Amortized Meta-Reasoning for Kilometre-Scale Audio QA

Venkata Sai Nikhil Thodupunuri and Ravi Vayuvegula

t.v.s.nikhil123@gmail.com, ravivayuvegula@google.com

Abstract. The KilometerAudio track of the Perception Test challenge asks what made a sound, when it started, how long it lasted, how often it recurred, and which instrument or song was playing, over 218 first-person walking-tour videos totalling 262.5 hours. The hard part is localisation, not recognition: a question never says when the event happened, and the five options lie close enough together that the nearest detected event does not discriminate between them. We reuse the explore-then-compile strategy of our KilometerVision entry, replacing 3D reconstruction with a multi-encoder acoustic index and a set of measurements taken directly from the waveform. Any detection the answer depends on is checked against the signal itself before it is used. An answer commits when two traces from *different* evidence families agree, a two-model ensemble supports them, and an adversarial pass fails to break them. An answer that cannot meet them is retried while budget remains, then committed at low confidence and marked as such. The system reaches **86.4**% on the test set.

1 Introduction

KilometerAudio poses 500 five-way questions over 218 hour-long walking-tour videos, answerable only from the audio track: which object made a sound, when it began, how long it ran, how often it repeated, which instrument or song was heard.

Recognition is not the bottleneck; localisation is. A tagger run over an hour of audio returns its top hits whether or not the queried sound is present, and the options are close enough together that an event of roughly the right kind will land near one of them. A clean match is therefore weak evidence that the right event was found.

We treat this as **amortized meta-reasoning**: an agent searches, per question, over several grounding strategies and reconciles independent estimates, leaving a trace; those traces are compiled into one policy covering the rest of the corpus without repeating the search. Our contributions are an operator bank that reads the signal directly, with no neural component, derived from exploration rather than designed in advance, which measures where the encoders only classify; and a verification policy that commits an answer only on two traces from different evidence families, a two-model ensemble and an adversarial review.

Table 1: The acoustic index.

encoder	output	role
AST [10]	527 classes	primary tagger
BEATs [4]	527 classes	second tagger
CLAP [31]	512-d	open-vocabulary search
CLAP-music [31]	512-d	instrument and song
CLAP-fused [3, 31]	512-d	fused, longer context
M2D-CLAP [21]	768-d	alternative audio embedding
SigLIP2 [28]	768-d	visual referent

2 Related Work

Video question answering is served by a large family of benchmarks [2, 7, 13–20, 22,25–27,30,32,33,35]. These typically emphasise recognising objects, identifying actions, and understanding events that are directly observable within individual frames or short clips. KilometerAudio differs on both counts: the evidence is an hour long, and it is carried by the audio track, so the question is where to look rather than what is in view.

Audio tagging on AudioSet [8] is dominated by transformer classifiers such as AST [10] and BEATs [4], with contrastive language–audio models such as CLAP [31] and M2D-CLAP [21] adding retrieval by text query. These give a class label rather than a measurement, whereas classical signal processing [1, 6] yields the quantities an answer needs. Agentic loops such as ReAct [34] interleave reasoning with tool calls, while Toolformer [24] instead teaches a model to emit tool calls in-line. Self-consistency [29] improves reliability by sampling one model's reasoning several times. We decorrelate differently: by changing what produces the evidence – a second model, and a second family of evidence – rather than by resampling one path. Explore-then-compile is meta-reasoning under a budget [23] and amortized inference [9], distinct from gradient-based meta-learning [5].

3 Method

3.1 Acoustic grounding layer

Built once per video, indexed to absolute video seconds so a hit in one representation can be checked against another. Several encoders are used rather than one (Table 1), covering closed-vocabulary tagging, open-vocabulary search by text query, and music. Visual embeddings support questions whose referent is named visually, and help fix timing where the audio is ambiguous.

A detection the answer depends on is not accepted on an encoder's word. It is settled by the measurement operators of §3.3, which read the signal directly.

Each video is also annotated in overlapping two-minute chunks. This gives audio events with labels and times, a caption per chunk, and an index from

on-screen text to time. It is a prior over where to look, never an answer: its timestamps carry several seconds of error, wider than the typical gap between neighbouring options.

