The Fourth Perception Test Challenge

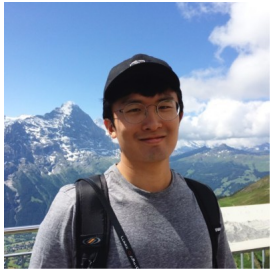
KilometerAudio (Task 3) @ECCV2026

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

Team IUCV



Edred Azziz · Undergraduate



Juhyung (Tony) Ha · Ph.D. student



David J. Crandall · Faculty

Indiana University Bloomington

KilometerAudio (Task 3)

218
test videos
~1-hour walking tours

500
test questions
multiple choice

~72 min
avg. video length
unnarrated, real-world

5 · 8
validation videos · questions
with ground truth



Example questions

Q Around what time of the day is there an ambulance heard?

Q How many times did the bell ring, and what was the time of day?

Q In front of Pizza Hut, which instrument is being played?

Temporal localization · **Audio–visual linking** · **Event counting** · **Duration estimation** · **Fine-grained recognition**

Method Overview

Foundation

Agentic harness **AGY** (antigravity-cli) with a **Gemini 3.7 Flash** — High Thinking backend.

Command-line tools + flexible problem-solving strategies for diverse question types.

The problem

Out of the box, agents under-investigate the video: they answer before gathering evidence, and reasoning is weakly supported.

01

Metadata-based retrieval

Query-agnostic temporal indices (sound events, objects, OCR) replace exhaustive frame search.

Slide 4

02

Loop engineering

Human-in-the-loop workflow search: iteratively refine the harness on the validation set.

Slide 5

03

Agentic long-video solver

Option-blind reasoning with checklists and targeted verification at test time.

Slide 6

COMPONENT 01

Query-Agnostic Metadata Extraction

The challenge

Early loops burned tokens on exhaustive search: N×N frame grids inspected until a cue appeared.

Our solution

Precompute structured, query-independent temporal indices over three modalities.

The impact

The solver queries indices first for initial temporal localization — no exhaustive visual search, far fewer tokens, far less time.

01

Sound events — BEATs

447 AudioSet classes at 40 ms resolution

A timeline of what sounds occur, and when

02

Objects — YOLOE

Open-vocabulary detection, 51 categories at 2 FPS

Class labels, confidence, boxes, timestamps

03

On-screen text — PP-OCRv6

Multilingual OCR at 2 FPS

Storefronts, street signs, contextual text

Loop Engineering

~50
automated iterations

3-5
human interventions

Using
5 · 8
validation videos · questions

Model weights stay fixed — only the workflow and prompts are optimized.

1
Draft harness W_0

"Here are the video and the question — solve it."

2
Run on validation

Solver agents attempt all 8 questions.

3
Audit failures

Evaluator agents review traces; humans step in when needed.

4
Refine the harness W_i

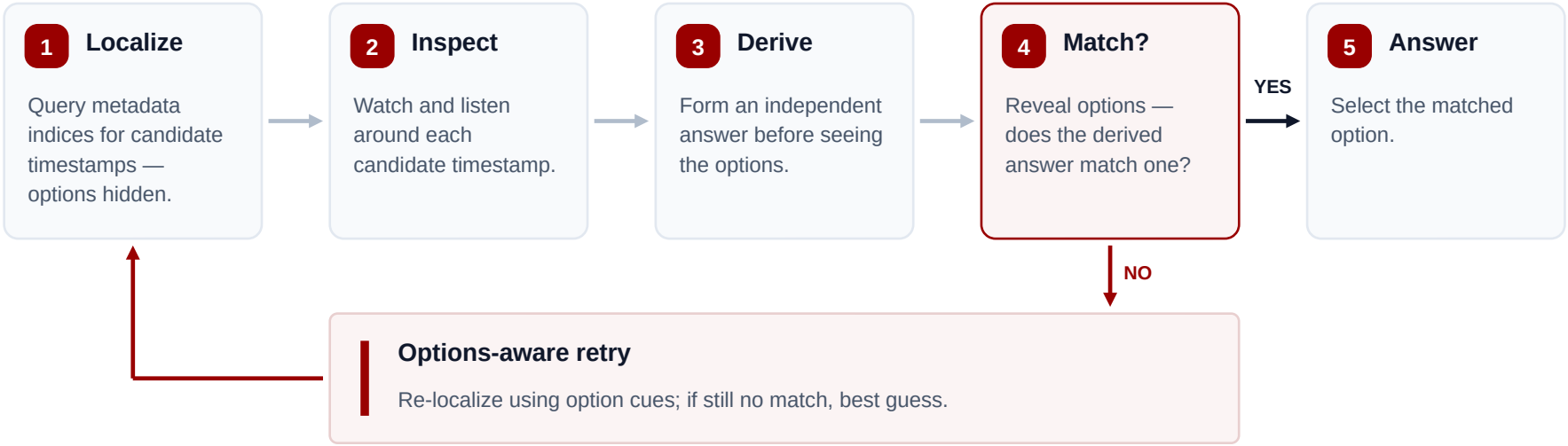
Rewrite instructions, checklists, constraints.

W^*
final workflow

Frozen before touching the test set.

COMPONENT 03 · TEST-TIME WORKFLOW

Agentic Long-Video Solver



✓ **Checklist per step**

required actions are ticked off, so media inspection can't be skipped.

✗ **Options concealed**

choices stay hidden until observations are logged, preventing anchoring.

PERFORMANCE AND STATISTICS

Results

78.2%

Top-1 accuracy

2nd place on the leaderboard

CONFIGURATION

AGY harness · Gemini 3.7 Flash (High Thinking)

30 agents in parallel over hours

>95%

less video watched

~3 min viewed of **72-min** videos

3.6

candidate timestamps

avg. selected per question

~17 min

avg. time per question

σ 22 min · range 2.5–197 min

428k

avg. tokens per session

402k input · 26k output

CLOSING

Thank You

Thank you!

For

Questions

Collaborations

Job Opportunity

We'd love to hear from you.

Team IUCV

Edred Azziz	*Undergraduate*	eazziz@iu.edu
Juhyung (Tony) Ha	*Ph.D. student*	juhha@iu.edu
David J. Crandall	*Faculty*	djcran@iu.edu

Luddy School of Informatics, Computing, and Engineering — Indiana University Bloomington

Open source

AVRA Solver
github.com/juhha/AVRA-Solver

Metadata Extractor
github.com/eazziz/AVRA-Metadata-Extractor