

Overtone: Fine-Grained Audio Reasoning via Diagnostic-Driven Delayed Acoustic Fusion

Zhuoyan Mi, Kangle Sun, Yikai Yang, Yujie Fu, Yuehai Wang*

Zhejiang University, Hangzhou, China

{22560261, 22531091, 22531108, 22547137, wyuehai}@zju.edu.cn

Abstract

Audio large language models (LLMs) have advanced rapidly in general audio understanding, yet consistently underperform on tasks requiring fine-grained acoustic discrimination. We investigate this gap through a systematic probing study, uncovering a pattern we term *representation reorganization*: in the LLM’s middle layers, semantic discriminability increases while fine-grained acoustic features such as speaker identity and pitch become progressively less recoverable by linear probes. Informed by this diagnosis, we propose Overtone, a delayed acoustic fusion architecture that bypasses the identified compression zone. Overtone processes identified audio segments through dual tracks—a semantic track from the LLM’s audio encoder and an acoustic track from a frozen self-supervised speech model—and injects acoustic features via gated cross-attention at post-compression layers. The model learns to autonomously decide when and where re-listening is needed during inference. Overtone achieves state-of-the-art performance on MMAU among open-source models at comparable scale and competitive results on MMAR, with gains concentrated in speech and music. Post-training probing confirms recovery of acoustic structure in the reorganized layers.

1 Introduction

Audio is among the richest information modalities in human communication, encoding not only linguistic content but also speaker identity and other paralinguistic cues (Schuller et al., 2011). Audio large language models (LLMs) have made remarkable strides in harnessing this richness, unifying speech, sound, and music comprehension within a single framework (Xu et al., 2025a; Comanici et al., 2025). Yet a notable asymmetry persists: while these models excel at semantic tasks such as speech recognition, they struggle

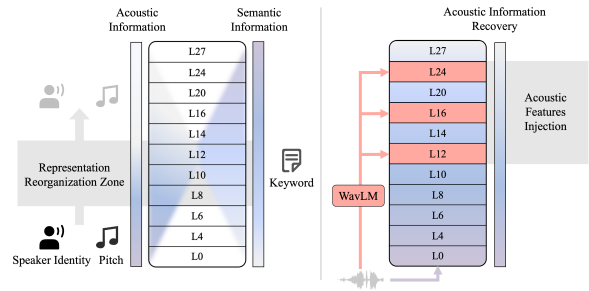


Figure 1: **Left:** In standard audio LLMs, acoustic features (speaker identity, pitch) are progressively compressed in the middle layers while semantic features are constructed. **Right:** Our method bypasses this compression zone by injecting frozen WavLM features, recovering acoustic information where it is needed.

gle with *fine-grained acoustic discrimination*—distinguishing speakers or subtle pitch differences, for instance (Sakshi et al., 2024; Kim et al., 2025). Given that current audio LLMs already operate at substantial scale and are trained on diverse corpora (Radford et al., 2022; Xu et al., 2025b), this raises the question: could the bottleneck lie in the architecture rather than in data or capacity?

We investigate this question through a controlled layer-wise probing study on a representative audio LLM (Figure 1, left). The results reveal a phenomenon that we term *representation reorganization*: the model’s middle layers do not simply degrade acoustic features—they appear to *reallocate* representational capacity from fine-grained acoustics to semantics, as evidenced by converging probing, ablation, and cross-model analyses. Speaker identity and pitch information collapse sharply across a narrow band of layers, while semantic discriminability is simultaneously constructed in the same layers. This reorganization is consistent with the language modeling objective, and it creates a representational bottleneck: any acoustic information that enters the LLM through the standard pathway must traverse these compres-

sion layers, where fine-grained details are systematically overwritten.

This finding reframes the challenge. Recent re-listening approaches (Wu et al., 2026a) allow models to revisit audio segments during reasoning, yet the re-injected signal still traverses the full LLM depth without any mechanism to preserve acoustic features beyond the compression zone, subjecting it to the same representational loss. Reasoning-enhanced methods (Wu et al., 2026b; Rouditchenko et al., 2025) improve deliberation quality through reinforcement learning but operate on representations already stripped of fine-grained detail. Without an architectural mechanism to preserve acoustic features beyond the compression zone, neither strategy addresses the root cause of the limitation.

We propose Overtone, a *diagnostic-driven* architecture that directly responds to the probing findings. When acoustic verification is needed during reasoning, it encodes the segment through two complementary tracks (Figure 1, right): a *semantic track* via the LLM’s own audio encoder, and an *acoustic track* via a frozen self-supervised speech model (Chen et al., 2022) that preserves the fine-grained features the LLM discards. The semantic track enters the LLM normally; the acoustic track is injected via gated cross-attention at post-compression layers, bypassing the identified bottleneck. Combined with reconstruction regularization, this delayed fusion preserves access to high-fidelity acoustic evidence during reasoning.

Our contributions are as follows:

- We identify and characterize *representation reorganization* in audio LLMs, where semantic discriminability increases as fine-grained acoustic information becomes less recoverable.
- We propose *delayed acoustic fusion*, a gated cross-attention bypass guided by the diagnostic findings, paired with reconstruction regularization to stabilize acoustic retrieval during training.
- We introduce learned segment localization and selective re-listening, achieving a mean segment IoU of 0.93.
- Overtone achieves state-of-the-art performance among open-source $\leq 8\text{B}$ audio LLMs on MMAU and competitive results on MMAR, with the largest gains in speech and music—the domains most dependent on fine-grained acoustic cues.

2 Related Work

Audio Language Models. Early audio processing relied on cascaded ASR-to-LLM pipelines (Zhang et al., 2023), which inherently stripped away vital non-linguistic cues such as emotion and environmental sounds. To mitigate this acoustic information loss, recent Audio LLMs have transitioned to end-to-end architectures that directly integrate continuous audio representations. This shift has driven diverse innovations, ranging from commercial flagships to open-source models tailored for long-context reasoning or parameter efficiency (Comanici et al., 2025; Ghosh et al., 2026; Wu et al., 2025). A parallel line of work explores dual-encoder architectures to decouple semantic and acoustic features (Tang et al., 2024; Hu et al., 2024). However, these approaches typically fuse outputs at the input level and apply both encoders on every forward pass. Consequently, despite the overall architectural advances, current models fail to target specific representational bottlenecks, leaving them susceptible to middle-layer degradation in highly complex acoustic scenarios.

Representation Analysis in Multimodal Models. As multimodal models scale, their internal representations naturally converge toward a shared semantic space (Lu et al., 2025), frequently at the cost of discarding fine-grained, modality-specific details (Chaudhuri et al., 2025). While existing works probe and optimize cross-modal encoding (Alex et al., 2026), layer-wise analysis of how audio LLMs progressively compress acoustic representations remains largely unexplored—a gap our diagnostic study directly addresses.

Dynamic Audio Reasoning. Recent audio LLMs incorporate multi-step reasoning but often suffer from internal text fallback, discarding acoustic details. To force genuine audio utilization, Step-Audio-R1 (Tian et al., 2025) distills logical chains anchored to acoustics, and Audio-Thinker (Wu et al., 2026b) uses adaptive metacognition for compute allocation. Echo (Wu et al., 2026a) advances “thinking with audio” via continuous acoustic engagement. However, these paradigms overlook layer-wise acoustic degradation, implicitly assuming encoded features remain intact throughout depth. Overtone secures this foundation: by bypassing the compression zone via delayed fusion, it resolves the bottleneck

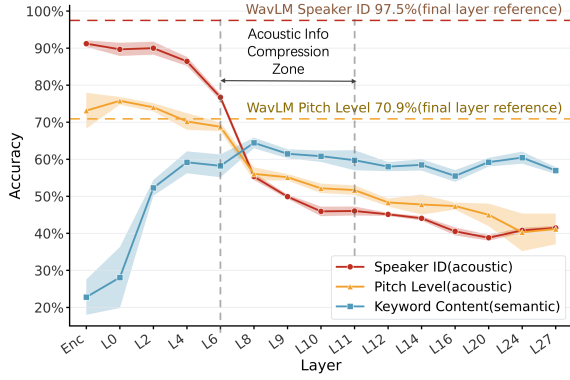


Figure 2: Linear probe accuracy across layers of Qwen2.5-Omni 7B. Fine-grained acoustic features undergo selective compression, with the steepest decline concentrated in layers 6–12, while semantic features are actively constructed—a *representation reorganization*. Dashed lines indicate WavLM-Large accuracy as an acoustic upper bound.

at its architectural source.

