FROM PROBES TO DECISIONS: MULTIMODAL EVIDENCE LEARNING FOR TRUSTWORTHY AUDIO REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Audio-language models can understand sound events and context in complex recordings, but being able to describe a recording does not imply having the evidence needed to answer a specific question. Question-guided acoustic measurements can provide targeted observations; however, even accurate observations may be insufficient to support the judgment at hand. We propose Multimodal Evidence Learning (MEL), a framework for trustworthy audio reasoning that connects question-relevant acoustic evidence acquisition with task-oriented evidence use. MEL decomposes a question's decision requirements into textual probes, guides acoustic tools to obtain observations with source context and scope, and combines these observations with the original audio and question for task decisions. The framework emphasizes what observations support and where their inferential limits lie, allowing local evidence to complement holistic audio understanding rather than be equated with task conclusions. We incorporate supervised learning into this process to strengthen the connection between a question's information needs and evidence acquisition, allowing local acoustic observations and holistic audio understanding to jointly inform task decisions. Experiments on MMAR and MMAU with two audio-language backbones of different sizes show that MEL improves overall accuracy over direct answering. Ablations and per-question analyses further reveal the contributions of evidence acquisition and learning, as well as how answer corrections and regressions jointly determine the net gains.

1 INTRODUCTION

Advances in audio-language models are expanding audio understanding beyond sound recognition and transcription toward complex question answering and natural-language retrieval (Xu et al., 2025b; Ma et al., 2026; Zhang et al., 2026a). These tasks require models to go beyond summarizing a recording and reason about sound attributes, event relations, and contextual meaning in response to a specific question. The same recording can support many different questions, each requiring different acoustic evidence, such as sound identities or temporal relations (Figure 1a). Audio reasoning therefore requires not only understanding what happens in a recording, but also identifying the evidence needed to answer the question at hand.

This process presents a dual challenge in connecting holistic understanding with targeted measurement. Holistic understanding captures context but may overlook subtle acoustic details that distinguish competing answers; targeted measurements reveal specific details but cannot, on their own, establish what those details mean in context. Even an accurate observation may support one claim while remaining insufficient to justify a stronger conclusion (Figure 1b). Reliable audio reasoning thus requires connecting evidence acquisition with evidence use: identifying critical acoustic information, relating otherwise fragmented observations, and interpreting their relevance to the current judgment within the full audio context.

Prior work has advanced audio understanding through stepwise reasoning and audio descriptions (Kuan & Lee, 2025), conditional tool use (Wang et al., 2026), and multi-source evidence fusion with reliability assessment and cross-checking (Olev & Alumäe, 2026). However, holistic descriptions or generic tool outputs may still leave the evidence required by a complex question

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

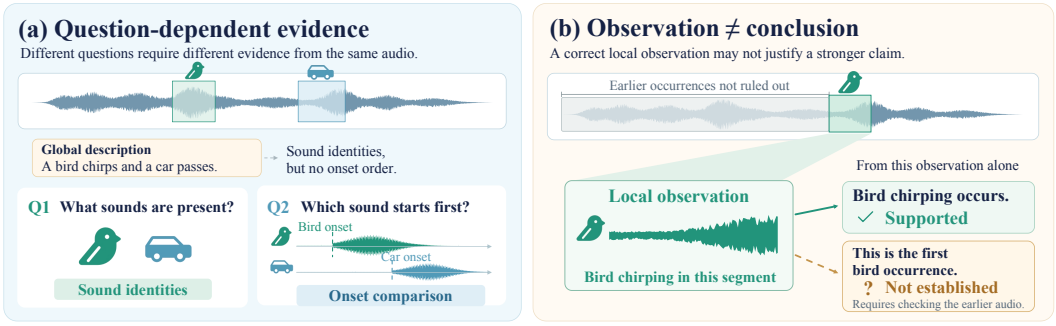

Figure 1: Two challenges in evidence-based audio reasoning. (a) Different questions about the same recording require different evidence: a description identifying sound sources does not necessarily establish their onset order. (b) A correct local observation supports the presence of an event, but does not establish its first occurrence without ruling out earlier occurrences.

unobserved. Even when relevant observations are obtained, their scope of support must be assessed in relation to the question's conditions and the full recording. We therefore ask: *How can models acquire acoustic evidence that matches a question's information needs, and improve the process from evidence acquisition to task judgment through learning to make well-supported decisions?*

To address this question, we propose **Multimodal Evidence Learning (MEL)** for trustworthy audio reasoning, a framework that connects question-guided acoustic evidence acquisition with task-oriented evidence organization and use. MEL expands a question into textual probes that target distinct information needs, guiding audio models and acoustic tools to obtain observations. It then combines these observations with the original audio and question for task decisions. The framework emphasizes the scope of support and inferential limits of observations, allowing local evidence to complement holistic audio understanding rather than be equated with conclusions. We incorporate supervised learning into this process to strengthen the connection between question requirements and evidence acquisition, enabling the acquired acoustic information to better serve task decisions.

We evaluate MEL on MMAR and MMAU, two audio understanding and reasoning benchmarks, using two audio-language backbones. Beyond overall performance, component ablations and per-question paired analyses examine the roles of evidence provision, question decomposition, and supervised learning, as well as how answer corrections and regressions jointly affect the net gains.

Our main contributions are as follows:

- **Question-guided acoustic evidence acquisition.** We translate the information needs of questions into textual probes that guide acoustic measurements, obtaining task-relevant observations with source context and scope while preserving access to the full recording.
- **Task-oriented evidence learning.** We incorporate supervised learning into a workflow connecting question decomposition, evidence acquisition, and task judgment, strengthening the alignment between question requirements and acquired evidence so that local observations and holistic audio understanding jointly inform decisions.
- **Empirical analysis across benchmarks and backbones.** We evaluate overall performance on two public benchmarks with two audio-language backbones. Component ablations and per-question answer analyses reveal the benefits of the complete pipeline, the roles of its components, and their limitations.

