

AVRA: An Audio-Visual Reasoning Agent for Long-Form Video QA

Team IUCV

Edred Azziz* Juhung Ha* David Crandall†
eazziz@iu.edu juhha@iu.edu djcran@iu.edu

Luddy School of Informatics, Computing, and Engineering
Indiana University Bloomington, Bloomington, IN 47408, USA

*Equal contribution. †Corresponding author.

Abstract

We present the Audio-Visual Reasoning Agent (AVRA), a training-free system for the KilometerAudio track of the Perception Test Challenge 2026. Frozen models are evaluated zero-shot without training or fine-tuning. Our approach has three components. First, a precomputed, query-agnostic index of audio events, visual objects, and on-screen text provides a structured and searchable representation of each long video, enabling agentic temporal localization. Second, a failure-driven, human-in-the-loop workflow search accumulates findings from validation runs to iteratively develop a general solver workflow while keeping the model fixed. Third, at test time, a fresh AVRA solver uses the index to select candidate intervals, directly inspects the corresponding raw media, and records evidence and completed steps before deriving an option-independent answer. It performs a targeted recheck when the observation and answer choices do not align. The selected workflow is fixed before test inference. Our submission achieved 0.78 top-1 accuracy and ranked second. Our code is publicly available.¹

1. Introduction

The KilometerAudio track of the Perception Test Challenge 2026 [4] evaluates audio-visual question answering over long walking-tour videos. The competition provides a labeled validation set of eight questions over six videos, while the test manifest used for our submission contains 500 questions over 218 videos. Given the duration of these recordings, exhaustively processing all raw audio and video is costly under practical inference budgets. Retrieval-based methods offer a practical way to narrow the temporal search space to a small set of potentially relevant intervals [2].

¹<https://github.com/juhha/AVRA-Solver>; <https://github.com/eazziz/AVRA-Metadata-Extractor>.

For a solver that uses temporal retrieval, the workflow also involves deciding how to inspect raw media, record evidence, track completed steps, and revisit uncertain observations. These choices interact across identification, counting, duration, and timing questions, making a reliable general workflow difficult to specify in advance.

We therefore develop the Audio-Visual Reasoning Agent (AVRA) through a failure-driven, human-in-the-loop search over general agent workflows (Fig. 1a). In each round, fresh solver agents apply the current pipeline to validation questions, after which independent agents and a human audit their evidence and answers. Based on these findings, we make question-agnostic revisions and repeat the loop on failed cases and clean controls. Model weights remain fixed; only the pipeline changes, and its final version is fixed before test-set inference.

The resulting AVRA workflow applies a single general template in a fresh session for each question. It combines metadata-assisted temporal localization, direct raw-media inspection, evidence logging, and checklist-guided execution. The solver commits its observation before reading the answer choices and rechecks the evidence when needed (Fig. 1b). Using this workflow, our submission achieved 0.78 top-1 accuracy and ranked second on the public leaderboard.

2. Method

2.1. Query-Agnostic Metadata Extraction

Each video is processed once to build a question-independent temporal index from sound events, visual objects, and on-screen text. BEATs [1, 5] produces confidence tracks for 447 AudioSet classes [3] at 40 ms intervals. YOLOE [6] detects 51 object categories on frames sampled at 2 fps, retaining each class, confidence, timestamp, and bounding box. PP-OCRv6 [7] extracts multilingual text at 2 fps with confidence, timestamp, and image location. The

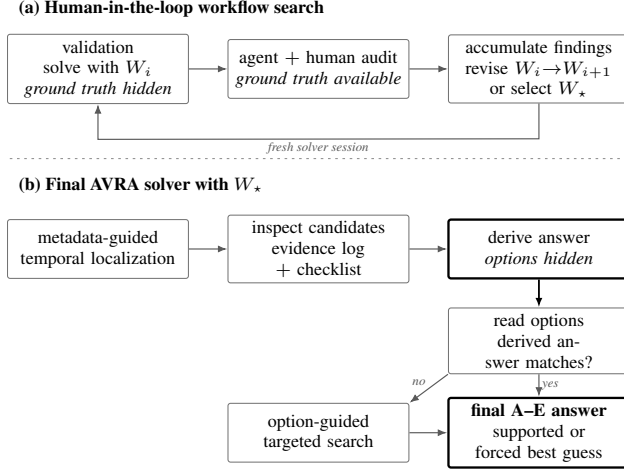


Figure 1. AVRA workflow search and inference. (a) Validation findings accumulate across fresh solver runs to revise W_i or select W_* . (b) A fresh solver localizes and inspects candidate media before reading the options; a mismatch triggers targeted search, followed by a supported answer or forced best guess.

three streams remain separate so the solver can inspect their complete inventories and select question-relevant cues for candidate localization.

2.2. Human-in-the-Loop Workflow Search

We use the AGY agent harness with Gemini 3.7 Flash (High). Among the configurations we evaluated, this combination provided the best balance of speed, cost, and utility for multimodal inspection, including direct audio. We keep the harness-model configuration fixed throughout workflow search and test inference. We search for a general solver workflow through the iterative process shown in Fig. 1a. At round i , fresh solver agents use workflow W_i to answer validation questions without access to ground-truth answers. After each round, agents and a human inspect the solver traces together with the ground truth to identify recurring failure modes.

