Evidence-Grounded Tool Use for Hour-Long Audio-Visual Question Answering

Yuyang Sun Yu Zhang Xiaogang Wang Xiao Dong Bing Wang
Bin Song Xuping Zhu Bin Zhang Kent Shen

2132115O6@seu.edu.cm, zhangyu@freedotech.com, wangxiaogang@swu.edu.cn,
dongxiao@freedotech.com, wangbing@freedotech.com, songbin@freedotech.com,
zhuxuping2025@gmail.com, zhangbin@smartcitysz.com, m15805155114@163.com

Abstract

KilometerAudio requires five-way question answering over hour-scale videos, where decisive evidence may be a brief sound, speech segment, repeated event, or an audio event linked to a visual source. Uniform visual sampling is inefficient and can miss such temporally sparse evidence. We build , a zero-shot agent that first creates reusable full-track speech and acoustic indexes, then lets a planner request bounded audio, speech, acoustic, and visual observations for one question at a time. Qwen3-Omni verifies short audio excerpts, LAION-CLAP retrieves candidate windows, Faster-Whisper supplies timestamped speech, and Gemini plans tool calls and inspects sparse frames. A conservative evidence gate returns an answer only after independent observations agree, while fixed compute and frame budgets make every decision auditable.

1. Task and Design Goals

The Perception Test benchmark evaluates multimodal perception and reasoning in video [1]. Task 3 focuses on audio-dependent questions over walking-tour videos: identifying a sound or speaker, transcribing speech, counting events, estimating duration, locating an event, or combining an audible event with a visible object. Each example contains one video, one question, and five options.

The central difficulty is temporal sparsity. Decoding an entire multi-hour video at high visual rate is expensive, while uniform sampling can miss a short bell, announcement, or interaction. Our design therefore follows three principles: index the entire audio track cheaply; spend generative-model compute only on retrieved windows; and require grounded evidence before replacing a conservative answer. The system processes each question independently during model inference and emits exactly one option ID.

2. Evidence-Grounded Agent

2.1. Reusable full-track evidence

We demultiplex audio without altering timestamps. Faster-Whisper `large-v3-turbo` [3, 4] produces timestamped segments together with language and reliability statistics. Empty transcripts are represented as missing evidence, not proof that speech is absent. We filter common low-information hallucinations and retain absolute timestamps for later retrieval.

In parallel, `laion/clap-htsat-fused` [2] embeds 10-s clips at a 5-s stride. About 120 acoustic descriptions (for example, bell, siren, footsteps, music, and vehicle sounds) are scored against each video's clip distribution. Per-video z-normalization is important because raw audio–text similarities are not calibrated across recordings. The resulting event timeline supports global search using one text encoding and matrix similarity, rather than repeatedly decoding a full track.

2.2. Question-conditioned tool use

Gemini 3.1 Pro Preview is the planner and visual reasoner. At each round it sees the question, options, compact index evidence, and previous observations, and selects exactly one action. Audio excerpts are capped at 30 s per Qwen call. For option-sensitive decisions, Qwen is queried with both original and reversed option order; a changed answer is marked order-sensitive and receives zero confidence. Gemini never receives the full video. Visual calls use targeted frames or normalized crops and are capped at 48 frames per question. The controller allows at most six or seven rounds depending on the run configuration.

2.3. Answer-type routing and evidence gates

Regex-based taxonomy identifies count, duration, timestamp, clock-reading, speech, sound-identity, and visually grounded questions. It changes tool preconditions, not the answer itself. Counts and durations require fresh audio inspection; identity questions require fresh audio evidence;

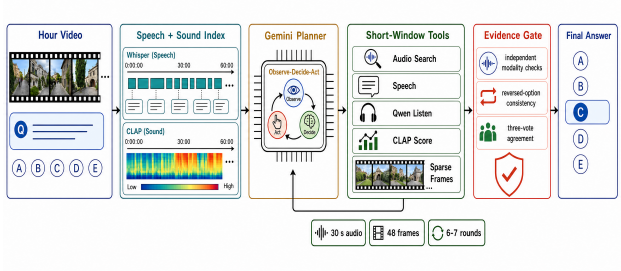

Figure 1. Overview of GroundedAV-Agent. The system first builds reusable full-track speech and acoustic indexes, then uses a Gemini planner to iteratively invoke question-conditioned tools over bounded evidence windows. Focused Qwen listening, speech retrieval, CLAP scoring, and sparse-frame inspection provide complementary observations. A conservative evidence gate checks cross-modal agreement, option-order consistency, and repeated planner votes before returning the final answer, under fixed budgets of 30-second audio clips, 48 frames, and six to seven reasoning rounds.

clock questions favor cropped visual checks near an audio-localized event.