2 Related Work

2.1 Audio-Language Models and Audio Reasoning

Audio-language models combine audio perception with language modeling to support the understanding of speech, environmental sounds, and music through a unified natural-language interface. SALMONN (Tang et al., 2024) and Qwen2-Audio (Chu et al., 2024) expand general-purpose audio

understanding, while Qwen3-Omni (Xu et al., 2025b) further supports multimodal interaction and reasoning. Beyond advances in foundation models, audio descriptions and stepwise reasoning offer another avenue for improving audio understanding. For example, MATCH (Kuan & Lee, 2025) first generates an audio description and then answers a specific question. Together, these efforts provide a foundation for perception and reasoning in complex audio tasks.

2.2 Question-Guided Audio Evidence Acquisition

Question decomposition and tool coordination provide effective ways to acquire task-relevant audio information. Speech-Copilot (Kuan et al., 2024) and AudioToolAgent (Wijngaard et al., 2026) translate language instructions into modular processing and coordinated tool use. Audio-Mind (Wang et al., 2026) and AuTAgent (Tong et al., 2026) obtain additional observations through conditional evidence acquisition and learned tool selection, respectively. PlanRAG-Audio (Someki et al., 2026) plans the information and temporal spans required by a query to support targeted retrieval for long-form audio understanding. These approaches connect question requirements with audio processing in different ways.

MEL organizes question decomposition around the evidence needed for the current decision, translating "what to answer" into "what to listen for." This decomposition produces acoustic measurement requests aligned with task requirements, making explicit which observations are needed to inform the final decision.

2.3 Learning to Reason with Evidence

At inference time, prior work uses cross-source corroboration, reliability modeling, and conflict resolution to make evidence use more verifiable. Multi-Source Evidence Fusion (Olev & Alumäe, 2026) combines observations from multiple audio models with acoustic tool outputs, organizing the basis for reasoning through predefined reliability tiers, confidence constraints, and contradiction checks. EChO-Agent (Zhang et al., 2026b) integrates tool observations into question-relevant evidence, reasons jointly over this evidence and the original audio, and checks consistency between the evidence and the answer. In visual question answering, ReCoVERR (Srinivasan et al., 2024) collects high-confidence, relevant visual evidence through follow-up questions to reduce unnecessary abstention on low-confidence predictions. These methods highlight the role of evidence organization and verification in multimodal reasoning.

Another line of work learns the relationship between evidence and answers through training. Self-RAG (Asai et al., 2024) incorporates judgments of retrieval relevance, evidential support, and response utility into training and inference. InstructRAG (Wei et al., 2025) synthesizes rationales from retrieved documents and ground-truth answers during supervision construction, then uses these rationales as in-context demonstrations or supervised fine-tuning data to support reasoning over noisy retrieval results.

MEL considers not only whether individual observations are trustworthy, but also whether they jointly support the decision at hand. We therefore incorporate actual tool observations and the original audio into decision-model training. Through evidence applicability supervision, we guide the model to organize and weigh evidence within the full audio context and understand what it supports and which inferences remain unjustified, connecting observation-level verification with conclusion-level judgment.

3 Method

Given an audio recording x and a textual question q, Multimodal Evidence Learning (MEL) connects question-guided acoustic evidence acquisition with task-oriented evidence use. As illustrated in Figure 2, MEL decomposes the information needs of a question into textual probes and uses acoustic tools to obtain relevant observations. It then combines these observations with the original audio and question to determine how they support task conclusions. The learning objectives follow these connections: acquiring evidence that addresses decision requirements and relating actual observations to justified decisions. The original audio remains available throughout final decision-making, while targeted observations complement holistic understanding.

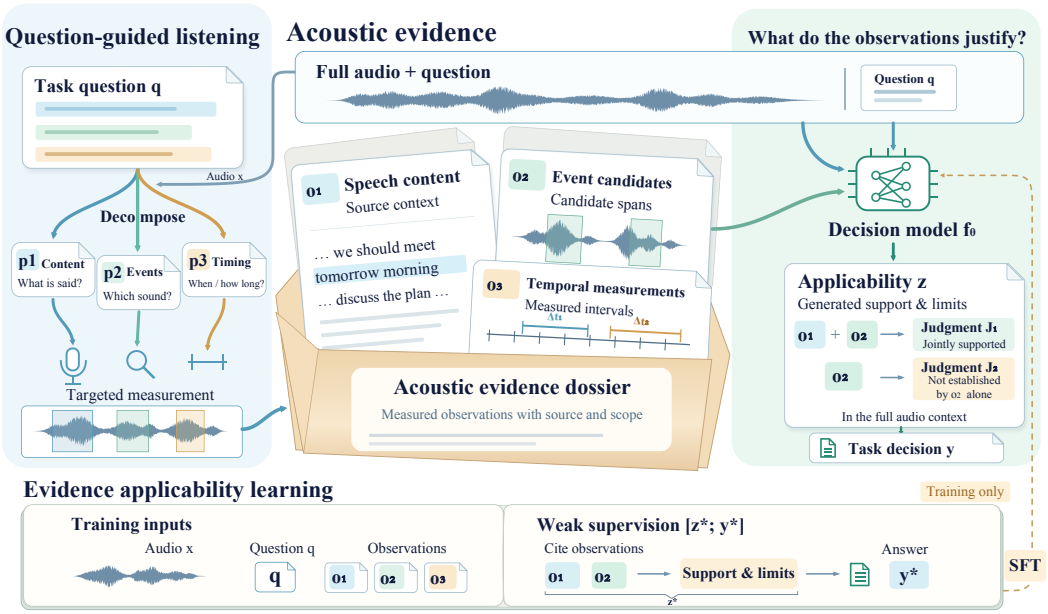

Figure 2: Overview of Multimodal Evidence Learning (MEL). Given the audio and a task question, MEL decomposes the judgment requirements into text probes that guide acoustic measurements. The resulting observations retain their source context and scope. Together with the original audio and question, these observations inform the decision model, which generates an evidence applicability explanation and the task output. During training, evidence applicability supervision teaches the model to relate observations to the judgments they support and the inferences they do not establish.

3.1 Question-Guided Acoustic Evidence Acquisition

MEL treats question decomposition as an elaboration of evidence needs. Considering the question, audio context, and available tool capabilities, the model identifies sound facts and relations needed to distinguish candidate answers and expresses them as textual probes. These probes specify the content, attributes, or temporal scope to inspect, translating "what to answer" into "what to listen for."