The index and the annotation are the whole of the offline stage. They locate events but do not produce the number an answer needs; the operators that do are written by the agent during exploration (§3.3).

3.2 Explore, compile, deploy

Search runs at high effort on a small explore split – the validation set plus about 10% of the test set – inside a coding-agent harness whose context directs it to write whatever measurement code the question needs. The traces are compiled into one policy, which a lighter orchestrator runs over the remaining ∼90% without per-question search. Each question is answered in an isolated session, as the competition requires. Every component is served by *Gemini 3.1 Pro* [11], except the two-model ensemble below, which queries *Gemini 3.1 Pro* and *Gemini 3.5 Flash* [12].

A session re-runs retrieval with queries it composes itself, shortlists regions from the annotation, measures, listens and looks, and applies the option grid as a constraint, recording an option, a confidence, a moment, its reasoning, and what would overturn the answer. Compilation labels every tool call in the promoted traces, takes their union as a weighted transition graph, and derives a routing rule from question kind to operator plus a commit gate. The operators it routes between are deterministic code; only routing and verification are delegated to the language model. What compilation carries forward is the *logical decomposition* – which operator applies to which kind of question, in what order, and under what verification. It carries no per-question or per-video content: no answers, no timestamps, no per-video observations. The compiled policy is question-agnostic, so each test question is answered independently from its own evidence, as the competition requires.

Three failure modes shaped the compiled rules. *Unverified detection*: an encoder can rank a spurious event highly, so any load-bearing detection is checked against the raw spectrum. *Hypothesis-shaped verification*: a filter chosen because it isolates the hypothesised sound confirms it regardless of the truth, so a method must pass a positive control (detect the target where it is known to occur) and a negative control (stay silent where it should be absent). *Measure then snap*: a question states several conditions at once – the kind of sound, what the camera was doing, which edge of the event is meant – and an event of the right kind can be measured precisely and matched to the nearest option while the other conditions go untested, so precision is mistaken for correctness. Each condition is therefore tested separately, and confidence reflects whether the right event was identified rather than how precisely some event was measured.

The gate operates at two levels. At the level of a single *perceptual reading* – asking a model what a clip contains, what an instrument is, or what a sign says – the same clip and prompt go to both *Gemini 3.1 Pro* and *Gemini 3.5 Flash*. Where the two agree the reading is used; where they disagree it is treated as

unresolved and a direct measurement decides. This is the two-model ensemble, and it applies within the listening and frame families only, since the other families involve no model judgement.

At the level of the *answer*, two further conditions apply. *Two independent traces*: evidence is partitioned into families – learned detectors, direct measurement, listening, frames, option grid – and the second derivation must come from a different family and be worked out afresh. *An adversarial review*: *Gemini 3.1 Pro* is given the answer and its evidence and instructed to argue it is wrong, and to say so plainly if it looks right. Where the evidence cannot meet these conditions, the session takes another route while budget remains, and only once the approaches are exhausted does it commit at low confidence with the shortfall recorded. An answer is never withheld.

3.3 Skills learned during exploration

Exploration yields three reusable things, and together they are what the deployed policy knows that a generic plan–act loop does not.

An operator bank. Every operator in Table 2 was written by an agent during exploration, kept because it survived verification, and reused by the compiled policy rather than rebuilt per question. Each answers a question the encoders cannot: the encoders say a sound resembles a class, the operators say where its edges are, how many there were, and whether the reading holds up.

A routing skill mapping question kind to the index queries and operators that settle it (Table 3); what is learned is which of them to reach for. The mapping is kind-specific: pitch tracking and harmonic tests carry most weight on instrument questions, the music-specialised encoders on song questions, and counting rests chiefly on onset intervals.

Strategy skills, the sequences that recur across successful traces: check the premise over the whole video before analysing any window; localise with the annotation but measure with signal processing; confirm the referent in the frames independently of the audio; let a direct measurement decide where the two models disagree; and run a control contrast before committing.

Only the operator bank, the routing and the strategies carry over. The per-video index of §3.1 is not regenerated per question, and no operator writes back into it.