The stop decision is sampled three times (greedy plus two temperature-0.7 samples). A proposed change must satisfy: (i) at least two independent fresh modalities support it; (ii) no fresh observation supports the previous answer; (iii) at least two planner votes agree; (iv) the answer type is not frozen by reliability calibration; and (v) the run used grounded observations that hide cached answer letters. Parse failures fall back to the existing candidate, never automatically to option A. A final audit verifies 500 unique IDs, option text/ID consistency, evidence provenance, frame budgets, and the absence of eliminated choices.

3. Models, Data, and Evaluation Mode

Table 1 lists all learned components. We did not update model parameters and used no challenge examples as in-context demonstrations. The per-question model treats validation and test videos only as inference inputs; we therefore characterize its evaluation mode as **zero-shot**. The pretrained checkpoints inherit the providers' original training corpora; the proprietary Gemini training mixture is not disclosed. Our only task-specific assets are prompts, deterministic routing rules, a vocabulary of acoustic descriptions, and indexes computed from the video being answered.

Local inference used at most four H100 GPUs. Model instances were co-resident per worker and local calls were serialized; multiple question threads overlapped only remote API latency. Cached indexes make repeated questions over one video inexpensive while preserving question-level inference isolation.

Table 1. Learned components; none was fine-tuned for Kilometer-Audio.

Role	Model
Planner + vision	Gemini 3.1 Pro Preview [6]
Audio	Qwen3-Omni 30B-A3B [5]
Speech	Whisper large-v3-turbo [3, 4]
Retrieval	LAION-CLAP HTSAT-fused [2]

4. Implementation and Reproducibility

The pipeline has two execution stages. The offline stage builds one ASR record and one CLAP matrix per video. The online stage loads only the index entries relevant to the current question and launches tools on demand. Each action appends a structured observation containing its source, time interval, answer hypothesis, confidence, and reliability flags. This trace is passed back to the planner before another action is allowed, yielding a literal observe–decide–act loop rather than a fixed cascade.

Question routing and budgets are deterministic. Full-track CLAP uses 10-s windows with 5-s overlap; Qwen receives excerpts no longer than 30 s per call; visual inspection is limited to 48 frames per question; and the controller stops after at most six or seven rounds. Four H100 GPUs host the local audio models, while visual reasoning uses a remote Gemini endpoint. Local model calls are lock-serialized and worker threads overlap only remote latency. All option IDs are checked against their option text before the prediction file is written.

For speech, we preserve segment timestamps, language probability, no-speech probability, compression ratio, and occupancy. A high-sensitivity voice-activity-detection pass is available when the first transcript is empty or unreliable. For acoustic retrieval, z-scores are always interpreted within the current video; an absolute CLAP similarity is never treated as a calibrated probability. These details are important because long recordings differ substantially in noise floor and event density.

References

[1] V. Pătrăucean et al. Perception Test: A diagnostic benchmark for multimodal video models. *NeurIPS Datasets and Benchmarks*, 2023. 1

[2] Y. Wu et al. Large-scale contrastive language–audio pretraining with feature fusion and keyword-to-caption augmentation. *ICASSP*, 2023. 1, 2

[3] A. Radford et al. Robust speech recognition via large-scale weak supervision. Technical report, OpenAI, 2022. 1, 2

[4] SYSTRAN. Faster-Whisper: CTranslate2-based Whisper inference. https://github.com/SYSTRAN/faster-whisper. 1, 2

[5] Qwen Team. Qwen3-Omni-30B-A3B-Thinking model card. https://huggingface.co/Qwen/Qwen3-Omni-30B-A3B-Thinking. 2

[6] Google DeepMind. Gemini API model documentation. https://ai.google.dev/gemini-api/docs/models. 2