Let P denote the decomposition and acquisition-planning model and $\mathcal{T}$ the available acoustic operations. Measurement requests and their resulting observations are represented as

$$\begin{aligned} \mathcal{R} &= P(x, q, \mathcal{T}), \\ o_i &= T_{t_i}(x; a_i), \qquad r_i = (p_i, t_i, a_i) \in \mathcal{R}, \end{aligned} \tag{1}$$

where p_i is a textual probe specifying an evidence need, t_i identifies an operation, and a_i contains its arguments, including the target content and audio scope. The question determines which information matters for the current judgment, the audio context helps identify what to inspect, and tool execution provides the corresponding observations. Decomposition therefore influences which acoustic information is acquired, rather than merely adding linguistic analysis.

Different decision requirements can call for different observation types, including speech content, event candidates, temporal measurements, and audio segments. Speech content, for example, can be inspected through transcription or direct listening, while temporal ordering requires locating and comparing relevant events. Audio segments with surrounding context allow closer inspection of local content. MEL does not require every observation to become a textual report; it preserves the information form appropriate for the judgment.

Each observation remains associated with its request and source scope, forming $\mathcal{E} = \{(r_i, o_i)\}_{i=1}^{m}$. This organization allows the decision model to identify which information need an observation addresses and which part of the recording it covers. Requests describe inspection targets rather than established facts, and returned observations must be interpreted within their measurement scope.

For example, the boundaries of a selected segment do not automatically specify the onset and offset of an event.

3.2 Task-Oriented Evidence Organization and Use

MEL focuses on the applicability of observations to specific task judgments, rather than accepting a tool report as a whole. An observation may support the presence of a sound without establishing its onset, occurrence count, or causal interpretation. Multiple observations may jointly support one judgment, while the same observation remains insufficient for another. Evidence applicability therefore depends on the relationship between observation content, the condition being assessed, and the full audio context.

The decision model jointly receives the original audio, question, and organized evidence, and generates a textual evidence applicability explanation z. This explanation connects observations to task judgments, identifying supported conditions, unresolved interpretations, and additional information on which a conclusion depends. The model then conditions on this explanation to generate the task output y:

$$p_\theta(z, y \mid x, q, \mathcal{E}) = p_\theta(z \mid x, q, \mathcal{E}) \cdot p_\theta(y \mid x, q, \mathcal{E}, z). \tag{2}$$

Here, z is an intermediate explanation generated before the answer by the same decision model, not an independent verification module or an externally supplied correct explanation at test time. It makes the observation–conclusion relationship explicit and provides context for the final decision. Its content depends on the question and available evidence; individual observations need not provide definitive support.

The original recording supplies the broader context for local observations. When tool outputs do not resolve a condition, the model can still consult the original audio. Observation retention and inferential support remain distinct: an observation need not be discarded because it is insufficient for a particular inference, and failing to obtain an observation of a sound does not establish its absence. Depending on the task, the final output is an answer or a candidate relevance score.

3.3 Evidence Applicability Learning

We formulate evidence applicability learning to teach the relationship between actual observations and task conclusions. Supervision concerns not only the answer, but also the observations underlying a judgment and the extent of their support. Compared with answer-only supervision or rationales without explicit requirements for evidence organization, this design makes support relations and inferential limits part of the learning target.

Given a training example (x, q, y^*) and actual observations $\mathcal{E}$, let z^* denote the corresponding evidence applicability demonstration. A demonstration references task-relevant observations, explains their support or limitations for the required judgments, and grounds the answer in the full audio context. Observation references may point to textual report content or to audio segments and their source scope; audio evidence need not have a complete textual description.

Demonstration construction must preserve the distinction between observations and inferences. Evidence of sound presence can support claims about counts, temporal order, or causality only when the necessary additional grounds are available. Unresolved conditions can remain explicit rather than forcing every observation to justify a definite conclusion. Training answers provide task supervision but must not be used to invent unmeasured properties as observations. The training data and demonstration construction procedure used in our experiments are described in Appendix B.1.

For the decision stage, we define the target sequence as an applicability explanation followed by an answer, $u^* = [z^*; y^*]$, with the standard autoregressive objective:

$$\mathcal{L}_{\text{app}}(\theta) = -\mathbb{E}_{(x,q,\mathcal{E},u^*)\sim\mathcal{D}} \left[\sum_{k=1}^{|u^*|} \log p_\theta(u_k^* \mid x, q, \mathcal{E}, u_{<k}^*) \right]. \tag{3}$$

This objective connects observation references, support and limitation statements, and task answers within a single generation process, without requiring an additional reliability classification head or a separate evidence filter.

Decomposition, acquisition planning, and final decision-making can use the same audio-language backbone, while their supervision targets serve different roles: the former concern question decomposition and evidence acquisition, and the latter concern observation interpretation and use. The framework does not require the decomposer to remain fixed. Training configurations and parameter update scopes are specified in Appendix B.2.

At inference time, the model formulates acquisition requests from the audio and question, tools return actual observations, and the decision model combines these observations with the original recording to generate the applicability explanation and task output. Question-guided acquisition determines which evidence to obtain, while evidence applicability learning addresses which judgments that evidence can support.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets. We evaluate MEL on audio understanding and reasoning benchmarks, MMAR (Ma et al., 2026) and MMAU (Sakshi et al., 2025). MMAR contains 1,000 questions from real-world recordings, covering sound, music, speech, and their mixtures. It emphasizes multi-step reasoning that combines acoustic details, contextual information, and domain knowledge. MMAU covers perception and reasoning across sound, music, and speech. We use its 1,000-question test-mini split from version v05.15.25. Both benchmarks use a multiple-choice question-answering format.

Evaluation metrics. We report overall and category-specific accuracy. For MMAR, the categories are sound, music, speech, and mixed audio, with the mixed category pooling examples from different modality combinations. For MMAU, we report sound, music, and speech results. Overall accuracy is computed over all 1,000 examples in each benchmark, rather than as an unweighted average of category accuracies. Failures to generate or parse a valid answer are counted as incorrect and remain in the evaluation denominator.

