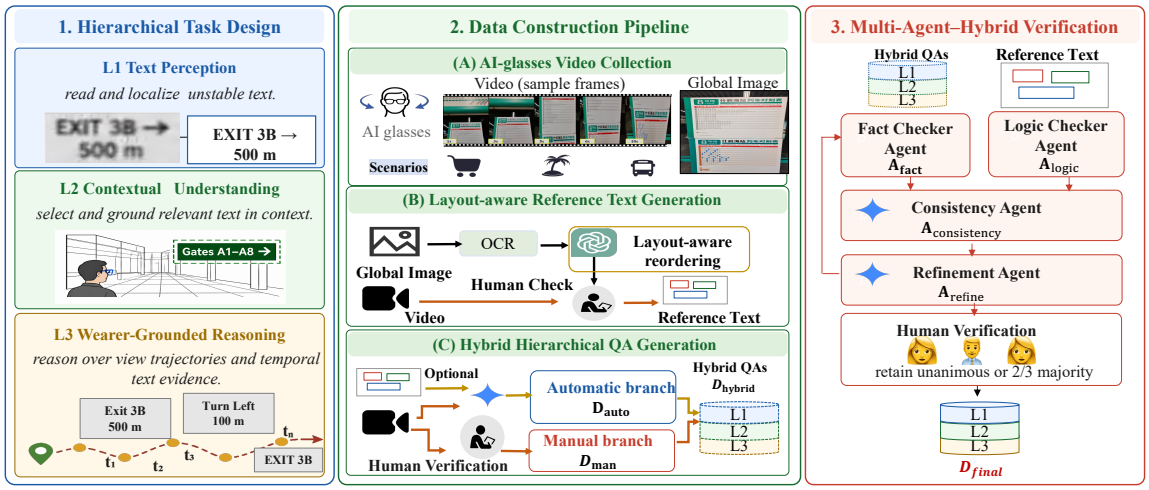



Figure 2: Overview of the WearerText construction pipeline. We first define a three-level hierarchy for AI-glasses text understanding, then collect videos and generate layout-aware reference text and hybrid QA pairs. The multi-agent–hybrid verification pipeline then performs data quality control and refinement before final human validation.

follows the key capabilities discussed above: unstable text reading, wearer-view alignment, and goal-oriented text evidence reasoning. **L1: Text Perception** evaluates whether models can recognize and localize unstable or briefly visible text in dynamic AI-glasses videos. **L2: Contextual Understanding** tests whether models can ground relevant text within the wearer's surrounding context, including dense signage, multilingual text, slogans, and scene-level semantics. **L3: Wearer-Grounded Reasoning** further examines whether models can reason over viewpoint trajectories and temporal text evidence to support goal-oriented assistance, such as navigation and viewpoint-motion understanding. Together, these levels provide a diagnostic framework for identifying where current MLLMs fail across perception, contextual grounding, and wearer-centered reasoning.

### 3.2 Data Construction Pipeline

To support scalable and grounded benchmark construction, we design an integrated pipeline that includes video collection, layout-aware reference text grounding, and hierarchical question–answer generation, as illustrated in Figure 2.

**AI-glasses Video Collection** To capture the structural complexity of real-world interactions, we utilize RayNeo AI Glasses to acquire a raw video dataset $\mathcal{V}$ consisting of 1,101 egocentric sequences. Each sequence, recorded at 30 FPS with a duration of 5–10 seconds, captures short interaction primitives such as approaching an object or scanning a menu. Our reasoning tasks therefore target goal-oriented selection of text evidence within these clips. For each video $v \in \mathcal{V}$, we simultaneously record a corresponding high-resolution global image $I_g$ to serve as a stable spatial reference for ground-truth text extraction. Detailed scenario coverage, sub-domains, and wearable-video condition statistics are provided in Appendix A and Table 5.

**Layout-aware reference text grounding.** Given the global image $I_g$, we construct a structured reference text $T_{\text{ref}}$ to serve as the factual anchor for subsequent question generation and verification. Specifically, we first apply OCR to obtain a raw token set $T_{\text{raw}} = \text{OCR}(I_g)$. To preserve the semantic reading order, we employ an LLM-based layout-aware transformation that reorganizes $T_{\text{raw}}$ into $T_{\text{layout}}$ using spatial coordinates $(x, y, w, h)$. Finally, a human verification is applied to produce a high-quality "gold standard" reference text $T_{\text{ref}}$.

**Hybrid hierarchical QA generation.** Based on $(v, T_{\text{ref}})$ pairs and the predefined three-level task hierarchy, we generate question–answer instances as candidate pairs $(q_i, a_i)$, where $q_i$ denotes a wearer-centered question and $a_i$ denotes its reference answer. We use a hybrid strategy, $\mathcal{D}_{\text{hybrid}} = \mathcal{D}_{\text{auto}} \cup \mathcal{D}_{\text{man}}$. The automatic branch $\mathcal{D}_{\text{auto}}$ leverages MLLMs for scalable synthesis, while the manual branch $\mathcal{D}_{\text{man}}$ provides expert-curated instances to complement automated generation with greater diversity and quality assurance. This hybrid design aims to balance large-scale coverage with reasoning-oriented question diversity.

### 3.3 MAH-V: Multi-Agent–Hybrid Verification

The reliability of task construction is critical for evaluating situated reasoning. We introduce the **M**ulti-**A**gent–**H**ybrid **V**erification (MAH-V) pipeline, which combines a committee of specialized agents $\mathcal{A} = \{A_{\text{fact}}, A_{\text{logic}}, A_{\text{consistency}}, A_{\text{refine}}\}$ with a final human-verification stage to filter $\mathcal{D}_{\text{hybrid}}$. The agent committee runs the multi-agent quality control stage shown in Figure 2, and its output is reviewed by expert annotators as described at the end of this section.

**Fact Checker Agent.** The Fact Checker $A_{\text{fact}}$ validates incorrect information by ensuring strict accuracy. For a candidate pair $(q_i, a_i) \in \mathcal{D}_{\text{hybrid}}$, $A_{\text{fact}}$ takes the video $v$ and reference text $T_{\text{ref}}$ together as inputs to compute a factuality score $S_{fact} \in [0, 1]$:

$$S_{fact} = A_{\text{fact}}(q_i, a_i, v, T_{\text{ref}}) \tag{1}$$

It verifies whether numerical values or entities in $a_i$ exist within the verified $T_{\text{ref}}$. Instances where $S_{fact} < \tau$ (threshold) are marked for improvement.

**Logic Checker Agent.** The Logic Checker $A_{\text{logic}}$ evaluates whether the reasoning process is grounded in the wearer-centered video context along three criteria. First, it assesses sequence necessity, determining whether temporal visual evidence is required to answer the question. Second, it performs indexical alignment by linking spatial references such as "the sign on the left" to specific video segments, ensuring that spatial descriptions are grounded in the wearer's viewpoint. Finally, it verifies temporal causality by checking that the textual cues supporting the answer have appeared in the observed video evidence, preventing reasoning that relies on unavailable information.

**Consistency Agent.** To improve the correctness of the generated data, the Consistency Agent $A_{\text{consistency}}$ performs cross validation. It detects contradictions between the factual evidence retrieved by $A_{\text{fact}}$ and the reasoning logic proposed by $A_{\text{logic}}$. For example, if $A_{\text{fact}}$ confirms a bus departs at "10:30" but $A_{\text{logic}}$ infers a wait time based on "10:35", $A_{\text{consistency}}$ identifies the mismatch and generates a conflict report $\mathcal{R}_{\text{conf}}$.

**Refinement Agent.** The Refinement Agent $A_{\text{refine}}$ serves as the final improvement step. Rather than simply discarding failed pairs, it uses the descriptive feedback and conflict reports $\{\mathcal{R}_{\text{conf}}, S_{fact}\}$ to improve the final QA pairs: $(q_i', a_i') = A_{\text{refine}}(q_i, a_i, \mathcal{R}_{\text{conf}}, S_{fact})$.