Across rounds, we maintain a cumulative development record of observed failures, their diagnoses, and the resulting revisions, while each solver evaluation starts in a fresh session. Based on the accumulated findings, question-agnostic revisions produce W_{i+1} , which is evaluated on failed cases and clean controls. Repeating this process yields the final workflow W_* , which is fixed before test-set inference.

2.3. Agentic Long-Video Solver

The final AVRA solver follows workflow W_* , selected through the search in Sec. 2.2, for every test question (Fig. 1b). It first uses metadata for temporal localization. Without reading the answer options, it inspects the candi-

Per-question metric	Mean $\pm$ SD	Min-max
Source-video duration (min)	72.30 $\pm$ 12.23	60.00–133.00
View duration (min)	2.96 $\pm$ 1.30	0.13–9.83
Candidate windows	3.64 $\pm$ 0.70	2–5
Input tokens (10^3)	402.01 $\pm$ 128.32	212.48–1208.48
Output tokens (10^3)	26.16 $\pm$ 7.83	13.26–70.94
Thinking output	10.77 $\pm$ 3.19	5.69–31.77
Assistant output	15.39 $\pm$ 5.00	6.83–40.81
Session time (min)	17.88 $\pm$ 22.34	2.54–197.14

Table 1. Detailed inference statistics. Source and view durations measure the available and inspected media, candidate windows are localized intervals, and session time is per-question wall-clock latency. Token counts are aggregated over each solver session.

date regions, records its observations in an evidence log, and derives an option-independent answer. It then reads the options and selects a matching choice. If no option matches, it performs an option-guided targeted search; if the answer remains unresolved, it returns a forced best guess. Checklist-guided execution, refined through workflow search, is part of W_* : each solver marks completed stages during inference.

Each question is assigned to a separate instance of the solver through the AGY harness. All instances share the same model, harness, instruction template, and workspace structure; only the question and source video change. To process the 500-question test set, we ran up to 30 solver agents in parallel.

3. Results

Our submission achieved 0.78 top-1 accuracy and ranked second on the public leaderboard. Detailed inference statistics are reported in Table 1.

4. Discussion and Limitations

Metadata-assisted localization improved validation accuracy and reduced inference cost by narrowing raw-media inspection. Audio and object indices supported most questions, while OCR helped with visible text. Checklists and explicit evidence logs reduced skipped steps and unsupported conclusions. The general-purpose tool harness and shared template allowed solver agents to autonomously construct and execute question-specific analysis procedures, rather than requiring hand-crafted tools for each question. Separate analysis agents reviewed validation results and solver traces, automating much of the routine evaluation work and thereby reducing the human effort required for workflow refinement. However, when agents over-focused on subproblems or reinforced shared errors, human guidance helped redirect the workflow search. Automatically generated visual-caption metadata was explored during the final refinement but could not be integrated within the available time.

References

- [1] Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, Wanxiang Che, Xiangzhan Yu, and Furu Wei. BEATs: Audio pre-training with acoustic tokenizers. In *Proceedings of the 40th International Conference on Machine Learning*, pages 5178–5193. PMLR, 2023.
- [2] Yeeun Choi, Youngbeom Yoo, Joon-Young Lee, Hyolim Kang, and Seon Joo Kim. Keep it simple: Multi-key episodic memory retrieval for ultra-long video understanding. In *European Conference on Computer Vision (ECCV)*, 2026.
- [3] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 776–780, 2017.
- [4] Viorica Pătrăucean, Lucas Smaira, Ankush Gupta, Adrià Recasens, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Joseph Heyward, Mateusz Malinowski, Yi Yang, Carl Doversch, Tatiana Matejovicova, Yury Sulsky, Antoine Miech, Alex Frechette, Hanna Klimczak, Raphael Koster, Junlin Zhang, Stephanie Winkler, Yusuf Aytar, Simon Osindero, Dima Damen, Andrew Zisserman, and João Carreira. Perception test: A diagnostic benchmark for multimodal video models. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2023.
- [5] Florian Schmid, Tobias Morocutti, Francesco Foscari, Jan Schlüter, Paul Primus, and Gerhard Widmer. Effective pre-training of audio transformers for sound event detection. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2025.
- [6] Ao Wang, Lihao Liu, Hui Chen, Zijia Lin, Jungong Han, and Guiguang Ding. YOLOE: Real-time seeing anything. In *IEEE/CVF International Conference on Computer Vision*, 2025.
- [7] Yubo Zhang, Xueqing Wang, Manhui Lin, Yue Zhang, Penglongyi Deng, Ting Sun, Tingquan Gao, Zelun Zhang, Jiaxuan Liu, Changda Zhou, Hongen Liu, Suyin Liang, Cheng Cui, Yi Liu, Dianhai Yu, and Yanjun Ma. PP-OCRv6: From 1.5m to 34.5m parameters, surpassing billion-scale VLMs on OCR tasks. *arXiv preprint arXiv:2606.13108*, 2026.