Models. We use Qwen2.5-Omni-7B (Xu et al., 2025a) and Qwen3-Omni-30B-A3B-Instruct (Xu et al., 2025b) as audio-language backbones to examine MEL across model scales and architectures. The inputs consist of the original audio, question, and answer options. MEL acquires question-relevant acoustic evidence and provides it alongside the original recording for final decision-making.

Baselines. We include direct answering (Direct) and chain-of-thought reasoning (CoT) as basic controls. We also compare against MATCH (Kuan & Lee, 2025), ThinkOmni (Guan et al., 2026), MATA (Wang et al., 2025), Self-Refine (Madaan et al., 2023), TwS (Xiong et al., 2025), EChO-Agent (Zhang et al., 2026b), and AGR (Rong et al., 2025). These methods cover description-assisted reasoning, reasoning enhancement, iterative refinement, and tool-assisted reasoning. We implement the baselines using their published descriptions and available code and configurations, with task- and backbone-specific adaptations that retain their core mechanisms.

Model and training configurations, audio preprocessing, and evidence acquisition settings are provided in the appendix. Evaluation settings, generation parameters, and baseline implementation details, including input ordering, auxiliary models, and tool configurations, are described in Appendix C. Full prompt templates are provided in the accompanying code.

4.2 EXPERIMENTAL RESULTS

Overall performance. As shown in Table 1, MEL achieves the highest overall accuracy among the compared methods on both benchmarks with both backbones. With Qwen2.5-Omni-7B, MEL reaches 57.1% on MMAR and 72.8% on MMAU, improving over Direct by 2.1 and 5.4 percentage points, respectively. With Qwen3-Omni-30B-A3B-Instruct, it achieves 64.0% and 73.8%, corresponding to improvements of 1.5 and 2.9 percentage points. Compared with the strongest non-MEL

Table 1: Accuracy (%) on MMAR and MMAU. Bold indicates the best result in each column within each backbone group. Dashes indicate unavailable results.

Model	Method	MMAR					MMAU			
		Sound	Music	Speech	Mixed	Overall	Sound	Music	Speech	Overall
Qwen2.5-Omni-7B	Direct	53.9	47.1	56.1	59.4	55.0	73.3	63.8	65.2	67.4
	CoT	53.9	44.7	55.4	57.9	53.8	73.3	65.3	70.9	69.8
	MATCH	**58.8**	**51.5**	56.8	57.6	56.3	73.3	62.9	66.7	67.6
	ThinkOmni	56.4	40.8	51.7	54.3	51.1	69.4	64.4	60.4	64.7
	MATA	49.7	38.3	59.9	52.8	51.4	73.9	63.2	64.9	67.3
	Self-Refine	45.5	35.9	52.4	45.4	45.5	64.0	63.5	62.5	63.3
	TwS	46.7	35.4	46.3	49.0	45.0	71.8	63.2	64.6	66.5
	EChO-Agent	39.4	28.6	62.6	53.4	48.7	73.0	47.0	66.4	62.1
	AGR	**58.8**	42.7	**67.3**	55.2	56.8	73.0	65.6	71.2	69.9
	MEL (ours)	56.4	50.5	56.8	**61.8**	**57.1**	**77.8**	**68.9**	**71.8**	**72.8**
Qwen3-Omni-30B -A3B-Instruct	Direct	57.0	51.9	68.7	66.3	62.5	**77.8**	67.1	67.9	70.9
	CoT	60.0	47.1	69.4	66.9	62.4	76.6	67.7	67.0	70.4
	MATCH	59.4	**52.9**	**70.7**	65.7	63.5	72.1	67.1	69.1	69.4
	ThinkOmni	45.5	38.8	54.4	51.6	48.8	72.4	63.8	65.5	67.2
	MATA	32.1	33.0	56.8	43.3	43.3	70.9	59.3	63.1	64.4
	Self-Refine	44.8	41.7	59.2	54.6	51.7	65.5	57.5	61.9	61.6
	TwS	41.8	42.7	52.0	46.0	46.4	61.3	56.6	60.4	59.4
	EChO-Agent	46.7	42.7	69.4	63.0	58.0	70.0	59.9	73.0	67.6
	AGR	46.1	42.2	66.3	58.2	55.3	69.7	61.7	**75.1**	68.8
	MEL (ours)	**64.8**	51.0	68.0	**68.1**	**64.0**	77.2	**72.2**	72.1	**73.8**

baseline in each setting, MEL improves MMAR accuracy by 0.3 and 0.5 percentage points for the two backbones, respectively, and MMAU accuracy by 2.9 percentage points for both. These results support combining question-relevant acoustic evidence with holistic audio understanding, with larger gains on MMAU in the current evaluation.

Comparison with existing methods. The main comparison covers several approaches to audio reasoning, including description-assisted reasoning such as MATCH, reasoning enhancements such as CoT and ThinkOmni, iterative refinement through Self-Refine, and tool-assisted methods such as EChO-Agent and AGR. These methods perform well in some audio categories, but their overall benefits vary across backbones and benchmarks. In contrast, MEL achieves the best overall accuracy in all four backbone–benchmark combinations, demonstrating more consistent overall competitiveness across the evaluated settings. Its aggregate advantage coexists with category-specific strengths of other methods, rather than requiring MEL to dominate every individual category.

Performance across audio categories. With the 7B backbone, MEL achieves the highest accuracy on all three MMAU categories, improving over Direct by 4.5, 5.1, and 6.6 percentage points on sound, music, and speech, respectively. On MMAR, MEL achieves the highest mixed-category accuracy with both backbones, while the 30B model improves sound accuracy over Direct by 7.8 percentage points. However, the gains are not uniform: some baselines remain stronger on individual music or speech categories, and 30B MEL performs slightly below Direct on MMAU sound. These differences indicate that the benefits of targeted evidence depend on the audio content and the distinctions required by the task.

4.3 Ablation Analysis

Ablation settings. Table 2 compares full MEL with four variants on the same evaluation examples. *w/o evidence* removes the selected segments and their accompanying information, retaining only the original audio, question, and answer options; it is equivalent to Direct under the same final answering prompt. *w/o decomposition* selects segments directly from the original question and

Table 2: Ablation results on MMAR and MMAU in accuracy (%). Bold indicates the best result in each column within each backbone group.