**Human Verification and Final Dataset.** After multi-agent verification, the refined QA pairs are passed to expert human annotators for final validation. Three domain experts independently review each sample, assessing both factual accuracy and logical coherence. Only samples achieving unanimous approval or majority consensus (2 out of 3) are retained in the final benchmark. This human-in-the-loop validation achieves an inter-annotator agreement of $\kappa = 0.89$, indicating high consistency in our quality assessment protocol.

### 3.4 VerEval: Verified-Evidence Evaluation

We propose **Ver**ified-Evidence **Eval**uation (**VerEval**), which adapts the MAH-V agent architecture for open-ended answer evaluation. While MAH-V uses OCR reference text for quality control during benchmark construction, VerEval directly compares model predictions with ground truth. VerEval uses three scoring agents. The *Fact Scorer* produces $S_{fact}$ by checking whether factual anchors such as numbers, timestamps, and entities match the visual evidence and ground truth. The *Logic Scorer* produces $S_{logic}$ by evaluating whether the answer's reasoning is compatible with spatio-temporal constraints in the video. The *Consistency Scorer* then produces $S_{consist}$ by jointly reviewing the fact and logic judgments, their justifications, and the model answer. Thus, $S_{consist}$ is the output of the consistency-scoring prompt, not a renamed factuality score.

We use $S_{fact}$ as an intermediate diagnostic signal, but exclude it from the final aggregate because its evidence is already checked by the consistency scorer and because the held-out human-alignment study showed lower discrimination for the raw fact score than for $S_{consist}$ and $S_{logic}$. The final VerEval score is therefore

$$S_{overall} = \lambda_1 \cdot S_{consist} + \lambda_2 \cdot S_{logic}. \tag{2}$$

In this paper, we set $\lambda_1 = 0.9$ and $\lambda_2 = 0.1$, selected on held-out human-judged samples; detailed weighting results are provided in Appendix D.2. We report the alignment between VerEval and human judgment in Figure 3, and provide all three scoring prompts in Appendix C.3.

## 4 Experiments

**Data Statistics.** We summarize dataset characteristics—scenario coverage, QA and video language distribution, and the test/train split—in Appendix Table 5, which illustrates the dataset's diversity across scenes and text settings.

Table 3: Overall VerEval scores. **Bold** indicates the best, Underline indicates the second best.

| Model | L1: Perception | | L2: Understanding | | | | | L3: Wearer-Grounded Reasoning | | | | | | Overall |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | L1.1 DTR | L1.2 DTL | L2.1 QTE | L2.2 CF | L2.3 CLU | L2.4 SSI | L2.5 WSI | L3.1 HII | L3.2 AIP | L3.3 STS | L3.4 STN | L3.5 GOC | L3.6 VTD | AVG |
| *Closed-source Models* | | | | | | | | | | | | | | |
| GPT-5.2 | 84.5 | **77.7** | 72.4 | 74.2 | **85.8** | **73.0** | **82.4** | **81.2** | **82.6** | 45.1 | **73.6** | **83.2** | 34.4 | **73.1** |
| Gemini-3-Pro | **91.4** | 75.8 | **82.7** | **82.8** | 83.2 | 70.2 | 81.6 | 77.1 | 79.5 | **46.2** | 73.1 | 80.6 | 23.0 | 72.9 |
| Claude-Opus-4.5 | 86.9 | 71.6 | 76.2 | 77.4 | 72.3 | 60.5 | 80.1 | 75.2 | 79.2 | 44.8 | 71.3 | 75.7 | 35.9 | 69.8 |
| Qwen3-VL-PLUS | 89.1 | 76.7 | 73.3 | 74.0 | 79.6 | 66.4 | 82.4 | 73.9 | 78.1 | 43.8 | 70.4 | 78.6 | **38.9** | 71.2 |
| Seed-1.6 | 84.0 | 70.4 | 71.8 | 75.0 | 61.1 | 63.3 | 60.2 | 58.6 | 54.3 | 46.0 | 55.4 | 57.6 | 35.1 | 61.0 |
| *Open-source Models* | | | | | | | | | | | | | | |
| InternVL3.5-8B | 54.4 | 49.6 | 43.9 | 50.4 | 57.0 | 28.1 | 49.6 | 40.4 | 25.5 | 7.9 | 43.3 | 35.3 | 21.4 | 39.0 |
| InternVL3.5-38B | 68.5 | 54.8 | 38.3 | 53.6 | 54.2 | 32.8 | 50.5 | 42.3 | 36.3 | 16.2 | 42.6 | 44.3 | 11.2 | 42.0 |
| Qwen2.5-VL-3B | 65.9 | 52.1 | 48.2 | 59.5 | 61.8 | 29.3 | 51.6 | 49.0 | 47.9 | 15.2 | 41.0 | 44.1 | 7.5 | 44.1 |
| Qwen2.5-VL-7B | 69.0 | 55.8 | 58.4 | 59.5 | 73.3 | 43.3 | 69.1 | 56.6 | 66.1 | 52.3 | 50.8 | 17.8 | 22.9 | 53.5 |
| Qwen2.5-VL-32B | 73.2 | 64.6 | 54.3 | 67.2 | 73.9 | 42.1 | 71.9 | 63.1 | 70.8 | 30.0 | 58.8 | 62.9 | 21.4 | 58.0 |
| Qwen3-VL-4B | 72.6 | 57.7 | 53.3 | 60.2 | 72.0 | 38.7 | 64.2 | 58.2 | 58.4 | 20.6 | 50.5 | 48.8 | 19.8 | 51.9 |
| Qwen3-VL-8B | 74.6 | 66.6 | 55.2 | 60.8 | 75.2 | 43.6 | 57.3 | 50.1 | 56.8 | 26.2 | 42.3 | 42.0 | 19.3 | 51.5 |
| Qwen3-VL-32B | 76.5 | 70.1 | 62.2 | 68.3 | 78.1 | 56.1 | 77.4 | 74.5 | 72.2 | 28.3 | 62.1 | 68.3 | 27.8 | 63.2 |
| GLM-4.6V-Flash-9B | 58.3 | 56.9 | 58.0 | 51.9 | 51.3 | 36.1 | 41.3 | 36.0 | 35.0 | 19.2 | 36.5 | 35.0 | 11.1 | 40.5 |
| GLM-4.6V-106B-A12B | 73.7 | 64.6 | 61.2 | 58.5 | 67.7 | 47.6 | 55.1 | 48.3 | 52.0 | 27.2 | 46.8 | 39.6 | 25.2 | 51.4 |
| MINICPM-V-4.5 | 76.5 | 63.3 | 62.7 | 69.6 | 46.6 | 48.4 | 60.5 | 53.5 | 50.6 | 22.2 | 46.2 | 46.0 | 9.5 | 50.4 |
| *Open-source Video Models* | | | | | | | | | | | | | | |
| VideoLLaMA3-7B | 51.1 | 55.0 | 47.1 | 52.8 | 53.8 | 28.5 | 49.0 | 43.1 | 50.0 | 13.9 | 39.7 | 42.7 | 4.9 | 40.9 |
| LLaVA-NeXT-Video | 8.6 | 18.3 | 10.6 | 18.3 | 32.9 | 7.5 | 21.6 | 17.0 | 19.8 | 2.6 | 15.1 | 21.5 | 2.1 | 15.1 |
| VideoR1-7B | 66.0 | 51.0 | 42.7 | 58.4 | 66.4 | 38.3 | 61.6 | 55.7 | 62.4 | 24.2 | 49.6 | 54.4 | 24.6 | 50.4 |
| MOLMO-2-8B | 27.0 | 51.0 | 30.8 | 48.6 | 63.0 | 22.1 | 43.7 | 34.5 | 46.9 | 11.3 | 35.7 | 35.8 | 11.9 | 35.6 |

**Model Evaluation.** We evaluate 5 closed-source API-based models (GPT-5.2 (OpenAI, 2025), Gemini-3-Pro (Google, 2025), Claude-Opus-4.5 (Anthropic, 2025), Qwen3-VL-PLUS (Bai et al., 2025a), Seed-1.6 (ByteDance, 2025)) and 15 advanced open-source MLLMs (InternVL3.5-8B/38B (Wang et al., 2025), Qwen2.5-VL-3B/7B/32B (Bai et al., 2025b), Qwen3-VL-4B/8B/32B (Bai et al., 2025a), GLM-4.6V-Flash-9B, GLM-4.6V-106B-A12B(Team et al., 2025), VideoLLaMA3-7B (Zhang et al., 2025), MINICPM-V-4.5 (Yu et al., 2025), LLaVA-NeXT-Video-7B (Zhang et al., 2024), VideoR1-7B (Feng et al., 2025), MOLMO-2-8B (Clark et al., 2026)).