4 Experiments

The test set is 500 five-way questions over 218 videos, scored as top-1 accuracy, so chance is one in five. Kinds are dominated by sound-source identity and timestamp, followed by instrument, count, duration and song, with a small remainder of other kinds.

Evaluation mode and data. No model is trained or fine-tuned for this task, so the system is zero-shot and prompted throughout. The encoders are used with their public released weights, and no challenge data is used to train, tune

Table 2: The operator bank derived during exploration. Each operator is deterministic code over the raw signal or the index, with no learned component.

operator	what it computes, and what that settles
Locating – narrowing an hour to a window	
Text-driven frame search	Embeds a written description of the object the question names and matches it against the frame embeddings, ranking moments by similarity. Answers "where in the video does this thing appear at all".
Whole-video class curve	Reduces a tagger's outputs for the queried labels to a single curve over the whole recording. Its *maximum* localises; its *absence everywhere* is the cheapest evidence that the question's premise fails.
Characterising – identifying the event and fixing its edges	
Harmonic/percussive separation	Splits the signal into sustained and transient parts. Footsteps are the largest contaminant in this corpus. They fall into the transient part, so measuring the sustained part removes them without touching bells, engines or music.
Band-limited envelope	Filters to the frequency band the event occupies and tracks energy in short steps. Where that energy rises and falls gives the onset and offset instants behind every timestamp and duration answer.
Narrowband tonality	Ratio of the strongest spectral peak in a band to the median of that band, in decibels. High values mean a pure tone against broadband noise, which surfaces bells, whistles and beepers that are inaudible in loudness terms.
Harmonic-series test	Checks whether spectral peaks sit at integer multiples of a common fundamental. Distinguishes a genuinely pitched source from a noise band that happens to peak in the same place.
Attack and decay	Rise time and decay slope of a single note's envelope. This is excitation type: bowed and blown sources rise over hundreds of milliseconds, plucked and struck ones in tens, which separates instrument families that sound alike at low bitrate.
Modulation rate	Frequency analysis of the envelope rather than the waveform, recovering periodic amplitude variation. Identifies rotor blade-pass, engine cycles and gated beepers, whose rate is diagnostic even when the source is faint.
Stereo balance	Left–right level difference over time. A source passing the camera sweeps this monotonically, which separates a moving source from a stationary one and pins the instant of closest approach.
Pitch tracking	Dominant fundamental over time, giving note identity and melodic contour for instrument and song questions.
Template matching	Cross-correlates one confirmed instance of an event against a longer span to find further instances, establishing whether an event is unique or one of several.
Onset detection + intervals	Picks transient onsets in a band, then reports the mean gap and its variability. The only route to a count, and the variability doubles as a check: irregular gaps mean the detector is mixing sources.
Verifying – deciding whether to believe the reading	
Control contrast	Runs the identical measurement on the event window and on windows where the event should be absent, and reports the difference. A measurement that fires equally in both is measuring something else.
Option scoring	Scores all five options against the same measurements, so the rejected ones are eliminated on the record rather than implicitly.

or prompt them. Both modalities are used. The audio track carries the answer and is indexed and measured directly. Frames confirm a referent the question names and, where the audio is ambiguous, fix timing from on-screen text and

Table 3: The learned routing skill: question kind to the index queries and derived operators that settle it.

question kind	operators
sound source	class curves, stereo balance, visual referent
timestamp, duration	band-limited envelopes, template matching
count	onset detection with interval statistics
instrument	pitch tracking, harmonic-series tests, music encoders
song	music encoders, option-by-option scoring

Table 4: Test accuracy (%) on the eval server.

System	Acc.
Single LLM answering from the metadata	54.0
Agentic policy over the grounding layer	76.8
+ dual-trace verification, two-model ensemble, adversarial review	**86.4**

scene changes. Questions are answered one at a time in isolated sessions, so no correlation between test questions can be exploited.

Main result. A single LLM answering directly from the indexed metadata reaches 54.0%: the representations alone do not answer these questions, because the answer-bearing seconds still have to be found and measured. An agentic policy over the same grounding layer reaches 76.8%, and adding the verification policy of §3.2 reaches 86.4% (Table 4). All three read the same grounding layer, so the differences isolate the answering policy.