3 Diagnostic Study: Representation Reorganization in Audio LLMs

Before introducing our method, we conduct a layer-wise probing study (Hewitt and Liang, 2019) to examine how audio LLMs transform acoustic representations internally. The findings reveal a representation reorganization phenomenon that directly motivates the architecture in Section 4.

3.1 Experimental Setup

We probe 14 representative layers of Qwen2.5-Omni 7B (Xu et al., 2025a), densely sampling the L6–L12 transition zone. Frozen WavLM-Large (Chen et al., 2022) features serve as an external acoustic reference.

We define three probing tasks spanning semantic and fine-grained acoustic attributes: **Speaker ID**, **Pitch Level**, and **Keyword Content**. For each layer, we train a single linear classifier on the extracted hidden states and report classification accuracy. Speaker ID and Pitch Level measure fine-grained acoustic information, while Keyword Content measures semantic discriminability.

All probes are trained on LibriSpeech (Panayotov et al., 2015). Speaker-disjoint splits are used for Pitch Level and Keyword Content to prevent identity leakage.

3.2 Key Findings

Figure 2 reveals a clear reorganization pattern across the middle layers.

Keyword Content accuracy rises sharply from 22.8% at the encoder output to 64.4% by L8, then remains relatively stable through deeper layers. In contrast, Speaker ID accuracy drops by nearly 50 points overall, with the steepest collapse concentrated between L6 and L12. Pitch Level accuracy follows a parallel trajectory. The onset of compression is gradual, beginning around L4, but the steepest concurrent change—where acoustic features decline most sharply while semantic features are constructed—is concentrated in L6–L12. We refer to this region as the core compression zone.

This trade-off becomes clearer when compared with WavLM-Large. While the LLM deep layers substantially lose speaker and pitch information, WavLM largely preserves both attributes, yet achieves much weaker semantic discrimination. The contrast suggests that the compression is objective-driven: the language modeling objective encourages semantic abstraction, whereas WavLM’s self-supervised speech objective preserves fine-grained acoustic structure.

Appendix A.3 shows a similar qualitative trade-off across three additional audio LLMs, suggesting that the bottleneck is not specific to Qwen2.5-Omni.

3.3 Design Implications

The probing results directly motivate three architectural constraints for recovering fine-grained acoustic reasoning.

(a) Injection should begin after the compression zone. Injecting acoustic features before the core compression zone would expose them to the same dynamics identified in Figure 2. We therefore begin delayed acoustic fusion at L12, a choice validated by the injection layer ablation in Section 6.2. The precise compression range may vary across architectures (Appendix A.3), and the injection point should be selected empirically for each base model.

(b) The acoustic source should remain independent of the LLM. Using shallow LLM representations as a bypass would still subject them to downstream reorganization during training. A frozen external encoder such as WavLM avoids this contamination while preserving the acoustic attributes the LLM progressively discards.

(c) Acoustic recovery should rely on adaptive retrieval rather than attribute-specific experts. Different acoustic attributes are compressed at different rates, making exhaustive expert decomposi-

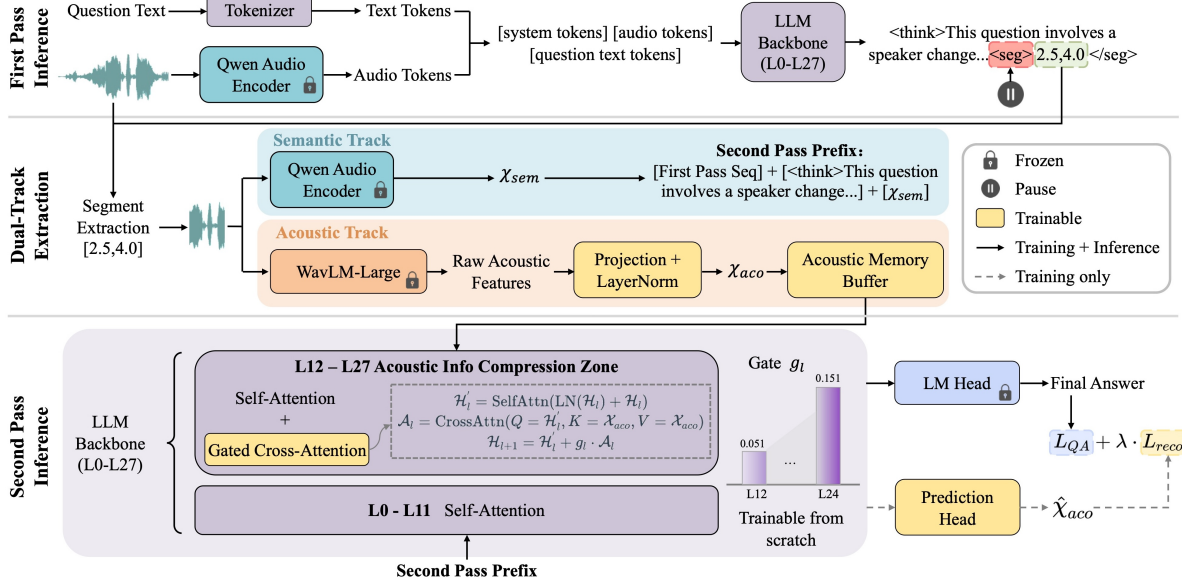


Figure 3: Overtone inference pipeline. The model generates a chain-of-thought with a `<seg>` tag (top), the identified segment is encoded through semantic and acoustic tracks (center), and the second pass injects acoustic features via gated cross-attention at layers 12, 16, 20, 24 (bottom). Dashed elements are active only during training.

tion impractical. A shared acoustic pathway with learned gating allows the model to dynamically retrieve task-relevant acoustic information without manually defining acoustic categories.

4 Method: Overtone

We now present Overtone, an architecture directly motivated by the diagnostic findings of Section 3.

4.1 Overview

Overtone is an *on-demand* mechanism that activates only when the model determines that fine-grained acoustic evidence is needed. The pipeline proceeds in three stages (Figure 3):

First pass. Given a full audio input and a question, the model generates a chain-of-thought (CoT). When acoustic verification is needed, it emits a structured tag `<seg>s, e</seg>` indicating the relevant time interval $[s, e]$.

Dual-track extraction. The segment $[s, e]$ is extracted from the original waveform and encoded through two independent paths: a *semantic track* (X_{sem} , from the Qwen audio encoder) and an *acoustic track* (X_{aco} , from a frozen WavLM-Large). The two tracks carry complementary information by design: the Qwen encoder is pre-trained for semantic alignment, while WavLM’s masked speech prediction objective preserves fine-grained acoustic detail.

Second pass. X_{sem} is appended to the token stream as standard audio tokens and participates in self-attention from L0 onward. X_{aco} is stored in an *acoustic memory buffer* and accessed via gated cross-attention at four designated layers. The model then generates a second-pass answer with access to both semantic context and high-fidelity acoustic evidence.

4.2 Delayed Acoustic Fusion

The acoustic track is extracted via a frozen WavLM (preventing objective-driven distortion) and aligned to the LLM dimension via a trainable projection and LayerNorm: $Z_{\text{aco}} = \text{WavLM}(\text{wav}_{s:e})$, $X_{\text{aco}} = \text{LN}(W_{\text{proj}} Z_{\text{aco}} + b_{\text{proj}})$.