### 4.1 Main Results

**Alignment Between VerEval and Human Judgment.** We evaluate VerEval on 390 QA pairs annotated by three experts and finalized by majority vote as *Correct* or *Wrong*. We compare against SAS, a transformer-based semantic answer similarity metric (Risch et al., 2021), and ANLS, a normalized edit-similarity metric widely used in scene-text VQA (Biten et al., 2019). As shown in Figure 3, our VerEval Score aligns more closely with human judgments. It yields a larger Correct–Wrong mean gap (0.402 vs. 0.112/0.094). These results indicate that VerEval provides a more reliable evaluation signal than ANLS and SAS for open-ended AI-glasses text understanding. We therefore report the main leaderboard with VerEval scores in Table 3, while detailed ANLS and SAS results are provided in Appendix Tables 6 and 7. All scores in our main results are reported on the raw $[0, 1]$ scale, and scores are multiplied by 100 for readability.

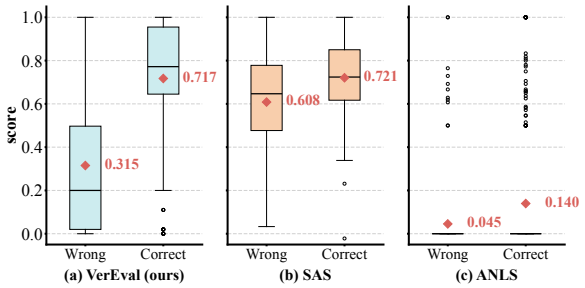



Figure 3: Distribution of automated metric scores grouped by human judgment (Wrong vs. Correct). (a) VerEval. (b) SAS. (c) ANLS. Red diamonds indicate mean values. The larger separation for VerEval indicates stronger alignment with human.

**Capability Boundaries of Mainstream MLLMs.** As detailed in Table 3, strong closed-source models such as GPT-5.2 and Gemini-3-Pro obtain the highest overall scores in our evaluation (73.1 and 72.9 respectively), but clear limitations remain in *Structured Text Synthesis (L3.3 STS)* and *Viewpoint Trajectory Description (L3.6 VTD)*. Specifically, the best scores in these two categories are 46.2 (Gemini-3-Pro) and 38.9 (Qwen3-VL-PLUS), indicating a substantial drop

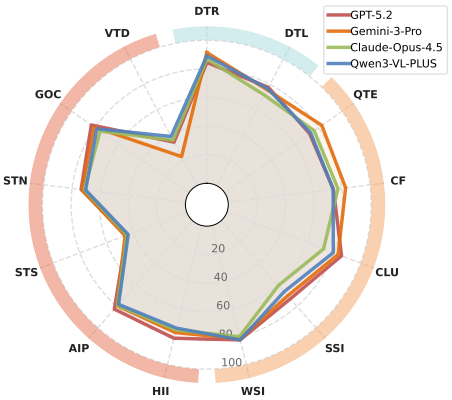



Figure 4: Capability Boundaries of Mainstream MLLMs.

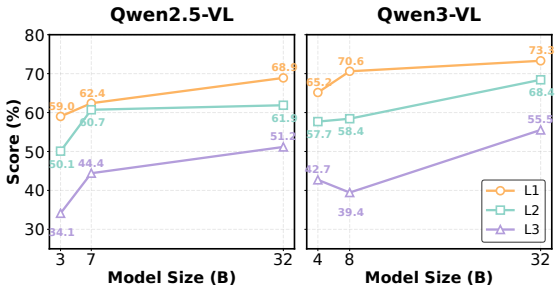



Figure 5: Performance across model sizes and task levels. Across both Qwen2.5-VL and Qwen3-VL families, L1 remains easier than L2, and L3 is consistently the most challenging, indicating that the three-level hierarchy preserves a clear difficulty separation across scales.

Table 4: Effectiveness of Scalable Verified Data Generation. **Bold** indicates the best performance; underline indicates the second best. ↑ and ↓ denote improvement or decline relative to the v0 baseline.

| Metric | v0 Zero-shot | v1 Unverified LLM | v2 MAH-V |
|---|---|---|---|
| **L1** | 62.4 | 65.0 ↑2.6 | **66.1** ↑3.7 |
| **L2** | 60.7 | 59.5 ↓1.2 | **62.5** ↑1.8 |
| **L3** | 56.5 | **59.0** ↑2.5 | 57.0 ↑0.5 |
| **Overall** | 59.5 | 60.3 ↑0.9 | **61.4** ↑1.9 |

relative to *Dynamic Text Recognition (L1.1 DTR)*. The results suggest that moving from unstable text perception to goal-oriented text evidence selection remains challenging for current MLLMs.

**Effectiveness of Training with Scalable Verified Data.** To test whether MAH-V data is useful beyond benchmark construction, we fine-tune Qwen2.5-VL-7B on (v1) unverified LLM-generated data and (v2) MAH-V-verified data—both drawn from videos disjoint from the test split—and compare against the zero-shot baseline (v0). Table 4 shows that v1 gains on L1/L3 but loses 1.2% on L2 *Contextual Understanding*, whereas v2 improves every level (+3.7% L1, +1.8% L2, +0.5% L3). Multi-agent verification therefore both raises overall data quality and removes the L2 degradation introduced by direct LLM synthesis.

### 4.2 Key Findings

We conduct extensive experiments to thoroughly analyze the challenges presented by WearerText. Our analysis centers on the following questions:

**Q1: How do tasks differ in difficulty?** As shown in Figure 5, across both Qwen2.5-VL and Qwen3-VL families, task difficulty increases from L1 to L3. L1 perception scores highest (59.0–73.1%), L2 understanding trails behind, and L3 reasoning remains most challenging. This clear separation ($L1 > L2 > L3$) confirms that task complexity grows with each level. Notably, the substantial performance gap between L1 and L3 (up to 20 points) suggests that current MLLMs struggle to move from reading unstable text to wearer-view alignment and goal-oriented text evidence selection.

**Q2: Why do current models fail?** To move beyond aggregate scores, we manually categorize GPT-5.2's incorrect predictions on the held-out test set. As shown in Figure 6(d), the dominant failure type is goal-oriented text evidence reasoning (57.4%), followed by unstable text recognition (26.6%) and wearer-view alignment (16.0%). These results suggest that the key bottleneck is not only OCR, but the ability to select and reason over the right textual evidence under the wearer's view. Figure 6 illustrates these failure modes. In (a), the model fails to recognize briefly visible scrolling text on a digital screen, showing the difficulty of unstable text reading under blur and short exposure. In (b), the model incorrectly includes years from visually similar awards, indicating weak temporal evidence accumulation and fine-grained text selection. In (c), the model gives an incorrect navigation suggestion because it fails to align visible shop signs with the wearer's spatial context. Overall, these cases show that AI-glasses text understanding requires the joint ability to read unstable text, ground it from the wearer's viewpoint, and accumulate goal-relevant evidence over time.