Model	Method	MMAR					MMAU			
		Sound	Music	Speech	Mixed	Overall	Sound	Music	Speech	Overall
Qwen2.5-Omni-7B	w/o evidence	53.9	47.1	56.1	59.4	55.0	73.3	63.8	65.2	67.4
	SFT-Direct	53.9	47.6	56.5	60.3	55.5	72.7	63.2	64.0	66.6
	w/o decomposition	**59.4**	46.1	56.8	60.3	56.2	77.5	67.4	67.0	70.6
	w/o SFT	52.1	48.5	**57.5**	60.9	55.9	75.4	68.3	65.5	69.7
	MEL (full)	56.4	**50.5**	56.8	**61.8**	**57.1**	**77.8**	**68.9**	**71.8**	**72.8**
Qwen3-Omni-30B -A3B-Instruct	w/o evidence	57.0	51.9	68.7	66.3	62.5	77.8	67.1	67.9	70.9
	SFT-Direct	57.6	**54.9**	68.0	66.0	62.9	76.3	68.3	67.0	70.5
	w/o decomposition	63.0	53.9	67.0	**69.3**	64.4	**78.7**	71.0	68.5	72.7
	w/o SFT	**64.8**	51.0	**70.4**	68.7	**64.9**	76.0	71.6	**72.1**	73.2
	MEL (full)	**64.8**	51.0	68.0	68.1	64.0	77.2	**72.2**	**72.1**	**73.8**

audio, without the initial audio index or structured subquestion planning, while retaining SFT. *w/o SFT* uses the original model weights for decomposition and segment selection, keeping the extraction rules and final answering procedure unchanged. *SFT-Direct* uses the same SFT checkpoint for direct answering without the acquisition branch, testing whether the training benefits transfer directly to answering without evidence acquisition.

Contribution of acoustic evidence. Full MEL outperforms *w/o evidence* on both benchmarks with both backbones, with gains of 5.4 and 2.9 percentage points on MMAU for 7B and 30B, respectively. This comparison retains the same final answering model and prompt, isolating the overall effect of supplying the acquisition branch outputs. Because the original recording remains available in both settings, the results support using targeted observations to supplement, rather than replace, holistic audio understanding.

Contribution of decomposition. With the 7B backbone, full MEL improves over *w/o decomposition* by 0.9 percentage points on MMAR and 2.2 on MMAU. With 30B, it improves MMAU accuracy by 1.1 percentage points but falls below direct segment selection by 0.4 points on MMAR. Explicitly organizing a question's information needs therefore benefits evidence acquisition in several settings, although its additional value varies by task. Since this ablation removes both the initial indexing step and structured planning, the comparison measures their combined contribution rather than the effect of adding subquestion text alone.

Contribution of SFT. Compared with *w/o SFT*, full MEL improves 7B accuracy by 1.2 percentage points on MMAR and 3.1 on MMAU. For 30B, SFT improves MMAU accuracy by 0.6 points but reduces MMAR accuracy by 0.9 points. The training benefits are thus more consistent for the 7B backbone in the current evaluation. Moreover, SFT-Direct improves over Direct by only 0.5 and 0.4 points on MMAR for 7B and 30B, respectively, while reducing MMAU accuracy by 0.8 and 0.4 points. Full MEL exceeds SFT-Direct in all four settings, including gains of 6.2 and 3.3 points on MMAU. These results suggest that the value of the current training is better understood in conjunction with evidence acquisition than as a general improvement in direct answering.

Backbone-dependent effects. Full MEL achieves the highest overall accuracy among the ablation settings in three of the four backbone–benchmark combinations. In particular, the 7B results consistently favor the complete pipeline over removing decomposition, evidence, or SFT. The 30B results are more task-dependent: full MEL remains strongest on MMAU, whereas *w/o decomposition* and *w/o SFT* reach 64.4% and 64.9% on MMAR, compared with 64.0% for the full model. One possible explanation is that the stronger backbone already possesses sufficient question understanding and audio localization ability to obtain useful evidence through a simpler acquisition procedure on some tasks, leaving less room for additional decomposition or the current supervision to help. Architecture, pretraining, and task characteristics may also contribute to this difference.

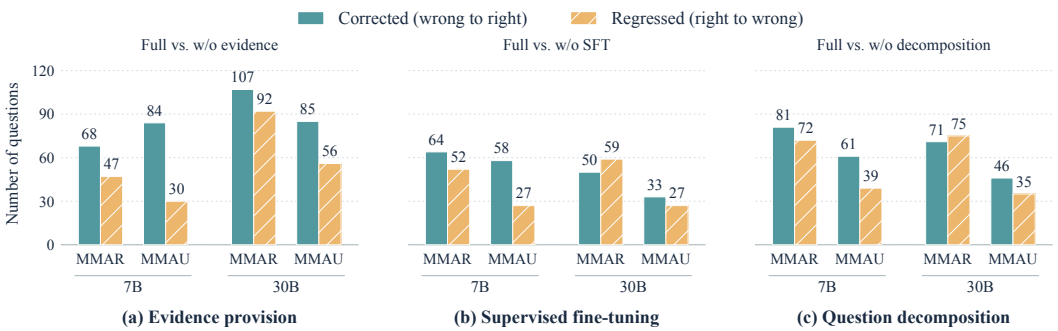

Figure 3: Paired answer changes between MEL and its ablated variants. Each panel compares the full method with the indicated control on the same 1,000 questions per benchmark and backbone. Corrected counts questions answered incorrectly by the control but correctly by MEL; regressed counts the reverse. The difference between the two counts determines the net accuracy change.

Overall, the complete design provides strong aggregate performance across the evaluated settings, particularly with the 7B backbone, without implying that every component independently improves every task. The 30B results further show that simpler acquisition procedures can remain effective for some model–task combinations.

4.4 Answer Corrections and Regressions

To understand the changes underlying aggregate performance, we compare predictions of full MEL with those of each ablated variant on the same questions. Figure 3 summarizes paired comparisons across both benchmarks and backbones. We distinguish corrections, where an incorrect answer becomes correct, from regressions, where a previously correct answer becomes incorrect.