At layers $l \in \{12, 16, 20, 24\}$, we append a gated cross-attention module to the standard self-attention output H'_l . This sparse placement covers the post-compression zone with modest overhead:

$$H'_l = \text{SelfAttn}_l(H_l)$$

$$A_l = \text{CrossAttn}_l(Q=H'_l, K=X_{\text{aco}}, V=X_{\text{aco}})$$

$$\hat{H}_{l,t} = H'_{l,t} + g_{l,t} \cdot A_{l,t}$$

where $g_{l,t} = \sigma(W_l \text{RMSNorm}(H'_{l,t}) + b_l)$ is a token-dependent gate. Building on Alayrac et al. (2022), we initialize $W_l=0$ and $b_l=-5$ to initially preserve pretrained LLM behavior. During training, tokens requiring fine-grained acoustic evidence learn to open the gate, while routine reasoning tokens keep it closed. When acoustic memory

is empty, the gate is forced to zero, skipping cross-attention entirely.

4.3 Reconstruction Regularization

Even with delayed injection, a risk remains: the QA loss alone provides an indirect learning signal for the acoustic bypass. Consequently, the model may develop semantic shortcuts that partially address acoustic tasks without fully utilizing the injected features. We address this with an explicit reconstruction objective:

$$\mathcal{L}_{\text{recon}} = \frac{1}{T} \sum_{t=1}^T \left(1 - \cos(f_{\theta}(H_L^{(t)}), \text{sg}(Z_{\text{aco}}^{(t)})) \right)$$

where T is the acoustic token count, $H_L^{(t)}$ is the layer L hidden state, f_{θ} is a prediction head, and sg denotes the stop-gradient operator. This encourages deep representations to retain information predictive of the WavLM features.

4.4 Learned Re-listening Trigger

Enabling autonomous segment localization and selective re-listening is a non-trivial problem. We decompose this into two training stages with separate objectives.

Stage 1: SFT cold start. We train on supervised examples where each sample pairs a question with an oracle segment annotation. The model learns both the trigger syntax and segment localization. The training loss combines standard next-token cross-entropy with the reconstruction regularization: $\mathcal{L}_{\text{SFT}} = \mathcal{L}_{\text{QA}} + \lambda \cdot \mathcal{L}_{\text{recon}}$.

Stage 2: GRPO refinement. We then apply Group Relative Policy Optimization (GRPO; [Shao et al., 2024](#)) with a composite reward $R = R_{\text{acc}} + R_{\text{format}} + R_{\text{seg}} + R_{\text{penalty}}$. Here, R_{acc} rewards answer correctness, R_{format} rewards valid segment formatting, R_{seg} rewards segment IoU, and R_{penalty} penalizes malformed outputs. This stage refines segment localization beyond SFT supervision without requiring gold chain-of-thought annotations. We set $\lambda = 0$ and drop reconstruction regularization in this stage.

At inference time, a lightweight post-hoc controller decides whether to execute the second pass to avoid unnecessary computational overhead.

5 Experimental Setup

5.1 Training Data

We construct a dataset of approximately 75k multiple-choice QA pairs spanning four audio domains (Table 1). Mixed-modality samples concatenate clips from different domains with short silence gaps, yielding questions that require cross-domain reasoning.

Domain	Sources	Samples
Speech	LibriSpeech VoxCeleb2	31k
Environmental	AudioSet-Strong FSD50K ESC-50	18.8k
Music	MusicCaps AudioSet-Strong	15k
Mixed	Cross-domain concatenation	11k

Table 1: Training data by domain, sourced from LibriSpeech ([Panayotov et al., 2015](#)), VoxCeleb2 ([Chung et al., 2018](#)), AudioSet-Strong ([Hershey et al., 2021](#)), FSD50K ([Fonseca et al., 2022](#)), ESC-50 ([Piczak, 2015](#)), MusicCaps ([Agostinelli et al., 2023](#)).

The data pipeline enforces two quality properties beyond standard filtering:

Cross-model consensus. Qwen2.5-Omni first captions each audio clip, prompting Gemini 2.5 Pro ([Comanici et al., 2025](#)) to generate multiple-choice questions with CoT rationales. Qwen then answers these questions based on the raw audio. A Gemini-based judge ([Zheng et al., 2023](#)) then evaluates each sample on both cross-model answer consensus and CoT-segment consistency, retaining only samples passing both checks for SFT training.

Audio-centricity filtering. To filter out questions answerable via text priors or world knowledge, we re-run Qwen using *silent* audio. Samples answered correctly without audio signal are classified as *weak* and discarded. Only the *strong* subset is retained to focus training on tasks requiring audio-text grounding.

5.2 Implementation Details

Base model and architecture. We build on Qwen2.5-Omni 7B, fine-tuning the LLM backbone via LoRA ([Hu et al., 2022](#)) (rank 16, $\alpha=32$) on all attention and MLP projections. The gated cross-attention modules (layers 12, 16, 20, 24)

Table 2: Comparison on MMAU and MMAR benchmarks. Speech, Sound, and Music are MMAU domain scores. Top three open-source $\leq 8B$ models are **bolded**, underlined, and wavy.

Model	Size	MMAU (%)				MMAR
		Speech	Sound	Music	Avg	Avg
<i>Proprietary</i>						
GPT-4o-Audio (Hurst et al., 2024)	–	69.33	63.20	49.93	60.82	64.09
Gemini-2.5-Pro (Comanici et al., 2025)	–	72.67	70.63	64.77	69.36	70.03
Step-Audio-2 (Wu et al., 2025)	–	72.75	80.60	68.23	73.86	–
<i>Open-source (large)</i>						
Qwen3-Omni (Xu et al., 2025b)	30B	70.70	77.87	66.83	71.80	65.50
Step-Audio-R1 (Tian et al., 2025)	33B	76.50	85.24	71.27	77.67	69.15
<i>Open-source ($\leq 8B$)</i>						
Qwen2.5-Omni (Xu et al., 2025a)	7B	68.90	76.77	67.33	71.00	57.33
Audio-Flamingo-3 (Goel et al., 2025)	8.2B	66.97	75.83	74.47	72.42	58.50
AF-Next-Captioner (Ghosh et al., 2026)	8B	72.13	79.87	75.30	75.76	63.00
MiMo-Audio (Zhang et al., 2025)	7B	68.47	82.58	73.65	74.90	63.60
Omni-R1 (Rouditchenko et al., 2025)	7B	75.80	78.30	70.80	74.97	63.46
Audio-Thinker (Wu et al., 2026b)	7B	76.60	79.03	70.53	75.39	67.25
Echo (Wu et al., 2026a)	7B	77.87	79.62	72.33	76.61	69.99
Overtone (Ours)	7.4B	80.97	77.87	74.13	77.66	69.58

and acoustic projection are trained from scratch. WavLM-Large (Chen et al., 2022) and the Qwen audio encoder remain frozen.

Training. Stage 1 (SFT) trains for 5 epochs with learning rate 5×10^{-6} , effective batch size 64, bf16 precision, and maximum sequence length 2048. The reconstruction weight is $\lambda=0.1$. Stage 2 (GRPO) uses 8 rollouts per prompt, a 0.04 KL coefficient, and temperature 1.0. All experiments are conducted on $8 \times A100$ 80GB GPUs.

5.3 Evaluation Benchmarks

We evaluate on four public benchmarks. **MMAU** (and its **MMAU-mini** split) (Sakshi et al., 2024) covers speech, sound, and music reasoning. **MMAR** (Ma et al., 2025) emphasizes cross-domain multimodal reasoning. **MMAU-Pro** (Kumar et al., 2026) extends MMAU with open-ended QA evaluated by an LLM judge. **MMSU** (Wang et al., 2026) targets fine-grained spoken language understanding across perception and reasoning dimensions. Ablation and analysis experiments use MMAU-mini for computational efficiency.

6 Results and Analysis

6.1 Main Results

Table 2 compares Overtone with state-of-the-art audio understanding models spanning proprietary systems, large open-source models, and recent

reasoning-enhanced approaches at the same parameter scale ($\leq 8B$).