**Q3: Does the benchmark require full AI-glasses video?** We compare GPT-5.2 with full video, sampled keyframes, and layout-aware OCR text. As shown in Figure 7(a), full video consistently outperforms reduced inputs across L1–L3 and overall. The gap is modest for image-only input but

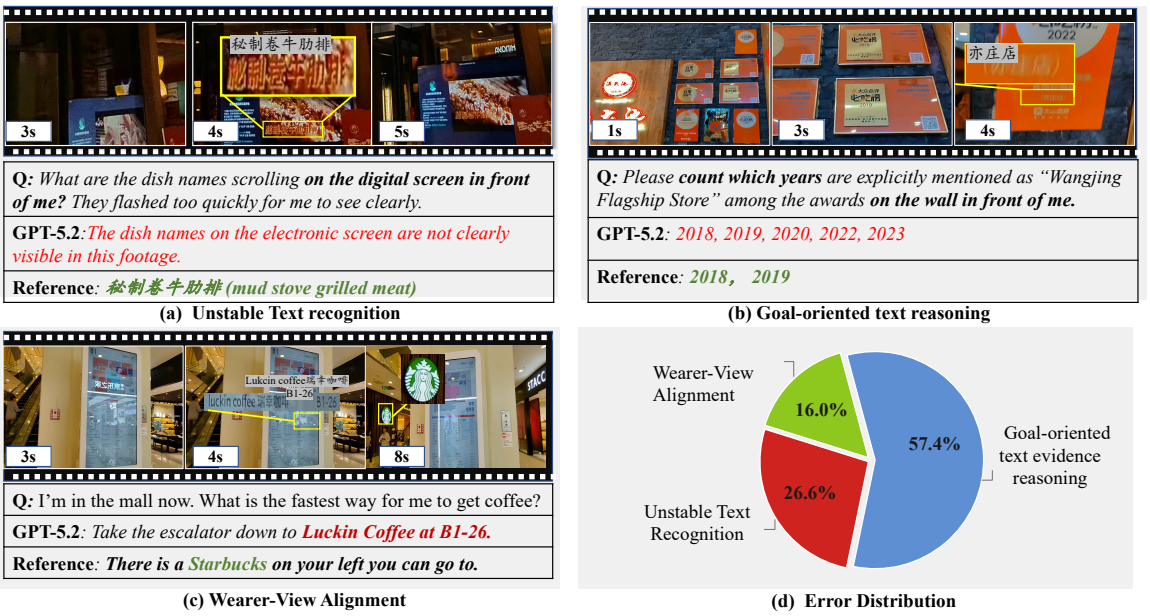



Figure 6: Representative failure cases and Error distribution on WearerText. Red text denotes model predictions, while green text denotes reference answers.

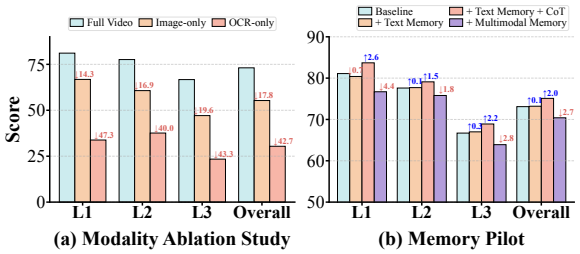



Figure 7: Diagnostic bar comparisons for GPT-5.2. (a) Modality ablation: full video, sampled keyframes, layout-aware OCR text. (b) Memory strategies: no memory, OCR text memory, structured text memory with reasoning, and raw multimodal memory. Arrows indicate changes relative to the full-video baseline.

much larger for OCR-only input, especially at L3, indicating that higher-level tasks require visual context and temporally ordered wearer-view signals beyond extracted text.

**Q4: Does memory help AI-glasses text understanding?** We compare four memory strategies on GPT-5.2 in Figure 7(b): no memory, unstructured multimodal memory with raw past frames, OCR-derived text memory with concatenated OCR tokens, and structured text memory with distilled per-frame text indexed by the wearer's viewpoint trajectory and paired with a reasoning prompt. Three task-specific patterns emerge. (i) **Memory helps when evidence is temporal.** Gains concentrate on L3 *Wearer-Grounded Reasoning* (e.g., L3.1 HII and L3.5 GOC), while frame-local L1 perception is largely unaffected. This suggests that the L3 deficit in Q1 partly reflects a memory-architecture limitation rather than model capacity alone. (ii) **Structure beats raw recall.** OCR-only memory brings little gain, whereas trajectory-indexed distilled text with a reasoning prompt improves overall performance by +2.0. It is also the only configuration that helps L3.4 STN and L3.6 VTD, both of which explicitly query the wearer's movement. (iii) **Raw multimodal memory hurts.** Stale frames degrade accuracy below the no-memory baseline, suggesting that re-perception competes with current-frame perception and amplifies the unstable-text errors. *Design implication:* future systems should maintain a low-bandwidth episodic record of distilled text indexed by the viewpoint trajectory rather than a verbatim multimodal buffer.

## 5 Conclusion

We introduce WearerText, a benchmark for evaluating AI-glasses text understanding in real-world wearable scenarios. Through our three-level task hierarchy, we assess 20 recent MLLMs on real-world public text-use challenges. Our findings have shown that current models remain limited on wearer-centered scene text understanding and reasoning, especially when answers require selecting relevant text evidence from partial observations according to the wearer's goal. These results suggest that progress in this setting requires stronger wearer-view alignment, structured memory over observed text, and intent-aware reasoning.

## Limitations

WearerTextis collected from real wearable-glasses videos, but its coverage is still bounded by the recording locations, participant behaviors, device characteristics, and public-space scenarios included in our study. Although the benchmark contains multilingual and mixed-language text, its distribution is not intended to represent all languages, regions, signage conventions, or accessibility needs. Future extensions should include broader geographic coverage, more diverse users, and additional low-resource languages.

Our task design focuses on offline question answering over short egocentric clips. Thus, "goal-oriented text evidence selection" refers to selecting relevant textual evidence within 5–10 second interaction primitives according to the wearer's current goal, rather than long-horizon memory over minutes or hours. This setting enables controlled comparison across models, but it does not fully capture interactive use cases in which a wearable assistant can ask clarification questions, adapt its guidance, or operate under real-time latency and energy constraints.

WearerTextis wearer-centered in the sense that questions are grounded in the camera-view trajectory of the person wearing the device. However, our RayNeo recordings do not include eye-tracking signals, so camera view should not be interpreted as true gaze. Future versions with synchronized gaze sensing could separate head direction, eye fixation, and user intention more precisely.

## Ethical Considerations

This work advances AI-glasses text understanding by providing a comprehensive and rigorous benchmark. All data contributors provided informed consent. To ensure privacy, we implemented rigorous anonymization (facial blurring) for all third-party individuals. Our pipeline prioritizes transparency and bias mitigation throughout the collection and annotation process.

## References


Anthropic. 2025. Introducing claude opus 4.5.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, and 1 others. 2025a. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

Shuai Bai, Keqin Chen, Xuejing Liu, and 1 others. 2025b. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Ali Furkan Biten, Ruben Tito, Andres Mafla, and 1 others. 2019. Scene text visual question answering. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4291–4301.

ByteDance. 2025. doubao-seed-1.6. Accessed: 2025-07-20.

Christopher Clark, Jieyu Zhang, Zixian Ma, and 1 others. 2026. Molmo2: Open weights and data for vision-language models with video understanding and grounding. *arXiv preprint arXiv:2601.10611*.

Cheng Cui, Ting Sun, Manhui Lin, and 1 others. 2025. Paddleocr 3.0 technical report. *Preprint*, arXiv:2507.05595.

Jakob Engel, Kiran Somasundaram, Michael Goesele, Albert Sun, Alexander Gamino, Andrew Turner, Arjang Talattof, Arnie Yuan, Bilal Souti, Brighid Meredith, and 1 others. 2023. Project aria: A new tool for egocentric multi-modal ai research. *arXiv preprint arXiv:2308.13561*.

Yulin Fei, Yuhui Gao, Xingyuan Xian, and 1 others. 2025. Do current video llms have strong ocr abilities? a preliminary study. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9860–9876.

Kaituo Feng, Kaixiong Gong, Bohao Li, and 1 others. 2025. Video-r1: Reinforcing video reasoning in mllms. *arXiv preprint arXiv:2503.21776*.

Google. 2025. A new era of intelligence with gemini 3.

Kristen Grauman, Andrew Westbury, Eugene Byrne, and 1 others. 2022. Ego4d: Around the world in 3,000 hours of egocentric video. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18973–18990.

Iason Karakostas, Aikaterini Valakou, Despoina Gavgiotaki, and 1 others. 2024. A real-time wearable ar system for egocentric vision on the edge. *Virtual Reality*, 28(1):44.