Corrections and regressions behind net gains. Corrections outnumber regressions in most comparisons, but both types of transitions occur frequently. For example, on MMAU with the 7B backbone, full MEL corrects 84 answers while regressing on 30 previously correct answers relative to the variant without evidence, yielding a net gain of 54 correct answers. On MMAR with the 30B backbone, the corresponding counts are 107 corrections and 92 regressions, leaving a net gain of only 15. Thus, similar aggregate improvements need not reflect similar prediction behavior, and a modest net accuracy gain can conceal substantial changes in individual answers.

Remaining room for improvement. These results suggest that the benefits of MEL are partly offset by regressions on previously correct answers, leaving room for further improvement. Although answer transitions alone do not identify the source of these regressions, more targeted question decomposition, more relevant and complete evidence, and better assessment of its applicability may help preserve corrections while reducing regressions. The observed gains therefore coexist with remaining challenges in consistently translating acquired evidence into reliable decisions.

5 Conclusion

This work studies how question-guided acoustic evidence can contribute to reliable audio reasoning. Experiments on MMAR and MMAU show that MEL improves overall accuracy over direct answering with both backbones. The ablations reveal that these gains depend on how evidence acquisition and learning are combined: the full method consistently outperforms its ablated variants with the 7B backbone, while individual components do not yield uniform improvements with the 30B backbone. Per-question analysis further shows that aggregate gains reflect a balance between answer corrections and regressions. Together, these findings highlight both the value of task-relevant evidence and the remaining difficulty of using it consistently. Reliable audio reasoning requires not merely additional observations, but an appropriate connection between what is observed and what can be concluded.

AI USE STATEMENT

In this work, we used generative AI tools to assist with synthetic data and training supervision construction, refinement of conceptual frameworks and research hypotheses, experimental design and execution, method implementation, data cleaning and formatting, qualitative feedback analysis, result interpretation, and manuscript translation. We did not use generative AI tools to formulate mathematical claims requiring proof, provide critical ingredients for mathematical proofs, or write proofs; formal theoretical model development and related proof tasks are not applicable to this work. Additionally, we used generative AI tools for literature search and summarization, code editing, manuscript structuring, drafting and polishing, and scientific figure ideation and draft design. The authors review all AI-assisted content included in the final work, including checking the cited literature, verifying code and experimental procedures, and examining result interpretations and research conclusions against experimental records. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide the code for our experiments in an anonymous repository at `https://anonymous.4open.science/r/MEL-A187/`. Our method and experimental setup are described in the main text to support reproduction of our approach and results.

REFERENCES

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=hSyW5go0v8.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen2-audio technical report, 2024. URL https://arxiv.org/abs/2407.10759.

Yiran Guan, Sifan Tu, Dingkang Liang, Linghao Zhu, Jianzhong Ju, Zhenbo Luo, Jian Luan, Yuliang Liu, and Xiang Bai. Thinkomni: Lifting textual reasoning to omni-modal scenarios via guidance decoding. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=pMpCOjzwI1.

Chun-Yi Kuan and Hung-Yi Lee. Can large audio-language models truly hear? tackling hallucinations with multi-task assessment and stepwise audio reasoning. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2025. doi: 10.1109/ICASSP49660.2025.10888384.

Chun-Yi Kuan, Chih-Kai Yang, Wei-Ping Huang, Ke-Han Lu, and Hung-Yi Lee. Speech-copilot: Leveraging large language models for speech processing via task decomposition, modularization, and program generation. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pp. 1060–1067, 2024. doi: 10.1109/SLT61566.2024.10832184.

Ziyang Ma, Yinghao Ma, Yanqiao Zhu, Chen Yang, Yi-Wen Chao, Ruiyang Xu, Wenxi Chen, Yuanzhe Chen, Zhuo Chen, Jian Cong, Kai Li, Keliang Li, Siyou Li, Xinfeng Li, Xiquan Li, Zheng Lian, Yuzhe Liang, Minghao Liu, Zhikang Niu, Tianrui Wang, Yuping Wang, Yuxuan Wang, Yihao Wu, Guanrou Yang, Jianwei Yu, Ruibin Yuan, Zhisheng Zheng, Ziya Zhou, Haina Zhu, Wei Xue, Emmanouil Benetos, Kai Yu, EngSiong Chng, and Xie Chen. MMAR: A challenging benchmark for deep reasoning in speech, audio, music, and their mix. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2026. URL https://openreview.net/forum?id=fgmrBJemlQ.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=S37hOerQLB.

Aivo Olev and Tanel Alumäe. Multi-source evidence fusion for audio question answering, 2026. URL https://arxiv.org/abs/2603.17822.

Yan Rong, Chenxing Li, Dong Yu, and Li Liu. Audiogenie-reasoner: A training-free multi-agent framework for coarse-to-fine audio deep reasoning, 2025. URL https://arxiv.org/abs/2509.16971.

S Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. MMAU: A massive multi-task audio understanding and reasoning benchmark. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=TeVAZXr3yv.

Masao Someki, Chien-yu Huang, Siddhant Arora, Samuele Cornell, Markus Müller, Nathan Susanj, Rupak Vignesh Swaminathan, Grant Strimel, Jing Liu, and Shinji Watanabe. PlanRAG-audio: Planning and retrieval augmented generation for long-form audio understanding. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 26167–26183, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.1304. URL https://aclanthology.org/2026.findings-acl.1304/.

Tejas Srinivasan, Jack Hessel, Tanmay Gupta, Bill Yuchen Lin, Yejin Choi, Jesse Thomason, and Khyathi Chandu. Selective "selective prediction": Reducing unnecessary abstention in vision-language reasoning. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 12935–12948, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.767. URL https://aclanthology.org/2024.findings-acl.767/.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun MA, and Chao Zhang. SALMONN: Towards generic hearing abilities for large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=14rn7HpKVk.

Siqian Tong, Xuan Li, Yiwei Wang, Baolong Bi, Yujun Cai, Shenghua Liu, Yuchen He, and Chengpeng Hao. Autagent: A reinforcement learning framework for tool-augmented audio reasoning, 2026. URL https://arxiv.org/abs/2602.13685.

Junyu Wang, Ziyang Ma, Zhengding Luo, Tianrui Wang, Meng Ge, Xiaobao Wang, and Longbiao Wang. Pay more attention to audio: Mitigating imbalance of cross-modal attention in large audio language models, 2025. URL https://arxiv.org/abs/2509.18816.