MMAU. Overtone achieves the highest overall MMAU score among all open-source $\leq 8B$ models, surpassing Echo by over 1 pp and even exceeding the 30B Qwen3-Omni. The per-domain breakdown reveals where the gain originates. On Speech, Overtone leads by a substantial margin (+3.1 pp over the second-best model), outperforming every model in the table including proprietary systems. It also ranks second among $\leq 8B$ models on Music, while remaining competitive on Sound.

The larger gains on Speech and Music are consistent with our diagnosis that fine-grained acoustic cues are preferentially compressed, whereas Sound relies more on coarser event-level information.

MMAR. On MMAR, Overtone ranks second among $\leq 8B$ open-source models (within 0.4 pp of Echo) and performs competitively with Gemini-2.5-Pro and the $4 \times$ larger Step-Audio-R1. The +12.3 pp gain over base Qwen2.5-Omni confirms the bypass generalizes beyond MMAU to structurally different benchmarks featuring mixed-modality and multilingual content.

MMAU-Pro and MMSU. Table 3 shows consistent gains on both benchmarks. On MMAU-Pro, Overtone improves overall accuracy by +7.6 pp, including +7.5 pp on Open-Ended QA. On MMSU, the largest gains occur in Phonology Per-

Table 3: Results on MMAU-Pro (Kumar et al., 2026) and MMSU (WANG et al., 2026). MMAU-Pro Open-Ended is scored by an LLM judge. MMSU reports Perception (Per.) and Reasoning (Rea.) across Semantics (Sem.), Phonology (Pho.), and Paralinguistics (Par.) dimensions. Both benchmarks are evaluated with identical inference settings and are not used for training or model selection.

	MMAU-Pro (%)					MMSU (%)						
	Spe.	Snd.	Mus.	Open	Avg	Sem. Per.	Pho. Per.	Par. Per.	Sem. Rea.	Pho. Rea.	Par. Rea.	Ovr.
Qwen2.5-Omni	58.1	47.8	62.7	56.3	54.6	55.4	37.6	39.5	88.0	81.8	52.2	61.0
Overtone	75.6	51.7	69.5	63.8	62.2	61.3	68.0	66.4	87.6	89.0	69.6	75.4
Δ	+17.5	+3.9	+6.8	+7.5	+7.6	+5.9	+30.4	+26.9	−0.4	+7.2	+17.4	+14.4

ception (+30.4 pp) and Paralinguistics Perception (+26.9 pp), while Semantics Reasoning remains stable (87.6% vs. 88.0%). These results further support the benefit of the acoustic bypass on fine-grained perceptual dimensions.

6.2 Ablation Studies

We conduct ablations on MMAU-mini (Table 4), verifying the effectiveness of each component and the generality of the encoder choice. We report Δ relative to 81.5% for consistency across groups.

Delayed fusion with reconstruction loss is essential. Delayed fusion already outperforms early fusion even without reconstruction, consistent with our diagnostic finding that early fusion exposes features to the compression zone. However, adding $\mathcal{L}_{\text{recon}}$ to early fusion *hurts* performance, likely because reconstruction pressure in the compression zone creates conflicting optimization signals. This establishes the delayed bypass as the necessary structural foundation.

WavLM bypass outperforms raw re-listening.

Raw re-listening, which re-injects segments via the standard audio encoder into all LLM layers, scores below even single-pass delayed fusion without reconstruction. In contrast, delayed WavLM bypass directly routes uncompressed acoustic features around the compression zone, achieving 82.3% with oracle segmentation. With learned segment triggers, Overtone closely matches this performance, demonstrating reliable autonomous localization.

Injection layer selection. Group (c) confirms L12 as the optimal fusion start. Performance peaks when fusion begins at L12 and degrades in both directions: earlier fusion remains subject to reorganization dynamics, while later fusion loses its intermediate-layer influence. The precise op-

timal layer may vary across architectures, as the compression zone location differs by model (Appendix A.3); we recommend selecting the injection point via a lightweight layer ablation when applying Overtone to a new base model.

Bypass carries causal information. We further verify the acoustic bypass causality under single-pass settings (Group (d)). Accuracy drops progressively from normal bypass to shuffled features to complete removal of acoustic input. The 9.8 pp gap between normal and no-acoustic conditions proves improvements stem from actual acoustic evidence, not merely additional compute.

Encoder generality. Group (e) disentangles the contributions of the delayed fusion architecture and the specific encoder choice by replacing WavLM with two alternative frozen encoders under the identical architecture (SFT stage only for controlled comparison). Both HuBERT-Large (+8.0 pp over SFT-only) and Whisper-Large-v3 (+5.3 pp) yield substantial gains, confirming that the delayed fusion architecture contributes independently of encoder choice. The stronger HuBERT gain (+8.0 pp vs. +5.3 pp for Whisper) is consistent with richer acoustic preservation benefiting the bypass.

6.3 Mechanism Analysis

6.3.1 Per-Domain and Subtask Analysis

Figure 4 tracks domain-level accuracy and inference latency from Base to SFT to Overtone. While standard SFT effectively captures coarser event-level structure like Sound, it plateaus on fine-grained domains. Overtone further improves Speech (+17.4 pp over SFT) and Music (+13.4 pp). This confirms that task fine-tuning alone cannot recover compressed acoustic details.

The per-sample re-listening trigger rate further supports selective invocation: Speech trig-

Table 4: Ablation studies on MMAU-mini. Each group isolates one design dimension. Groups (a), (c)–(e) use the SFT stage only for controlled comparison, while Group (b) evaluates the two-pass re-listening strategy. Δ is reported relative to 81.5%.

Configuration	Acc. (%)	Δ
<i>(a) Fusion strategy: single forward pass, varying injection depth and reconstruction loss</i>		
Base model (zero-shot)	56.5	−25.0
SFT-only (no bypass)	70.3	−11.2
Early fusion, no $\mathcal{L}_{\text{recon}}$	75.0	−6.5
Early fusion + $\mathcal{L}_{\text{recon}}$	72.4	−9.1
Delayed fusion, no $\mathcal{L}_{\text{recon}}$	76.8	−4.7
Delayed fusion + $\mathcal{L}_{\text{recon}}$	81.5	0.0
<i>(b) Re-listening strategy: two-pass pipeline, varying encoder and segmentation source</i>		
Raw re-listen via standard encoder	74.4	−7.1
Oracle segmentation WavLM bypass	82.3	+0.8
Learned WavLM bypass (Overtone)	81.5	0.0
<i>(c) Injection layer selection: delayed fusion, varying starting layer</i>		
Early fusion (L0 Start)	75.0	−6.5
L4 Start (4, 8, 12, 16)	73.8	−7.7
L8 Start (8, 12, 16, 20)	79.9	−1.6
L12 Start (12, 16, 20, 24)	81.5	0.0
L16 Start (16, 20, 24, 27)	76.7	−4.8
L20 Start (20, 22, 24, 26)	73.2	−8.3
<i>(d) Acoustic content verification: single-pass with controlled acoustic inputs</i>		
Normal bypass (real features)	78.7	−2.8
Shuffled features	71.1	−10.4
Mean-pooled bypass	75.5	−6.0
No acoustic input	68.9	−12.6
<i>(e) Frozen encoder selection: identical delayed fusion architecture, SFT stage only</i>		
Whisper-Large-v3 (ASR, supervised)	75.6	−5.9
HuBERT-Large (SSL, self-supervised)	78.3	−3.2
WavLM-Large	81.5	0.0

gers most frequently (73.4%), followed by Music (56.7%) and Sound (51.9%). This ordering mirrors the diagnostic findings and performance gains.

Subtask breakdowns (Appendix Table 10) show the largest gains in acoustically demanding categories (e.g., Phonemic Stress Pattern Analysis +28.3 pp, Rhythm and Tempo Understanding +23.9 pp). Conversely, knowledge-dependent tasks like Eco-Acoustic Knowledge show slight decreases, consistent with the bypass providing limited benefit when the bottleneck is world knowledge.