OpenAI. 2025. Introducing gpt-5.2.

OpenAI. 2025. Introducing OpenAI o3 and o4-mini. `https://openai.com/index/introducing-o3-and-o4-mini/`. Accessed 2025.

Toby Perrett, Ahmad Darkhalil, Saptarshi Sinha, Omar Emara, Sam Pollard, Kranti Kumar Parida, Kaiting Liu, Prajwal Gatti, Siddhant Bansal, Kevin Flanagan, and 1 others. 2025. Hd-epic: A highly-detailed egocentric video dataset. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 23901–23913.

Sangeeth Reddy, Minesh Mathew, Lluis Gomez, Marçal Rusinol, Dimosthenis Karatzas, and CV Jawahar. 2020. Roadtext-1k: Text detection & recognition dataset for driving videos. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11074–11080. IEEE.

Julian Risch, Timo Möller, Julian Gutsch, and Malte Pietsch. 2021. Semantic answer similarity for evaluating question answering models. In *Proceedings of the 3rd Workshop on Machine Reading for Question Answering*, pages 149–157.

Yang Shi, Huanqian Wang, Xie Xie, Huanyao Zhang, Lijie Zhao, Xinfeng Li, Chaoyou Fu, Zhuoer Wen, Wenting Liu, Zhuoran Zhang, and 1 others. 2026. Mme-videoocr: Evaluating ocr-based capabilities of multimodal llms in video scenarios. *Advances in Neural Information Processing Systems*, 38.

V Team, Wenyi Hong, Wenmeng Yu, and 1 others. 2025. Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *Preprint*, arXiv:2507.01006.

George Tom, Minesh Mathew, Sergi Garcia-Bordils, and 1 others. 2023. Reading between the lanes: Text videoqa on the road. In *International Conference on Document Analysis and Recognition*, pages 137–154. Springer.

Ethan Waisberg, Joshua Ong, Mouayad Masalkhi, Nasif Zaman, Prithul Sarker, Andrew G Lee, and Alireza Tavakkoli. 2024. Meta smart glasses—large language models and the future for assistive glasses for individuals with vision impairments. *Eye*, 38(6):1036–1038.

Weiyun Wang, Zhangwei Gao, Lixin Gu, and 1 others. 2025. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*.

Weijia Wu, Yuanqiang Cai, Debing Zhang, and 1 others. 2021. A bilingual, openworld video text dataset and end-to-end video text spotter with transformer. *arXiv preprint arXiv:2112.04888*.

Weijia Wu, Yuzhong Zhao, Zhuang Li, and 1 others. 2023. Icdar 2023 competition on video text reading for dense and small text. In *International Conference on Document Analysis and Recognition*, pages 405–419. Springer.

Junbin Xiao, Nanxin Huang, Hao Qiu, and 1 others. 2025. Egoblind: Towards egocentric visual assistance for the blind people. *arXiv preprint arXiv:2503.08221*.

Rui Yang, Hanyang Chen, Junyu Zhang, and 1 others. 2025a. Embodiedbench: Comprehensive benchmarking multi-modal large language models for vision-driven embodied agents. *arXiv preprint arXiv:2502.09560*.

Zhoufaran Yang, Yan Shu, Zhifei Yang, and 1 others. 2025b. Vidtext: Towards comprehensive evaluation for video text understanding. *arXiv preprint arXiv:2505.22810*.

Tianyu Yu, Zefan Wang, Chongyi Wang, Fuwei Huang, Wenshuo Ma, Zhihui He, Tianchi Cai, Weize Chen, Yuxiang Huang, Yuanqian Zhao, Bokai Xu, Junbo Cui, Yingjing Xu, Liqing Ruan, Luoyuan Zhang, Hanyu Liu, Jingkun Tang, Hongyuan Liu, Qining Guo, and 15 others. 2025. Minicpm-v 4.5: Cooking efficient mllms via architecture, data, and training recipe. *Preprint*, arXiv:2509.18154.

Boqiang Zhang, Kehan Li, Zesen Cheng, and 1 others. 2025. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*.

Yuanhan Zhang, Bo Li, haotian Liu, and 1 others. 2024. Llava-next: A strong zero-shot video understanding model.

Minyi Zhao, Bingjia Li, Jie Wang, and 1 others. 2022. Towards video text visual question answering: Benchmark and baseline. *Advances in Neural Information Processing Systems*, 35:35549–35562.

Sheng Zhou, Junbin Xiao, Qingyun Li, and 1 others. 2025. Egotextvqa: Towards egocentric scene-text aware video question answering. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3363–3373.

# A Dataset Statistics

We summarize the statistics of **WearerText** in Table 5. The dataset consists of 1,101 real-world egocentric videos collected in three text-rich scenarios—shopping centers, transportation hubs, and tourist destinations—and contains 15,336 question–answer pairs spanning diverse perceptual, contextual, and reasoning-oriented tasks.

The collection spans sub-domains such as restaurants, supermarkets, menus, subway and railway stations, airports, street scenes, parks, and museums. Common wearable-video conditions related to the challenges discussed in the main paper are intentionally preserved: 72.0% of videos contain viewpoint changes, 33.7% contain briefly visible text, and 4.8% contain visible motion blur. These conditions are not mutually exclusive, since a single video may contain multiple challenges. We view this coverage as diagnostic rather than exhaustive; broader geographic and linguistic expansion remains future work.

The benchmark primarily covers Chinese and English language settings. Most QA pairs are in Chinese (67.59%), while English and mixed Chinese-English QA pairs account for 16.40% and 15.19%, respectively. At the video level, 29.22% of the visible scene text is Chinese-only, while 69.15% contains mixed Chinese-English text, reflecting the multilingual nature of everyday embodied text environments.

We reserve 1,391 QA pairs as the held-out test split used for all zero-shot model evaluations and human-alignment analysis in the main paper. The remaining 13,945 QA pairs form a separate train split that is reserved for the MAH-V SFT ablation (Sec. 4.1) and is *not* used to tune the zero-shot baselines on the leaderboard; the fine-tuned Qwen2.5-VL-7B numbers therefore should not be compared directly with the zero-shot rows in Table 3.

Table 5: Statistics of **WearerText**. Wearable-video condition percentages are non-exclusive: one video may contain multiple conditions.

| **Statistic** | **Value** |
|---|---|
| Total Videos | 1,101 |
| Total Questions | 15,336 |
| **By Scenario** | |
| Shopping Centers | 378 videos |
| Transportation Hubs | 412 videos |
| Tourist Destinations | 311 videos |
| **Wearable-Video Conditions (non-exclusive)** | |
| Viewpoint Changes | 72.0% videos |
| Briefly Visible Text | 33.7% videos |
| Visible Motion Blur | 4.8% videos |
| **QA Language** | |
| Chinese | 67.59% |
| English | 16.40% |
| Mixed Chinese-English | 15.19% |
| Other | 0.82% |
| **Video Text Language** | |
| Chinese | 29.22% |
| Mixed Chinese-English | 69.15% |
| **Verification Split** | |
| Test Set | 1,391 QAs / 107 videos |
| Train Set | 13,945 QAs / 994 videos |

# B Multi-Faceted Per-Task Evaluation

To complement VerEval, we also report two surface-similarity metrics so that readers can compare WearerText models under traditional VQA scoring conventions: (1) **ANLS** (Average Normalized Levenshtein Similarity) measures exact string accuracy via normalized Levenshtein similarity (edit distance divided by the longer string length), shown in Table 6. (2) **SAS** (Semantic Answer Similarity) measures the semantic similarity between the predicted answer and the ground truth answer, so paraphrased but equivalent answers still score highly, shown in Table 7.

# C Implementation Details

## C.1 Data Construction Pipeline

The OCR tool used in this paper is PaddleOCR (Cui et al., 2025), and the layout-aware transformation, which reorganizes OCR tokens into reading-order reference text $T_{\text{ref}}$ conditioned on detected bounding boxes, is performed by o4-mini (OpenAI, 2025); this is a separate stage from scoring and does not interact with the judge. The Fact and Consistency Agents return continuous scores in $[0, 1]$; QA pairs whose Fact score falls below the acceptance threshold $\tau = 0.7$ or whose Consistency Agent emits a conflict report are routed to the Refinement Agent, which is allowed up to 2 revision passes before any still-failing pair is dropped. The retained pairs then proceed to the human-verification stage.