5 Conclusion

We apply explore-then-compile to KilometerAudio, replacing 3D reconstruction with a multi-encoder acoustic index and a layer of direct signal measurement. An encoder's label is never the answer on its own; verification crosses evidence families, crosses models, and must survive an adversarial pass. Accuracy is 86.4% with that verification and 76.8% without it.

References

1. Bello, J., Daudet, L., Abdallah, S., Duxbury, C., Davies, M., Sandler, M.: A tutorial on onset detection in music signals. IEEE Transactions on Speech and Audio Processing **13**(5), 1035–1047 (2005). https://doi.org/10.1109/TSA.2005.851998
2. Cai, M., Tan, R., Zhang, J., Zou, B., Zhang, K., Yao, F., Zhu, F., Gu, J., Zhong, Y., Shang, Y., et al.: TemporalBench: Benchmarking fine-grained temporal understanding for multimodal video models. arXiv:2410.10818 (2024)

3. Chen, K., Du, X., Zhu, B., Ma, Z., Berg-Kirkpatrick, T., Dubnov, S.: Hts-at: A hierarchical token-semantic audio transformer for sound classification and detection. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 646–650. IEEE (2022)
4. Chen, S., Wu, Y., Wang, C., Liu, S., Tompkins, D., Chen, Z., Wei, F.: Beats: Audio pre-training with acoustic tokenizers. arXiv preprint arXiv:2212.09058 (2022)
5. Finn, C., Abbeel, P., Levine, S.: Model-agnostic meta-learning for fast adaptation of deep networks. In: International conference on machine learning. pp. 1126–1135. PMLR (2017)
6. Fitzgerald, D.: Harmonic/percussive separation using median filtering (2010), https://api.semanticscholar.org/CorpusID:52834812
7. Fu, C., Dai, Y., Luo, Y., Li, L., Ren, S., Zhang, R., Wang, Z., Zhou, C., Shen, Y., Zhang, M., et al.: Video-MME: The first-ever comprehensive evaluation benchmark of multi-modal LLMs in video analysis. arXiv:2405.21075 (2024)
8. Gemmeke, J.F., Ellis, D.P., Freedman, D., Jansen, A., Lawrence, W., Moore, R.C., Plakal, M., Ritter, M.: Audio set: An ontology and human-labeled dataset for audio events. In: 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP). pp. 776–780. IEEE (2017)
9. Gershman, S.J., Goodman, N.D.: Amortized inference in probabilistic reasoning. Cognitive Science **36** (2014), https://api.semanticscholar.org/CorpusID:924780
10. Gong, Y., Chung, Y.A., Glass, J.: Ast: Audio spectrogram transformer. arXiv preprint arXiv:2104.01778 (2021)
11. Google: Gemini 3.1 pro: A smarter model for your most complex tasks. https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-1-pro/ (2026), accessed 31 August 2026
12. Google: Gemini 3.5: Frontier intelligence with action. https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-5/ (2026), accessed 31 August 2026
13. Kesen, I., Pedrotti, A., Dogan, M., Cafagna, M., Acikgoz, E.C., Parcalabescu, L., Calixto, I., Frank, A., Gatt, A., Erdem, A., Erdem, E.: ViLMA: A zero-shot benchmark for linguistic and temporal grounding in video-language models. In: ICLR (2024)
14. Lei, J., Yu, L., Bansal, M., Berg, T.L.: TVQA: Localized, compositional video question answering. In: EMNLP (2018)
15. Li, J., Wei, P., Han, W., Fan, L.: IntentQA: Context-aware video intent reasoning. In: ICCV. pp. 11963–11974 (2023)
16. Li, K., Wang, Y., He, Y., Li, Y., Wang, Y., Liu, Y., Wang, Z., Xu, J., Chen, G., Luo, P., Wang, L., Qiao, Y.: MVBench: A comprehensive multi-modal video understanding benchmark (2023)
17. Liang, J., Meng, X., Zhang, H., Wang, Y., Wei, J., Zhao, D.: ReasVQA: Advancing VideoQA with imperfect reasoning process. In: NAACL-HLT. pp. 1696–1709 (2025)
18. Liu, H., Ilievski, F., Snoek, C.G.M.: Commonsense video question answering through video-grounded entailment tree reasoning. arXiv:2501.05069 (2025)
19. Liu, Y., Li, S., Liu, Y., Wang, Y., Ren, S., Li, L., Chen, S., Sun, X., Hou, L.: TempCompass: Do video LLMs really understand videos? arXiv:2403.00476 (2024)
20. Meng, J., Li, X., Wang, H., Tan, Y., Zhang, T., Kong, L., Tong, Y., Wang, A., Teng, Z., Wang, Y., Wang, Z.: Open-o3 Video: Grounded video reasoning with explicit spatio-temporal evidence. arXiv:2510.20579 (2025)