Yucheng Wang, Jing Peng, Hanqi Li, Chenghao Wang, Wenming Tu, Yu Xi, Zhaokai Sun, Kai Yu, and Shuai Wang. Audio-mind: An auditable agentic framework for audio understanding, 2026. URL https://arxiv.org/abs/2605.28480.

Zhepei Wei, Wei-Lin Chen, and Yu Meng. InstructRAG: Instructing retrieval-augmented generation via self-synthesized rationales. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=P1qhkp8gQT.

Gijs Wijngaard, Elia Formisano, Michel Dumontier, and Jenia Jitsev. Audiotoolagent: An agentic framework for audio-language models, 2026. URL https://arxiv.org/abs/2510.02995.

Zhen Xiong, Yujun Cai, Zhecheng Li, Junsong Yuan, and Yiwei Wang. Thinking with sound: Audio chain-of-thought enables multimodal reasoning for large audio-language models, 2025. URL https://openreview.net/forum?id=nn1CV8VXcI.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-omni technical report, 2025a. URL https://arxiv.org/abs/2503.20215.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, Baosong Yang, Bin Zhang, Ziyang Ma, Xipin Wei, Shuai Bai, Keqin Chen, Xuejing Liu, Peng Wang, Mingkun Yang, Dayiheng Liu, Xingzhang Ren, Bo Zheng, Rui Men, Fan Zhou, Bowen Yu, Jianxin Yang, Le Yu, Jingren Zhou, and Junyang Lin. Qwen3-omni technical report, 2025b. URL https://arxiv.org/abs/2509.17765.

Honglei Zhang, Yuting Chen, Chenpeng Hu, Siyue Zhang, and Yilei Shi. Reasonaudio: A benchmark for evaluating reasoning beyond matching in text-audio retrieval, 2026a. URL https://arxiv.org/abs/2605.03361.

Siyuan Zhang, Jian Zong, Junyu Wang, Peiyuan Jiang, Jiahao Yan, Jingyu Zhang, Tianrui Wang, Xiaobao Wang, Longbiao Wang, and Jianwu Dang. Echo-agent: Evidence chain orchestration agent for audio reasoning, 2026b. URL https://arxiv.org/abs/2606.15141.

A MEL IMPLEMENTATION DETAILS

A.1 QUESTION-GUIDED EVIDENCE ACQUISITION

MEL connects the information needs of a question to the original audio through audio indexing, question decomposition, and targeted segment selection. The model first constructs a chronological audio index without access to the question or answer options. The index contains up to eight contiguous intervals, describing their acoustic content and uncertainties to support subsequent localization.

The model then receives the original audio, question, answer options, and index, and produces up to four information needs and three candidate segments. Each need specifies a sound target, attribute, or relation; selected segments are associated with the corresponding need identifiers. The model may choose not to provide additional segments. Otherwise, it specifies start and end times in the coordinate system of the original recording. Both the index and the selection plan are model-generated, and segment selection retains access to the original audio.

In these experiments, evidence takes the form of context-expanded audio segments extracted by a deterministic cropping procedure, with their temporal metadata retained. The final answering input includes these segments and metadata, the question and options, and the complete original recording. The index and decomposition guide segment selection rather than being supplied to the answerer as an additional report.

A.2 CONTEXT EXPANSION AND INPUT ORGANIZATION

To preserve local context, each selected interval is expanded around its center to approximately 1.5 times its selected duration, subject to a maximum of 60% of the recording duration per segment. Windows near recording boundaries are shifted to remain within the valid range. Segments are ordered by start time and may overlap; no additional cap is imposed on their combined duration.

Cropping preserves the original sampling rate. Multichannel audio is averaged into a mono waveform, and segments are stored as floating-point WAV files before processing by the corresponding model's audio processor. Predicted intervals define the selected context, rather than annotations of precise event boundaries.

During indexing and planning, the user input places the original audio before the text. Final answering receives segment metadata, segment audio, the question and options, and the full recording, in that order. When no valid segments are available, only the question, options, and original recording are retained. A malformed plan receives at most one format-repair attempt without access to the reference answer. Questions without valid evidence remain in the answering and evaluation sets.

The answering prompt instructs the model to assess whether observations support the distinctions required by the question in the full audio context, without treating missing observations as evidence of absence. Answers use the format <answer>X</answer>.

Table 3: Training data statistics.

Statistic	Count
Candidate audio sources	1,000
Retained training audio sources	944
Audio indexing demonstrations	944
Question-conditioned planning demonstrations	303
Total SFT records	1,247

Table 4: Training configuration shared by Qwen2.5-Omni-7B and Qwen3-Omni-30B-A3B-Instruct.

Setting	Value
Trainable modules	LoRA on attention q/k/v/o projections in the Thinker text layers
LoRA rank / alpha / dropout	16 / 32 / 0.03
Other backbone parameters	Frozen
Learning rate	2×10^{-5}
Weight decay	0.01
Learning-rate schedule	Cosine
Warmup ratio	0.05
Per-device batch size	1
Number of devices	4
Gradient accumulation steps	4
Global batch size	16
Training steps	78
Maximum sequence length	32,768
Precision	BF16
Random seed	42
Checkpoint	Final step

B Training data and configuration

B.1 Training data

The candidate training pool is constructed from existing audio question-answering data and contains 1,000 audio sources. Screening excludes identified overlaps and source-linked matches with MMAR and MMAU evaluation audio.

We use the MiMo-V2.5 service offline to construct demonstrations of audio indexing and evidence acquisition planning. After checks of format, temporal ranges, and weak evidence support, we retain 944 indexing records and 303 planning records, totaling 1,247 records from 944 audio sources (Table 3). An audio source can contribute both types of records. Teacher judgments are used only for weak-supervision construction and filtering; the teacher does not provide guidance during benchmark evaluation.

B.2 Training configuration

Both backbones use the same supervision data, with LoRA parameters trained separately for each backbone. The adapted weights are used for indexing and evidence acquisition planning; final answering retains the original weights of the corresponding backbone. Table 4 summarizes the training configuration.

Table 5: Number of evaluation questions by audio category.