## C.2 VerEval Judge Setup

All VerEval scorers share a single fixed judge model Qwen3-VL-32B (Bai et al., 2025a). The judge receives, per QA pair, the model's predicted answer, the human-written reference answer, the verified reference text $T_{\text{ref}}$, and the original ques-

Table 6: ANLS Performance. **Bold** indicates the best, Underline indicates the second best. Task abbreviations: DTR (Dynamic Text Recognition), DTL (Dynamic Text Localization), QTE (Query-Guided Text Extraction), CF (Confirmation), CLU (Cross-Lingual Understanding), SSI (Situated Semantic Interpretation), WSI (Wearer-Centered Situational Inference), HII (History-Conditioned Intent Inference), AIP (Action-Trajectory Intent Prediction), STS (Structured Text Synthesis), STN (Spatio-Temporal Navigation Reasoning), GOC (Goal-Oriented Comparison), VTD (Viewpoint Trajectory Description).

| Model | L1: Perception | | L2: Understanding | | | | | L3: Wearer-Grounded Reasoning | | | | | | Overall |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | L1.1 DTR | L1.2 DTL | L2.1 QTE | L2.2 CF | L2.3 CLU | L2.4 SSI | L2.5 WSI | L3.1 HII | L3.2 AIP | L3.3 STS | L3.4 STN | L3.5 GOC | L3.6 VTD | AVG |
| *Closed-source Models* | | | | | | | | | | | | | | |
| GPT-5.2 | **62.4** | **19.7** | 20.3 | 8.6 | **16.4** | **23.0** | 15.9 | 0.5 | 0.0 | 0.6 | 0.0 | 0.0 | 0.0 | 12.9 |
| Gemini-3-Pro | 59.0 | 18.7 | **29.9** | **21.0** | 19.1 | 30.3 | **20.0** | **3.2** | **0.0** | **1.5** | **0.0** | **1.6** | 6.6 | **16.2** |
| Claude-Opus-4.5 | 68.8 | 20.6 | 27.9 | 9.8 | 6.5 | 18.3 | 19.6 | 0.0 | 0.0 | 0.0 | 0.0 | 0.5 | 0.0 | 13.2 |
| Qwen3-VL-PLUS | 73.1 | 20.0 | 29.9 | 13.9 | 10.5 | 21.4 | 18.2 | 0.5 | 0.0 | 0.5 | 0.0 | 0.0 | 0.0 | 14.5 |
| Seed-1.6 | 72.4 | 17.5 | 32.9 | 10.7 | 7.5 | 19.7 | 20.7 | 1.9 | 0.0 | 0.0 | 0.0 | 0.0 | **5.1** | 14.5 |
| *Open-source Models* | | | | | | | | | | | | | | |
| InternVL3.5-8B | 34.6 | 2.7 | 6.6 | 3.4 | 3.7 | 3.9 | 1.1 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.2 | 4.4 |
| InternVL3.5-38B | 51.2 | 16.9 | 10.5 | 1.6 | 2.7 | 7.7 | 7.2 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 14.9 | 8.7 |
| Qwen2.5-VL-3B | 43.9 | 7.0 | 7.8 | 5.4 | 3.4 | 6.1 | 6.4 | 0.7 | 0.0 | 0.5 | 0.0 | 0.0 | 22.9 | 8.0 |
| Qwen2.5-VL-7B | 50.4 | 3.8 | 13.3 | 5.9 | 9.3 | 10.8 | 5.0 | 0.6 | 0.0 | 0.0 | 0.0 | 0.0 | 20.0 | 9.1 |
| Qwen2.5-VL-32B | 61.1 | 12.2 | 15.0 | 4.6 | 8.4 | 9.6 | 11.6 | 0.5 | 0.0 | 0.0 | 0.0 | 0.0 | 12.7 | 10.4 |
| Qwen3-VL-4B | 55.3 | 20.7 | 13.0 | 4.0 | 9.8 | 11.0 | 8.5 | 0.6 | 0.0 | 0.5 | 0.0 | 0.0 | 9.9 | 10.3 |
| Qwen3-VL-8B | 61.1 | 16.4 | 20.2 | 2.8 | 8.9 | 14.1 | 11.2 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 10.4 |
| Qwen3-VL-32B | 59.7 | 15.1 | 16.0 | 4.4 | 9.2 | 11.6 | 12.7 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 2.4 | 10.1 |
| GLM-4.6V-Flash-9B | 51.4 | 15.7 | 19.9 | 1.6 | 7.8 | 10.7 | 8.9 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 8.9 |
| GLM-4.6V-106B-A12B | 62.1 | 22.1 | 21.8 | 8.3 | 9.7 | 17.8 | 10.7 | 0.5 | 0.0 | 0.0 | 0.0 | 0.0 | 12.1 | 12.7 |
| MINICPM-V-4.5 | 51.3 | 14.7 | 23.1 | 11.2 | 3.8 | 9.6 | 6.8 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 9.3 |
| *Open-source Video Models* | | | | | | | | | | | | | | |
| VideoLLaMA3-7B | 16.7 | 5.2 | 1.9 | 2.2 | 1.8 | 2.7 | 2.6 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 2.5 |
| LLaVA-NeXT-Video | 0.0 | 0.5 | 0.5 | 0.0 | 0.0 | 0.5 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.1 |
| VideoR1-7B | 1.5 | 1.2 | 2.4 | 0.0 | 1.1 | 1.7 | 0.5 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.8 | 0.7 |
| MOLMO-2-8B | 15.7 | 14.0 | 8.3 | 0.8 | 4.8 | 2.4 | 3.7 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 3.8 |

tion; it does not receive video frames, so its grounding is determined by the structured reference rather than re-perceiving the scene. We deliberately keep the judge text-only: keep, but ensure earlier mentions are consistent, and re-feeding noisy egocentric frames would introduce a second source of perception error that we cannot control across the 20 MLLMs we evaluate. VerEval therefore measures answer consistency with respect to the verified evidence record, not re-perception of the video; the per-model perception capability is captured separately by the L1 perception tasks in the leaderboard. Decoding uses temperature 0 and a 1,024-token cap, and each pair is scored once with a deterministic seed. Note that Qwen3-VL-32B also appears as one of the evaluated open-source MLLMs in Table 3; we discuss the resulting judge–evaluatee overlap and its implications in the Limitations section.

### C.3 VerEval Scoring Prompts

We employ three prompts for model evaluation in the VerEval framework. The Fact Scoring Prompt generates a diagnostic Fact Score ($S_{fact}$) by detecting discrepancies in key information (e.g., numbers, timestamps, entities) against video content and ground truth. The Logic Scoring Prompt assesses the structural integrity of the reasoning process and its alignment with spatio-temporal constraints, outputting the Logic Score ($S_{logic}$). The Consistency Scoring Prompt jointly reviews the fact judgment, logic judgment, their rationales, and the model answer, producing the Consistency Score ($S_{consist}$). In the final VerEval score, $S_{consist}$ refers specifically to the output of the consistency-scoring prompt in Figure 8c. It is not a renamed version of $S_{fact}$. We treat $S_{fact}$ as an intermediate diagnostic because the consistency scorer already checks factual evidence while also assessing whether the fact and logic judgments are mutually coherent. Empirically, raw fact scoring showed weaker separation from human judgments than the consistency and logic scores. Therefore, $S_{fact}$ is excluded from the aggregate, and $S_{overall}$ is computed only from $S_{consist}$ and $S_{logic}$. The full scoring prompt templates are shown in Figure 8.

# Fact Scoring
You are a rigorous expert in fact scoring. Please evaluate the accuracy of the answer to be assessed based on the ground truth and video.