21. Niizumi, D., Takeuchi, D., Yasuda, M., Nguyen, B.T., Ohishi, Y., Harada, N.: M2d-clap: Exploring general-purpose audio-language representations beyond clap. IEEE Access (2025)
22. Rawal, R., Saifullah, K., Basri, R., Jacobs, D., Somepalli, G., Goldstein, T.: CinePile: A long video question answering dataset and benchmark. arXiv:2405.08813 (2024)
23. Russell, S., Wefald, E.: Principles of metareasoning. Artificial Intelligence **49**(1), 361–395 (1991). https://doi.org/https://doi.org/10.1016/0004-3702(91)90015-C, https://www.sciencedirect.com/science/article/pii/000437029190015C
24. Schick, T., Dwivedi-Yu, J., Dessì, R., Raileanu, R., Lomeli, M., Hambro, E., Zettlemoyer, L., Cancedda, N., Scialom, T.: Toolformer: Language models can teach themselves to use tools. Advances in neural information processing systems **36**, 68539–68551 (2023)
25. Shangguan, Z., Li, C., Ding, Y., Zheng, Y., Zhao, Y., Fitzgerald, T., Cohan, A.: TOMATO: Assessing visual temporal reasoning capabilities in multimodal foundation models. arXiv:2410.23266 (2024)
26. Swetha, S., Gupta, R., Kulkarni, P.P., Shatwell, D.G., A Chan Santiago, J., Siddiqui, N., Fioresi, J., Shah, M.: VRR-QA: Visual relational reasoning in videos beyond explicit cues. In: CVPR. pp. 32840–32849 (2026)
27. Swetha, S., Kuehne, H., Shah, M.: TimeLogic: A temporal logic benchmark for video QA. arXiv:2501.07214 (2025)
28. Tschannen, M., Gritsenko, A., Wang, X., Naeem, M.F., Alabdulmohsin, I., Parthasarathy, N., Evans, T., Beyer, L., Xia, Y., Mustafa, B., et al.: Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. arXiv preprint arXiv:2502.14786 (2025)
29. Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., Zhou, D.: Self-consistency improves chain of thought reasoning in language models. arXiv preprint arXiv:2203.11171 (2022)
30. Wu, B., Yu, S., Chen, Z., Tenenbaum, J.B., Gan, C.: STAR: A benchmark for situated reasoning in real-world videos. In: NeurIPS (2021)
31. Wu, Y., Chen, K., Zhang, T., Hui, Y., Berg-Kirkpatrick, T., Dubnov, S.: Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2023)
32. Xiao, J., Shang, X., Yao, A., Chua, T.S.: NExT-QA: Next phase of question-answering to explaining temporal actions. In: CVPR. pp. 9777–9786 (2021)
33. Xu, J., Mei, T., Yao, T., Rui, Y.: MSR-VTT: A large video description dataset for bridging video and language. In: CVPR (2016)
34. Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., Cao, Y.: React: Synergizing reasoning and acting in language models. arXiv preprint arXiv:2210.03629 (2022)
35. Yu, Z., Xu, D., Yu, J., Yu, T., Zhao, Z., Zhuang, Y., Tao, D.: ActivityNet-QA: A dataset for understanding complex web videos via question answering. In: AAAI. pp. 9127–9134 (2019)