In terms of overhead, Overtone adds trainable cross-attention modules and an acoustic projection (WavLM is frozen), with +0.36 s average inference latency ($\sim 18\%$ increase over SFT-only). When the bypass is not triggered, Overtone skips

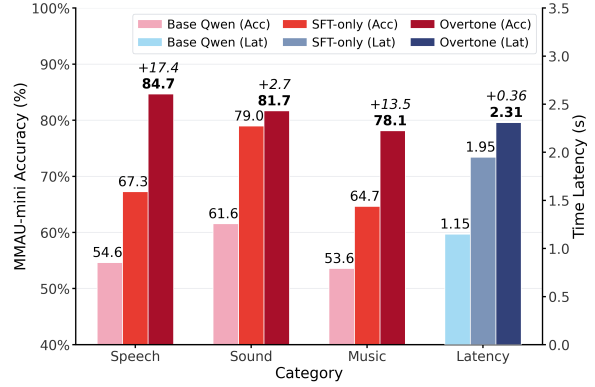


Figure 4: MMAU-mini accuracy by domain and time latency across Base, SFT, and Overtone. Largest gains appear in Speech (+17.4 pp over SFT), while introducing a marginal latency overhead (+0.36 s).

WavLM encoding and second-pass acoustic recovery entirely.

6.3.2 Learned Trigger Quality

Overtone with learned segment triggers achieves 81.5% on MMAU-mini, closely matching the oracle-segmentation setting (82.3%). Beyond accuracy, we directly evaluate segment localization quality using IoU (intersection over union between predicted and oracle time intervals): the learned triggers achieve a mean IoU of 0.93 with oracle segments, with over 96% of segments exceeding $\text{IoU} \geq 0.5$. The invalid-segment fallback rate remains below 7% across benchmarks, indicating reliable localization.

6.3.3 Post-training Probing

Figure 5 compares post-training probes for Base, SFT-only, and Overtone.

For Speaker ID, the base model approaches chance-level performance at L12. SFT-only partially restores this, but Overtone recovers substantially further (+24.7 pp over the base). Pitch Level follows a similar trend (+15.3 pp gain) at L12.

The consistent gap over SFT-only indicates recovery beyond standard fine-tuning.

6.3.4 Gate Activation Analysis

Figure 6 reports mean gate values across injection layers and subtasks, with stronger activation in deeper layers.

Interestingly, gate usage defies a strict acoustic-versus-semantic dichotomy. Tasks such as event counting and co-occurrence, despite appearing largely semantic, still exhibit strong gate activation, likely because they require precise temporal

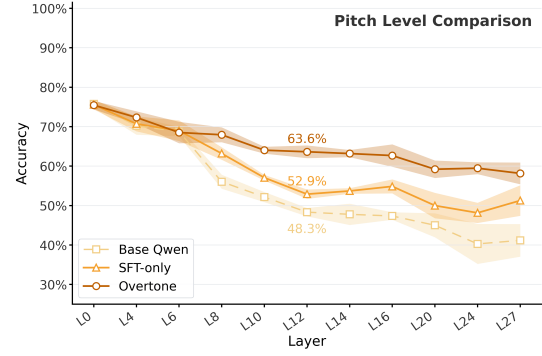
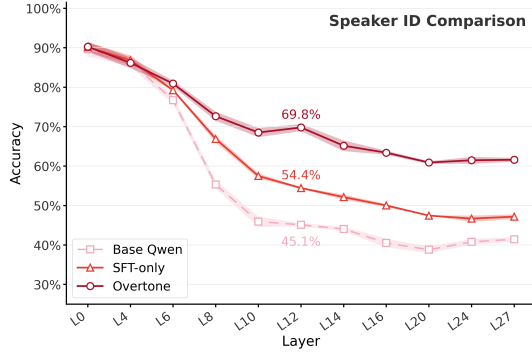


Figure 5: Post-training linear probe accuracy across layers. Overtone recovers acoustic information in the compression zone (L6–L12) relative to both the base model and SFT-only baselines.

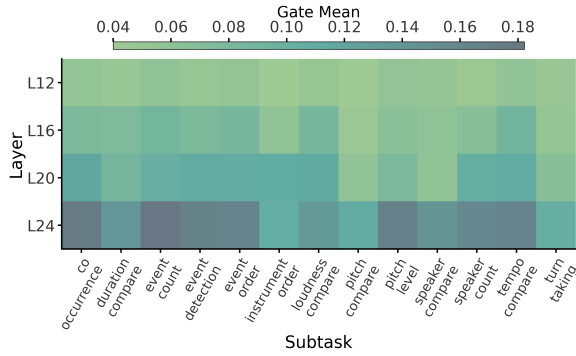


Figure 6: Learned gate activation values across injection layers and subtasks. Gates open progressively in deeper layers, confirming active utilization of the acoustic bypass.

localization. This suggests gates learn nuanced usage patterns beyond binary task categorization, presenting an avenue for future investigation.

6.3.5 Failure Analysis

We identify three failure modes where the acoustic bypass provides limited or negative benefit: (1) *Global temporal questions*. Case 4 (Appendix E) shows that segment-localized re-listening can amplify salient local detail at the expense of holistic context: the model correctly localizes a tempo-relevant segment but over-interprets local rhythmic variation as a global tempo trend. (2) *World-knowledge-dependent tasks*. Eco-Acoustic Knowledge (Table 10) shows a slight decrease (−2.1 pp from SFT), as this subtask relies on factual knowledge that the acoustic bypass cannot supply. (3) *Short clips*. On clips shorter than 3 s, the learned trigger typically defaults to the full clip, providing no localization benefit over a naive full-clip baseline.

These failure modes clarify the scope of the approach: the acoustic bypass is most effective when

the diagnostic evidence is local and fine-grained, and provides diminishing returns for tasks requiring global integration or external knowledge.

7 Conclusion

We identify a *representation reorganization* bottleneck in audio LLMs, where semantic abstraction coincides with reduced recoverability of fine-grained acoustic information. Guided by this diagnosis, Overtone uses delayed acoustic fusion to bypass the affected layers. Overtone achieves the highest MMAU score among open-source $\leq 8B$ models, while post-training probes confirm recovery of acoustic structure.

Limitations

Although Overtone improves fine-grained acoustic reasoning, gains are smaller on tasks dominated by global integration or world knowledge. In addition, learned gates do not exhibit a clean acoustic-versus-semantic separation, and their routing behavior remains difficult to interpret.

Ethical Considerations

Overtone improves fine-grained acoustic discrimination, particularly speaker identity recognition. While this capability is developed for benign audio understanding benchmarks, it could in principle facilitate speaker profiling or surveillance if misapplied. We note that Overtone operates on pre-existing open-source models and public datasets, and does not introduce new speaker identification capabilities beyond what dedicated speaker verification systems already provide. Nevertheless, practitioners deploying similar architectures should consider privacy safeguards and adhere to applicable data protection regulations.