## QA to be Scored
**Question**: *{question}*
**Ground Truth**: *{ground_truth}*
**Answer to be Evaluated**: *{answer}*

## Scoring Task Requirements
### 1. Fact Assessment (Core: Hallucination Detection)
Verify every word/number in the answer to check for:
- Information that contradicts the ground truth
- Information inconsistent with the video, especially significant text recognition errors
- Fabricated non-existent content
- Errors in key information such as numbers, names, times, locations

### 2. Completeness Assessment
Check for omissions of key information in the ground truth and important content in the video

## Scoring Criteria
- Total score = 1.0, minimum 0
- **Critical Error**: Factual contradiction with the ground truth or video in key information -> Deduct 0.4-0.8 points
- **Minor Error**: No critical errors, but secondary errors or minor deviations in supplementary statements -> Deduct 0.1-0.3 points
- **Incomplete Information**: Answers the question without factual contradictions but does not fully cover the ground truth and video -> Deduct 0.05-0.15 points
- **Perfect Match**: Fully consistent with the ground truth and video -> 1.0 points

## Output Format
Strictly output in the following JSON format: ```{ "score": "xxx", "reason": "xxx" } ```
Where "score" is a two-decimal-place score, and "reason" explains the factual errors or inconsistencies with the video in the answer to be evaluated. Please rigorously verify every specific piece of text and validate it with the video information!

(a) Fact scoring prompt.

# Logic Scoring
You are a logic scoring expert specializing in evaluating the logical rigor of question design and answers, and verifying them in conjunction with video.

## QA to be Scored
**Question**: *{question}*
**Ground Truth**: *{ground_truth}*
**Answer to be Evaluated**: *{answer}*

## Scoring Task Requirements
### 1. Evaluation of Logical Rigor in Answers
Quality of logical reasoning in the answer:
- Whether the reasoning process is reasonable and the conclusion is natural
- Whether the logical chain is complete and rigorous
- Whether there are logical leaps or contradictions

### 2. Accuracy of Question Understanding
Assess whether the core requirements of the question are accurately understood

### 3. Logical Consistency with video
Check whether the answer's logic aligns with the video:
- Whether the understanding of the video scenes is correct
- Whether the reasoning based on the video is reasonable
- Whether there are logical judgments that contradict the video

## Scoring Criteria
- Total score = 1.0, minimum 0
- **Logical Confusion**: The reasoning process contains serious logical errors or contradictions with the video -> Deduct 0.4-0.8 points
- **Misunderstanding**: Misinterpretation of the question, leading to an incorrect answer direction -> Deduct 0.4-0.8 points
- **Logical Laxity**: The reasoning process is mostly correct but lacks rigor or is poorly integrated with the video -> Deduct 0.1-0.3 points
- **Logical Rigor**: The reasoning process is reasonable, the conclusion is natural, and fully consistent with the video -> 1.0 points

## Output Format
Strictly adhere to the following JSON format for output: ```{ "score": "xxx", "reason": "xxx" } ```
Where "score" is a two-decimal-place score, and "reason" is the justification for the score, explaining the logical errors in the answer to be evaluated and any inconsistencies with the video. Please strictly follow the logical standards for evaluation and verify them in conjunction with the video information!

(b) Logic scoring prompt.

# Consistency Scoring
You are a comprehensive judgment expert, providing final judgments based on factual and logical scoring results.

## QA Information
**Question**: *{question}*
**Ground Truth**: *{ground_truth}*
**Answer to be Evaluated**: *{answer}*

## Factual Scoring Results
- **Fact Score**: *{fact_score}*
- **Fact Scoring Reason**: *{fact_reason}*

## Logical Scoring Results
- **Logic Score**: *{logic_score}*
- **Logic Scoring Reason**: *{logic_reason}*

## Scoring Criteria
- Total Score = 1.0, minimum 0

## Evaluation Task
Based on the above factual and logical scoring results and video, provide a final consistency judgment:
- Balance factual accuracy and logical rigor, considering both dimensions
- Evaluate the overall consistency between the answer and the video
- Assess the overall quality and consistency level of the answer
- Consider primary and secondary issues
- Provide a final comprehensive score

## Output Format
Strictly adhere to the following JSON format for output: ```{"score": "xxx","reason": "xxx"}```
Where "score" is a two-decimal-place score, and "reason" is a detailed scoring rationale, explained in conjunction with the video. Please provide a strict and fair judgment!

(c) Consistency scoring prompt.

Figure 8: Scoring prompt templates used in VerEval.

Table 7: SAS Performance. **Bold** indicates the best, Underline indicates the second best. Task abbreviations: DTR (Dynamic Text Recognition), DTL (Dynamic Text Localization), QTE (Query-Guided Text Extraction), CF (Confirmation), CLU (Cross-Lingual Understanding), SSI (Situated Semantic Interpretation), WSI (Wearer-Centered Situational Inference), HII (History-Conditioned Intent Inference), AIP (Action-Trajectory Intent Prediction), STS (Structured Text Synthesis), STN (Spatio-Temporal Navigation Reasoning), GOC (Goal-Oriented Comparison), VTD (Viewpoint Trajectory Description).

| Model | L1: Perception | | L2: Understanding | | | | | L3: Wearer-Grounded Reasoning | | | | | | Overall |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | L1.1 DTR | L1.2 DTL | L2.1 QTE | L2.2 CF | L2.3 CLU | L2.4 SSI | L2.5 WSI | L3.1 HII | L3.2 AIP | L3.3 STS | L3.4 STN | L3.5 GOC | L3.6 VTD | AVG |
| *Closed-source Models* | | | | | | | | | | | | | | |
| GPT-5.2 | 84.4 | **74.4** | 69.5 | **77.3** | **76.5** | **76.6** | **56.1** | **71.8** | **70.2** | **74.1** | **71.5** | **71.8** | **70.4** | 72.6 |
| Gemini-3-Pro | 84.1 | 75.2 | **77.0** | 81.7 | 77.8 | 78.7 | 64.7 | 76.4 | 71.4 | 76.9 | 76.4 | 76.7 | 72.4 | **76.1** |
| Claude-Opus-4.5 | 86.8 | 73.1 | 76.1 | 81.5 | 72.5 | 73.4 | 63.6 | 74.0 | 72.8 | 75.3 | 76.6 | 74.4 | 74.3 | 75.0 |
| Qwen3-VL-PLUS | **89.1** | 72.1 | 73.1 | 77.0 | 74.0 | 72.2 | 65.8 | 71.0 | 69.7 | 74.3 | 73.5 | 73.2 | 70.4 | 73.5 |
| Seed-1.6 | 87.9 | 70.0 | 73.1 | 76.5 | 67.7 | 72.3 | 65.9 | 67.3 | 63.9 | 70.2 | 67.3 | 67.0 | 66.3 | 70.4 |
| *Open-source Models* | | | | | | | | | | | | | | |
| InternVL3.5-8B | 66.6 | 59.6 | 54.4 | 62.9 | 54.9 | 54.5 | 26.5 | 60.7 | 50.1 | 61.6 | 60.9 | 58.7 | 55.3 | 55.9 |
| InternVL3.5-38B | 79.2 | 68.7 | 54.9 | 55.5 | 58.4 | 54.6 | 36.9 | 51.2 | 51.9 | 54.6 | 51.1 | 52.8 | 59.3 | 56.1 |
| Qwen2.5-VL-3B | 75.0 | 67.7 | 64.8 | 73.6 | 62.0 | 59.7 | 34.7 | 62.0 | 64.5 | 61.2 | 62.1 | 61.0 | 54.9 | 61.8 |
| Qwen2.5-VL-7B | 79.8 | 65.0 | 65.4 | 74.7 | 69.5 | 65.3 | 30.0 | 68.0 | 69.4 | 68.6 | 68.3 | 65.7 | 82.4 | 67.1 |
| Qwen2.5-VL-32B | 83.2 | 69.2 | 64.1 | 76.7 | 70.0 | 65.3 | 51.7 | 64.2 | 66.5 | 67.2 | 65.9 | 65.2 | 64.5 | 67.2 |
| Qwen3-VL-4B | 80.4 | 72.1 | 65.1 | 73.2 | 71.3 | 65.6 | 48.6 | 67.6 | 66.0 | 68.3 | 64.3 | 64.5 | 58.3 | 66.6 |
| Qwen3-VL-8B | 82.4 | 67.7 | 65.1 | 60.8 | 70.6 | 64.5 | 56.9 | 56.1 | 62.8 | 57.1 | 55.6 | 53.1 | 64.7 | 62.9 |
| Qwen3-VL-32B | 83.1 | 70.4 | 66.4 | 76.3 | 71.7 | 69.3 | 57.1 | 65.3 | 68.7 | 69.1 | 69.8 | 66.0 | 69.2 | 69.4 |
| GLM-4.6V-Flash-9B | 80.3 | 63.4 | 60.3 | 44.0 | 57.1 | 57.8 | 42.7 | 46.1 | 51.5 | 47.1 | 47.0 | 46.5 | 44.7 | 53.0 |
| GLM-4.6V-106B-A12B | 83.8 | 73.2 | 64.1 | 73.7 | 71.7 | 66.7 | 51.0 | 60.4 | 62.0 | 64.8 | 61.5 | 59.4 | 62.5 | 65.8 |
| MINICPM-V-4.5 | 78.0 | 64.0 | 64.6 | 74.6 | 52.9 | 61.9 | 45.9 | 61.6 | 59.2 | 61.9 | 60.5 | 57.7 | 42.3 | 60.4 |
| *Open-source Video Models* | | | | | | | | | | | | | | |
| VideoLLaMA3-7B | 58.4 | 59.9 | 61.3 | 67.9 | 61.1 | 55.5 | 35.6 | 60.9 | 61.8 | 65.3 | 64.3 | 62.1 | 42.5 | 58.2 |
| LLaVA-NeXT-Video | 39.6 | 56.4 | 46.5 | 60.7 | 56.0 | 48.1 | 26.5 | 54.6 | 58.3 | 57.3 | 60.3 | 57.6 | 22.7 | 49.6 |
| VideoR1-7B | 56.8 | 61.0 | 54.8 | 64.4 | 61.4 | 63.9 | 44.1 | 55.2 | 56.1 | 54.4 | 55.9 | 52.7 | 67.9 | 57.6 |
| MOLMO-2-8B | 56.7 | 65.2 | 54.8 | 60.2 | 63.3 | 53.9 | 34.4 | 54.7 | 56.6 | 55.3 | 53.9 | 52.0 | 37.5 | 53.7 |