Dataset	Sound	Music	Speech	Mixed	Total
MMAR	165	206	294	335	1,000
MMAU test-mini	333	334	333	–	1,000

C Evaluation protocol and baselines

C.1 Evaluation data and scoring

MMAR and MMAU test-mini each contain 1,000 questions. We use MMAU version v05.15.25. Table 5 gives the category counts. Reference answers are used only for offline scoring and are not included in indexing, planning, or answering requests during benchmark evaluation.

The MMAR mixed category pools sound–speech, music–speech, sound–music, and three-way mixtures. Category and overall accuracies are computed using the corresponding question counts. Missing or invalid answers, parsing failures, and execution failures are counted as incorrect. MMAR scoring compares option letters; MMAU maps valid options to text and applies the matching rules of the corresponding dataset version.

For final answering, MEL uses temperature 0, random seed 42, a maximum context length of 32,768 tokens, and a maximum generation length of 2,048 tokens. Top-p and the repetition penalty are both 1. Indexing and planning have generation limits of 4,096 and 2,048 tokens, respectively. Final answers are extracted by rules, without an additional model solving the question again.

One MMAR question is presented with four options in MEL, its matched Direct control, and the ablations, whereas some baselines retain its six original options.

C.2 Baseline implementations

Baselines are implemented from published descriptions, available code, and configurations, with adaptations for audio multiple-choice answering. We summarize their implementations below. Full prompts, input ordering, and generation budgets are specified in the accompanying code.

Direct and CoT. Direct uses the same final answering prompt as MEL, with only the question, options, and original audio, and does not execute evidence acquisition. CoT requests reasoning before the final option. The 7B version uses the baseline template, while the 30B version modifies the answering instruction in the matched Direct prompt.

MATCH. MATCH first obtains an audio description and then answers using the description, question, options, and original audio. Evaluation reuses cached descriptions.

ThinkOmni. ThinkOmni uses text-model-guided contrastive decoding with Qwen3-8B as the auxiliary text model. The 7B implementation is based on the authors' code, and the 30B implementation uses a local decoding implementation.

MATA. MATA adjusts attention to audio tokens following the published mechanism. We scale the corresponding raw attention scores by 1.1 in layers 10–20.

Self-Refine. Self-Refine performs initial answering followed by up to two rounds of feedback and revision. For 7B, outputs that fail the original parser receive an additional answer-extraction step using DeepSeek-0731. This extractor receives only the original output and answer options, not the question, audio, or reference answer. The 30B version does not use this postprocessing.

TwS. TwS uses the authors' tool engine and 21 digital signal processing operations, with the task instruction adapted for multiple-choice answering. We allow at most five model calls.

EChO-Agent. We reimplement question routing, tool measurement, evidence organization, candidate generation, and verification, allowing at most one correction. The implementation uses a DeepSeek text coordinator together with Whisper, YAMNet, and emotion and music analysis components.

AGR. AGR adapts the authors' iterative reasoning procedure, combining Whisper transcription, DeepSeek text coordination, and Qwen audio understanding, with at most three iterations.

Each method retains its reasoning or tool-use mechanism; generation budgets and auxiliary components are not forced to be identical. ThinkOmni and EChO-Agent use different input orders across the two backbones. Detailed input organization and auxiliary configurations are given in the accompanying code.

C.3 Ablation settings

Without evidence. The answerer receives no outputs from the acquisition branch and retains only the original audio, question, and options. This condition corresponds to Direct with the matched answering prompt.

Without decomposition. We remove the audio index and information-need decomposition. The same adapted model selects segments directly from the original question and audio. This comparison measures the combined contribution of indexing and structured planning.

Without SFT. Indexing and segment selection use the original model weights, while segment extraction rules and the final answerer remain unchanged.

SFT-Direct. The same adapted checkpoint answers questions directly without evidence acquisition. It is not a separately trained answer-supervision model.

The full method, the variant without decomposition, and the variant without SFT share the same final answerer and cropping rules, although their selected segments may differ.

D Paired answer analysis and examples

D.1 Paired statistics

Figure 3 pairs full MEL with the variants without evidence, without SFT, and without decomposition by question identifier. Each backbone–benchmark comparison includes all 1,000 questions. A correction occurs when the control is incorrect and MEL is correct; a regression occurs when the control is correct and MEL is incorrect. We also count questions on which both methods are correct or both are incorrect. These four counts sum to 1,000. Both-incorrect cases need not have the same predicted option.

Let $N_{\text{corrected}}$ and $N_{\text{regressed}}$ denote corrections and regressions, respectively, and let N denote the number of questions. The net accuracy change, in percentage points, is

$$\Delta\text{Acc} = 100\,\frac{N_{\text{corrected}} - N_{\text{regressed}}}{N}. \tag{4}$$

Execution and parsing failures are treated as incorrect. All transitions are defined from the control condition to full MEL.

For the 30B MMAR comparison against the variant without evidence, corrections, regressions, both-correct cases, and both-incorrect cases number 107, 92, 533, and 268, respectively, yielding a net gain of 1.5 percentage points. This analysis describes answer changes rather than independently assessing the correctness of acoustic evidence.

D.2 Illustrative examples

Table 6 presents a correction and a regression from the 7B comparison between full MEL and the condition without evidence. The original model outputs are answer letters. The interpretations

Table 6: Examples of an answer correction and regression with the 7B backbone. Segment times refer to the original recording.

Item	Correction: word identification	Regression: utterance-level counting
Dataset	MMAR	MMAU
Question	Identify the two words recognized as British	From the given utterance, count the number of words that contain at least one unstressed phoneme
Options	A. chips and water; B. fries and juice; C. crisps and soda; D. tea and lemonade	A. twelve; B. one; C. eighteen; D. nine
Reference	A	D
Supplied segment	9.5–12.5 s	0–2.928 s
Without evidence	C	D
Full MEL	A	B

below are based on the task requirements and supplied audio ranges, rather than model-generated rationales or independent acoustic annotations.

The word-identification example shows an incorrect answer becoming correct when local audio is supplied. The counting example illustrates that adding a local segment does not guarantee improvement: a local observation need not cover all information required for an utterance-level count. These examples illustrate answer transitions, but the records alone do not establish their specific acoustic causes.