References

- Andrea Agostinelli, Timo I. Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, Matt Sharifi, Neil Zeghidour, and Christian Frank. 2023. [Musiclm: Generating music from text](#). *Preprint*, arXiv:2301.11325.
- Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L Menick, Sebastian Borgeaud, and 8 others. 2022. [Flamingo: a visual language model for few-shot learning](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 23716–23736. Curran Associates, Inc.
- Tony Alex, Wish Suharitdamrong, Sara Atito, Armin Mustafa, Philip J. B. Jackson, Imran Razzak, and Muhammad Awais. 2026. [Pal: Probing audio encoders via llms – audio information transfer into llms](#). *Preprint*, arXiv:2506.10423.
- Abhra Chaudhuri, Anjan Dutta, Tu Bui, and Serban Georgescu. 2025. [A closer look at multimodal representation collapse](#). In *Forty-second International Conference on Machine Learning*.
- Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. 2022. [Wavlm: Large-scale self-supervised pre-training for full stack speech processing](#). *IEEE Journal of Selected Topics in Signal Processing*, 16(6):15051518.
- Joon Son Chung, Arsha Nagrani, and Andrew Senior. 2018. [VoxCeleb2: Deep Speaker Recognition](#). In *Interspeech 2018*, pages 1086–1090.
- Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Naveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3416 others. 2025. [Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities](#). *Preprint*, arXiv:2507.06261.
- Eduardo Fonseca, Xavier Favory, Jordi Pons, Frederic Font, and Xavier Serra. 2022. [Fsd50k: An open dataset of human-labeled sound events](#). *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:829–852.
- Sreyan Ghosh, Arushi Goel, Kaousheik Jayakumar, Lasha Koroshinadze, Nishit Anand, Zhifeng Kong, Siddharth Gururani, Sang gil Lee, Jaehyeon Kim, Aya Aljafari, Chao-Han Huck Yang, Sungwon Kim, Ramani Duraiswami, Dinesh Manocha, Mohammad Shoeybi, Bryan Catanzaro, Ming-Yu Liu, and Wei Ping. 2026. [Audio flamingo next: Next-generation open audio-language models for speech, sound, and music](#). *Preprint*, arXiv:2604.10905.
- Arushi Goel, Sreyan Ghosh, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang gil Lee, Chao-Han Huck Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, and Bryan Catanzaro. 2025. [Audio flamingo 3: Advancing audio intelligence with fully open large audio language models](#). *Preprint*, arXiv:2507.08128.
- Shawn Hershey, Daniel P W Ellis, Eduardo Fonseca, Aren Jansen, Caroline Liu, R Channing Moore, and Manoj Plakal. 2021. [The benefit of temporally-strong labels in audio event classification](#). In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 366–370.
- John Hewitt and Percy Liang. 2019. [Designing and interpreting probes with control tasks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2733–2743, Hong Kong, China. Association for Computational Linguistics.
- Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. [LoRA: Low-rank adaptation of large language models](#). In *International Conference on Learning Representations*.
- Shujie Hu, Long Zhou, Shujie Liu, Sanyuan Chen, Lingwei Meng, Hongkun Hao, Jing Pan, Xunying Liu, Jinyu Li, Sunit Sivasankaran, Linqun Liu, and Furu Wei. 2024. [WavLLM: Towards robust and adaptive speech large language model](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4552–4572, Miami, Florida, USA. Association for Computational Linguistics.
- Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mdry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, and 399 others. 2024. [Gpt-4o system card](#). *Preprint*, arXiv:2410.21276.
- Jaeyeon Kim, Heeseung Yun, Sang Hoon Woo, Chao-Han Huck Yang, and Gunhee Kim. 2025. [Wow-bench: Evaluating fine-grained acoustic perception in audio-language models via marine mammal vocalizations](#). In *TTIC Summer Workshop on Foundations of Speech and Audio Foundation Models 2025*.
- Sonal Kumar, Šimon Sedláček, Vaibhavi Lokegaonkar, Fernando López, Wenyi Yu, Nishit Anand, Hyeon-gon Ryu, Lichang Chen, Maxim Plička, Miroslav

- Hlaváček, William Fineas Ellingwood, Sathvik Udupa, Siyuan Hou, Allison Ferner, Sara Barahona, Cecilia Bolaños, Satish Rahi, Laura Herrera-Alarcón, Satvik Dixit, and 15 others. 2026. [Mmaupro: a challenging and comprehensive benchmark for holistic evaluation of audio general intelligence](#). In *Proceedings of the Fortieth AAAI Conference on Artificial Intelligence and Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence and Sixteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'26/IAAI'26/EAAI'26. AAAI Press.
- Jianglin Lu, Hailing Wang, Yi Xu, Yizhou Wang, Kuo Yang, and Yun Fu. 2025. [Representation potentials of foundation models for multimodal alignment: A survey](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pages 16669–16684. Association for Computational Linguistics.
- Ziyang Ma, Yinghao Ma, Yanqiao Zhu, Chen Yang, Yi-Wen Chao, Ruiyang Xu, Wenxi Chen, Yuanzhe Chen, Zhuo Chen, Jian Cong, Kai Li, Kelian Li, Siyou Li, Xinfeng Li, Xiquan Li, Zheng Lian, Yuzhe Liang, Minghao Liu, Zhikang Niu, and 15 others. 2025. [Mmar: A challenging benchmark for deep reasoning in speech, audio, music, and their mix](#). *Preprint*, arXiv:2505.13032.
- Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. [Librispeech: An asr corpus based on public domain audio books](#). In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210.
- Karol J. Piczak. 2015. [Esc: Dataset for environmental sound classification](#). *Proceedings of the 23rd ACM international conference on Multimedia*.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. [Robust speech recognition via large-scale weak supervision](#). *Preprint*, arXiv:2212.04356.
- Andrew Rouditchenko, Saurabhchand Bhati, Edson Araujo, Samuel Thomas, Hildegard Kuehne, Rogério Feris, and James R. Glass. 2025. [Omni-r1: Do you really need audio to fine-tune your audio llm?](#) *2025 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–7.
- S Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. 2024. [Mmau: A massive multi-task audio understanding and reasoning benchmark](#). *Preprint*, arXiv:2410.19168.
- Björn Schuller, Anton Batliner, Stefan Steidl, and Dino Seppi. 2011. [Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge](#). *Speech Commun.*, 53:1062–1087.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. [Deepseekmath: Pushing the limits of mathematical reasoning in open language models](#). *Preprint*, arXiv:2402.03300.
- Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun MA, and Chao Zhang. 2024. [Salmonn: Towards generic hearing abilities for large language models](#). In *International Conference on Learning Representations*, volume 2024, pages 16607–16629.
- Fei Tian, Xiangyu Tony Zhang, Yuxin Zhang, Haoyang Zhang, Yuxin Li, Daijiao Liu, Yayue Deng, Donghang Wu, Jun Chen, Liang Zhao, Chengyuan Yao, Hexin Liu, Eng Siong Chng, Xuerui Yang, Xiangyu Zhang, Daxin Jiang, and Gang Yu. 2025. [Step-audio-r1 technical report](#). *Preprint*, arXiv:2511.15848.
- Dingdong WANG, Junan Li, Jincenzi Wu, Dongchao Yang, Xueyuan Chen, Tianhua Zhang, and Helen Meng. 2026. [Mmsu: A massive multi-task spoken language understanding and reasoning benchmark](#). In *International Conference on Learning Representations*, volume 2026, pages 31374–31410.
- Boyong Wu, Chao Yan, Chen Hu, Cheng Yi, Chengli Feng, Fei Tian, Feiyu Shen, Gang Yu, Haoyang Zhang, Jingbei Li, Mingrui Chen, Peng Liu, Wang You, Xiangyu Tony Zhang, Xingyuan Li, Xuerui Yang, Yayue Deng, Yechang Huang, Yuxin Li, and 90 others. 2025. [Step-audio 2 technical report](#). *Preprint*, arXiv:2507.16632.
- Daiqing Wu, Xuan Zhang, Dongbao Yang, Jiashu Yao, Longfei Chen, Qingsong Liu, Sicheng Zhao, Can Ma, Yangyang Kang, and Yu ZHOU. 2026a. [Echo: Towards advanced audio comprehension via audio-interleaved reasoning](#). In *The Fourteenth International Conference on Learning Representations*.
- Shu Wu, Chenxing Li, Wenfu Wang, Hao Zhang, Hualei Wang, Meng Yu, and Dong Yu. 2026b. [Audio-thinker: Guiding large audio language model when and how to think via reinforcement learning](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(40):3396233970.
- 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. 2025a. [Qwen2.5-omni technical report](#). *Preprint*, arXiv:2503.20215.
- Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfu Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, and 19 others. 2025b. [Qwen3-omni technical report](#). *Preprint*, arXiv:2509.17765.

- Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2023. [SpeechGPT: Empowering large language models with intrinsic cross-modal conversational abilities](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15757–15773, Singapore. Association for Computational Linguistics.
- Dong Zhang, Gang Wang, Jinlong Xue, Kai Fang, Liang Zhao, Rui Ma, Shuhuai Ren, Shuo Liu, Tao Guo, Weiji Zhuang, Xin Zhang, Xingchen Song, Yihan Yan, Yongzhe He, Cici, Bowen Shen, Chengxuan Zhu, Chong Ma, Chun Chen, and 80 others. 2025. [Mimo-audio: Audio language models are few-shot learners](#). *Preprint*, arXiv:2512.23808.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph Gonzalez, and Ion Stoica. 2023. [Judging llm-as-a-judge with mt-bench and chatbot arena](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc.

A Probing Study Details

A.1 Probe Training Configuration

All linear probes are single nn.Linear classifiers trained on frozen hidden states extracted from the target model. The probe training configuration is summarized in Table 5.

Table 5: Linear probe training configuration.

Item	Value
Probe architecture	single linear layer
Optimizer	Adam
Learning rate	1×10^{-3}
Batch size	64
Max epochs	50
Early stopping patience	5
Loss	cross-entropy
Seeds	5

Hidden states are obtained as the layer output averaged over audio token positions. For the audio encoder reference (ENC), we use the Qwen audio encoder output prior to projection into the LLM hidden space. For WavLM-Large, we mean-pool the final layer output in the same way. We probe 14 layers across the LLM (L0, L2, L4, L6, L8, L9, L10, L11, L12, L14, L16, L20, L24, L27), with denser sampling in the L6–L12 transition zone.

Sample sizes per probe task (train / val / test splits) are listed in Table 6. Pitch Level and Keyword Content use speaker-disjoint splits to prevent identity leakage; Speaker ID splits utterances within speakers.

Table 6: Probe dataset sizes.

Task	Train	Val	Test	Total
Speaker ID	3514	753	753	5020
Pitch Level	1419	291	288	1998
Keyword Content	2039	464	461	2964
Pitch Direction	712	145	143	1000
Overlap Detection	314	68	66	448
Hard Overlap	1040	226	234	1500

A.2 Supplementary Probes

Beyond the three core probes reported in the main paper, we conduct additional probes to verify that the compression pattern is (i) consistent across pitch encoding formats and (ii) selective with respect to feature granularity.

Pitch Direction. A 2-class probe classifying whether the F0 trajectory is rising or falling. Like the absolute Pitch Level probe, accuracy declines

Table 7: Supplementary probe accuracies (% , mean over 5 seeds) at representative layers. Pitch Direction is fine-grained and compresses with depth; Overlap variants are coarser and remain stable.

Layer	Pitch Dir.	Overlap	Hard Overlap
ENC	65.0	95.5	97.0
L0	65.2	94.8	97.0
L4	64.3	95.5	96.6
L8	55.0	95.5	94.3
L12	49.2	95.5	93.0
L16	49.9	96.1	92.0
L20	51.1	94.2	90.7
L24	47.3	97.9	87.4
L27	47.0	93.9	86.1

through the compression zone—from 65% at the encoder to roughly 50% at L12 and below—confirming that both absolute pitch and pitch dynamics are compressed together.

Overlap Detection. A 2-class probe detecting whether two voices are simultaneously present. In contrast to fine-grained acoustic attributes, this coarser pattern remains essentially intact across all 14 layers (94–98% accuracy, no decline through L6–L12). A more challenging *Hard Overlap* variant with tighter overlap conditions shows a more gradual decline through deeper layers (97% at ENC → 86% at L27) but does not collapse. Probe accuracies are summarized in Table 7.

These results reinforce the selectivity of the compression: information types that contribute to next-token prediction (coarse overlap structure) are preserved, while fine-grained acoustic attributes (pitch direction, like absolute pitch and speaker identity) are systematically compressed in the same L6–L12 zone identified in the main text.

A.3 Cross-Model Generalization

To verify that representation reorganization is a general phenomenon rather than a Qwen2.5-Omni-specific artifact, we replicate the probing study on three additional audio LLMs spanning different scales and architectures: Audio-Flamingo-3 (8B, 28 layers), Qwen3-Omni (30B, 48 layers), and Step-Audio-R1 (33B, 64 layers).

Two consistent phenomena emerge across all three additional models (Figure 7). First, semantic discriminability (Keyword Content) is actively constructed in early LLM layers (L0–L8) across all architectures, rising from near-chance encoder-level accuracy. Second, and most critically for our hypothesis, fine-grained acoustic information

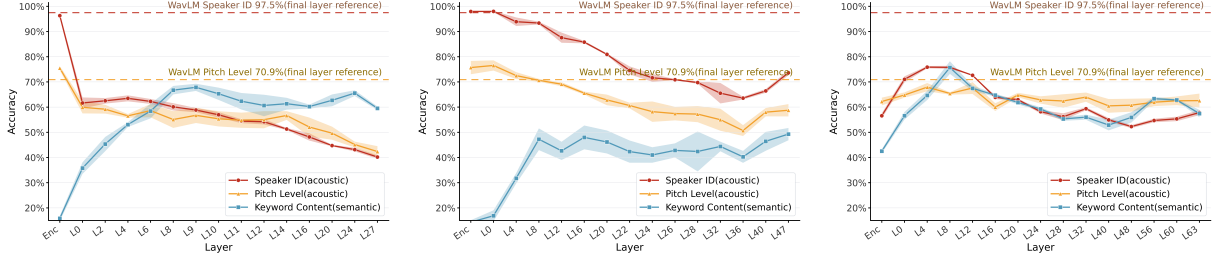


Figure 7: Cross-model probing results. **Left:** Audio-Flamingo-3 (8B). **Center:** Qwen3-Omni (30B). **Right:** Step-Audio-R1 (33B). Across all three models, fine-grained acoustic features in deep LLM layers fall substantially below the WavLM reference, while semantic features are actively constructed in early layers. The specific shape of acoustic decline varies by architecture.

in deep LLM layers is substantially lower than the WavLM reference (97% on Speaker ID), regardless of model scale—deep-layer Speaker ID ranges from 40% (Audio-Flamingo-3) to 74% (Qwen3-Omni), all well below WavLM. This holds even for Step-Audio-R1, which operates at over four times Overtone’s parameter count.

The precise location where acoustic information is lost varies by architecture: Audio-Flamingo-3 shows the steepest drop at the encoder-to-LLM transition, Qwen3-Omni exhibits gradual decline distributed across mid-to-deep layers, and Step-Audio-R1 displays a non-monotonic trajectory with early-layer enhancement followed by deep-layer compression. We interpret this architectural diversity as evidence *strengthening* rather than complicating our central claim: four audio LLMs with different encoders, different scales (7B–33B), and different training pipelines all arrive at the same end-state—deep representations that have substantially lower fine-grained acoustic discriminability than dedicated acoustic encoders. The phenomenon is therefore not an artifact of any specific model design choice, but a general property of audio language modeling.

This diversity also motivates Overtone’s design philosophy. Because the location of compression is model-dependent, a robust architectural intervention must be *learnable and adaptive* rather than tied to any specific layer range. Our gated cross-attention with diagnostic-driven injection layers (L12–L24 for Qwen2.5-Omni) implements exactly this principle: the gates learn where and how strongly to retrieve acoustic information, allowing the same architectural template to be re-targeted to other audio LLMs by re-running the diagnostic probe to identify their specific compression profile.

B Method Details

B.1 Restart Controller

The Restart Controller is a lightweight post-hoc binary classifier that decides whether to execute the second-pass re-listening after the first pass emits a `<seg>` tag. It is trained offline and independently from Overtone’s main weights; its gradient does not flow back to the model.

Input features. The controller consumes observable signals from the first-pass output, including: first-pass text length, presence and count of `<answer>` and `<seg>` tags, parsed answer validity, segment start/end/duration/center, out-of-bounds margin, segment-to-audio duration ratio, answer log-probability, and answer margin between top-1 and top-2 options.

Architecture. A 3-layer MLP: Linear $\rightarrow$ ReLU $\rightarrow$ Linear $\rightarrow$ ReLU $\rightarrow$ Linear(1), with hidden dimension 64. Numeric features are concatenated with a learned audio domain embedding (e.g., speech/sound/music; embedding dim 8).