Table 8: Ablation study evaluating the impact of each quality control agent on the acceptability of the synthesized tasks.

| Fact Checker | Logical Checker | Consistency Agent | Human Acceptance |
|---|---|---|---|
| ✗ | ✗ | ✗ | 84.1% |
| ✗ | ✓ | ✓ | 86.5% |
| ✓ | ✗ | ✓ | 92.7% |
| ✓ | ✓ | ✗ | 88.4% |
| ✓ | ✓ | ✓ | 95.5% |

Table 9: OCR layout ablation. Scores are reported by task level and overall average.

| Input | L1 | L2 | L3 | Overall |
|---|---|---|---|---|
| OCR+Layout | **33.8** | **37.6** | 23.4 | **30.4** |
| OCR-NoLayout | 27.2 | 35.6 | **23.6** | 28.7 |

# D Additional Analyses

## D.1 MAH-V Quality Control Ablation

We perform an ablation analysis over the three quality control agents to validate their effectiveness, as shown in Table 8. For each ablation, we sampled 1,391 QA pairs and calculated the average human acceptance over three trained annotators. Removing any quality control module decreases human acceptance: removing the Fact Checker causes the largest drop ($-9.0$ percentage points, to 86.5%), followed by the Consistency Agent ($-7.1$ points) and the Logic Checker ($-2.8$ points). Removing all agents further degrades acceptance to 84.1%.

Complementing the end-to-end MAH-V SFT comparison in main-paper Table 4, we additionally ablate the layout-aware OCR transformation in Table 9. Although the overall gap is modest, removing layout-aware reordering lowers the overall score by 1.7 points and causes a large drop on L1.2 Dynamic Text Localization (19.6 $\rightarrow$ 7.3), confirming that spatial structure is especially important for text grounding.

## D.2 VerEval Weighting Ablation

We select the final VerEval aggregation weight by maximizing alignment with human judgments on held-out annotated samples. Since $S_{consist}$ is produced by the consistency prompt and subsumes factual conflicts as well as fact–logic coherence, it serves as the primary term; $S_{logic}$ provides a secondary penalty for flawed reasoning chains. We report AUC (area under the ROC curve), which measures how well a metric ranks human-labeled *Cor-*

Table 10: VerEval weighting ablation measured by AUC against human annotations.

| Weight | Ann-1 | Ann-2 | Ann-3 | Avg. |
|---|---|---|---|---|
| 0.5 / 0.5 | 0.827 | 0.827 | 0.837 | 0.830 |
| 0.7 / 0.3 | 0.827 | 0.826 | 0.840 | 0.831 |
| 0.8 / 0.2 | **0.828** | **0.827** | **0.840** | 0.831 |
| 0.9 / 0.1 | **0.828** | **0.827** | **0.840** | **0.832** |

*rect* answers above *Wrong* answers; higher values indicate stronger agreement with human judgments. Table 10 shows that the $0.9/0.1$ consistency–logic weighting achieves the highest average AUC across annotator sets. The variance across configurations is below 0.002, indicating that VerEval is robust to the exact weighting choice rather than sensitive to parameter tuning.

### D.3 Model-Size Scaling

Figure 5 shows scaling effects for Qwen2.5-VL and Qwen3-VL. Qwen2.5-VL shows steady gains across all levels, with L1 rising by 9.9 points and L3 by 17.1 points. Conversely, Qwen3-VL displays non-monotonic patterns: while L1 and L2 scale upward, L3 performance regresses at 8B (42.7 $\rightarrow$ 39.4) before recovering at 32B. These results indicate that perception scales predictably, whereas understanding and reasoning exhibit instability at scales.

### D.4 Video-Pretrained Models

Video-native models show mixed results, with most underperforming image-based counterparts. LLaVA-NeXT-Video exhibits catastrophic failure (15.1 overall), while VideoLLaMA3-7B (40.9) and MOLMO-2-8B (35.6) trail behind similarly-sized image models like Qwen3-VL-4B (51.9). VideoR1-7B achieves competitive performance (50.4) with notable strength in L3.2 (62.4), but still struggles with Structured Text Synthesis (L3.3: 24.2) and Viewpoint Trajectory Description (L3.6: 24.6). These results suggest that conventional video pre-training on third-person, motion-centric datasets does not directly transfer to first-person, text-dense wearable scenarios where frame-level text clarity, wearer-view alignment, and semantic consistency matter more than generic motion modeling.

## E Privacy Protection and Ethical Considerations

**Data Collection Ethics.** All data collection procedures were conducted in accordance with institutional ethical guidelines and privacy protection standards. We obtained appropriate consent from all participants involved in the first-person video recording process, ensuring full transparency about data usage and research purposes.

**Face Anonymization Protocol.** To safeguard individual privacy, we implemented comprehensive face anonymization measures across our dataset. All human faces appearing in first-person perspective videos and images were automatically detected and masked using state-of-the-art face detection algorithms. This process ensures complete de-identification while preserving the visual context necessary for multimodal reasoning tasks. Following automated processing, we conducted systematic manual verification across all processed frames to ensure complete face anonymization coverage. Any frames with incomplete anonymization were reprocessed until satisfactory blur coverage was achieved.

**Data Usage and Distribution.** The anonymized dataset is exclusively intended for non-commercial scientific research purposes. The dataset will be made available to the research community under strict usage agreements that: (1) limit usage to academic and non-commercial research only; (2) explicitly prohibit attempts to reverse the anonymization process or identify individuals; (3) require proper attribution and responsible use of the data. All data sharing protocols comply with international privacy regulations and academic research standards, ensuring the dataset contributes solely to advancing scientific knowledge in multimodal AI research. These comprehensive privacy protection measures ensure that our benchmark contributes to multimodal AI research while maintaining the highest standards of participant privacy and data ethics. The anonymization process does not compromise the scientific validity of the tasks, as the core visual reasoning capabilities being evaluated remain fully intact.