Label construction. For each first-pass output, we additionally run the second pass and assign $y = 1$ if the first pass was incorrect and the second pass corrects it, and $y = 0$ otherwise (i.e., when restarting would not help or would harm). Labels are constructed on a separate validation pool drawn from the same data sources as SFT training but disjoint from both the SFT/GRPO training data and all evaluation benchmarks (MMAU, MMAR, MMAU-Pro, MMSU).

Training. AdamW optimizer with learning rate 1×10^{-3} , weight decay 1×10^{-4} , 200 epochs, and weighted binary cross-entropy loss with positive class weight set to the negative-to-positive ratio in the training data.

Threshold selection. We sweep the decision threshold over $[0.05, 0.95]$ in steps of 0.01 on a validation set, selecting the threshold that maximizes controller-gated final accuracy. Ties are broken in favor of fewer harmful restarts.

B.2 Gated Cross-Attention

At each injection layer $l \in \{12, 16, 20, 24\}$, a gated cross-attention module is inserted after the standard self-attention. The cross-attention uses grouped-query attention (GQA) with 28 query heads, 4 key/value heads, and head dimension 128, matching the Qwen2.5-Omni decoder’s GQA configuration. Queries are projected from the current decoder hidden state $H'_l \in \mathbb{R}^{T \times d}$ (with $d = 3584$); keys and values are projected from the acoustic memory X_{aco} , with the 4 KV heads repeat-interleaved to 28 to align with the query heads before attention computation.

Acoustic projection. The acoustic memory X_{aco} is produced by a trainable projection of frozen WavLM features: a linear layer mapping $1024 \rightarrow 3584$ followed by LayerNorm. WavLM remains frozen across all training stages.

Gate initialization. Gates are initialized with $W_l = \mathbf{0}$ and $b_l = -5$ for all injection layers, yielding initial gate value $g_{l,t} = \sigma(-5) \approx 0.007$ uniformly across token positions at training start. The zero-initialized projection means the input-dependent routing is inactive at initialization—the gate behaves as a constant near zero—and gradient updates progressively unlock token-specific gating throughout training.

Empty acoustic memory. During the first pass, or when the Restart Controller suppresses the second pass, the acoustic memory buffer is empty. In this case, the cross-attention pathway is bypassed entirely and the layer reduces to its standard self-attention output, incurring no additional cost.

B.3 Reconstruction Loss

The reconstruction loss L_{recon} is computed on the final-layer hidden states—specifically, the output of the LLM’s final RMSNorm—at positions corresponding to audio tokens in the input sequence.

Prediction head f_θ . A 2-layer MLP with architecture $3584 \rightarrow 2048 \rightarrow 1024$ (GELU activation after the first layer). The 1024-dimensional output matches WavLM’s feature dimension. f_θ

Table 8: Data pipeline funnel. The first two rows (audio sourcing) are counted at the clip level; subsequent rows count generated QA pairs. “Overall” denotes retention from raw QA generation to the final SFT training set.

Stage	Count	Pass rate / ratio
Audio clips collected	98,513	—
Captions generated	98,400	99.9%
Raw QA generated	258,793	2.63 / clip
Structural postprocessing	254,086	98.2%
Judge-routed to SFT	154,689	60.9%
Silent-audio strong subset	75,851	49.0%
Final SFT training set	75,851	29.3% (overall)

is trained jointly with the cross-attention modules during Stage 1 and discarded at inference.

Loss schedule. $\lambda = 0.1$ during Stage 1 (SFT), $\lambda = 0$ during Stage 2 (GRPO), and L_{recon} is not used during the Restart Controller training stage. We disable L_{recon} during GRPO to allow the policy to specialize on the task reward signal without continued representation-level constraints; by the end of SFT, the regularization has served its purpose of anchoring deep-layer representations to WavLM-derived targets.

No-recon ablation. The “w/o recon” variant referenced in the main paper’s ablations uses identical Stage 1 hyperparameters but sets $\lambda = 0$ throughout. All other components, including the cross-attention modules and their gate initialization, are unchanged.

C Data and Training Details

C.1 Data Pipeline Statistics

Table 8 summarizes the multi-stage filtering pipeline used to construct the SFT training set.

Audio-centricity filtering. After judge routing, the 154,689 SFT-routed samples are re-evaluated on a silent-audio variant of each input. Samples answered correctly without audio (78,838, 51.0%) are classified as *weak* and discarded; the remaining 75,851 (49.0%) constitute the audio-grounded *strong* subset used for SFT training. The high weak-sample proportion indicates that judge validation alone is insufficient as a quality filter—over half of judge-validated samples could be answered from textual priors—making this step essential to ensure the training signal genuinely depends on audio.

Table 9: GRPO reward weights.

Component	Weight
w_{acc}	+0.50
w_{fmt}	+0.20
w_{seg}	+0.15
w_{oob}	−0.30
no-seg penalty	−0.10
multi-answer penalty	−0.05
trailing/prefix-text penalty	−0.05
structure penalty	−0.02
KL coefficient	0.04

GRPO training data. The GRPO stage uses a separate pool of 20,207 examples disjoint from the SFT training set. This pool comprises 10,762 samples routed by the judge for reward-based refinement (where the reader models agreed on the answer but the CoT trace warranted improvement) and 9,445 additional samples selected from a harder pool where the two reader models initially produced different answers despite a well-formed question. The composition balances cleaner reward signal with harder cases where the outcome reward can supervise learning without requiring gold-standard CoT, consistent with our motivation for the GRPO stage in Section 4.4 of the main paper.

C.2 GRPO Reward Function Details

The composite reward used in our experiments is:

$$R = R_{\text{acc}} + R_{\text{format}} + R_{\text{seg}} + R_{\text{penalty}}.$$

Component definitions.

- R_{acc} : $+w_{\text{acc}}$ if the parsed answer matches the ground-truth option, 0 otherwise.
- R_{format} : $+w_{\text{fmt}}$ if the output contains a well-formed `<seg>s, e</seg>` tag, 0 otherwise.
- R_{seg} : $w_{\text{seg}} \cdot \text{IoU}(\text{pred}, \text{oracle})$ when an oracle segment is available; subject to an OOB penalty w_{oob} if the predicted segment exceeds the audio duration (tolerance 0.5 s).
- R_{penalty} : small penalties for malformed structure—multiple answers, trailing/prefix text outside the structured template.

Default weights. Table 9 lists the reward weights used in our experiments.

OOB conditions trigger when $s < 0$, $e > T + 0.5$, $s \geq e$, or $e - s < 0.05$ s (where T is the audio duration). Within each group of G rollouts per prompt, the GRPO advantage is the standardized reward $A_i = (R_i - \mu_R) / \sigma_R$, and the policy is updated to increase the probability of high-advantage

rollouts subject to a KL constraint against the SFT-initialized reference policy.

D Additional Experimental Results

D.1 MMAU-mini Full Subtask Breakdown

Table 10 provides the complete per-subtask accuracy of Base Qwen, SFT-only, and Overtone on MMAU-mini (1,000 samples, 27 subtasks across three domains).

D.2 MMAU Full Set Per-Subtask Results

Table 11 reports Overtone’s per-subtask accuracy on the full MMAU benchmark (9,000 samples), spanning the same 27 subtasks at substantially larger per-category sample sizes than the mini split.

The pattern observed on MMAU-mini is preserved at the larger sample sizes: Speech remains Overtone’s strongest domain (81.0%), driven by tasks requiring fine-grained acoustic discrimination (e.g., Multi Speaker Role Mapping 99.1%, Dissonant Emotion Interpretation 95.1%). The hardest subtasks—Emotion Flip Detection (36.9%, Speech) and Temporal Reasoning (44.4%, Music)—both involve global structural reasoning that goes beyond local segment-level acoustic verification, consistent with the limitations discussed in the main text.

E Case Studies

We present four representative examples illustrating Overtone’s behavior on different task types. Cases 1–3 are success cases where second-pass re-listening corrects a first-pass error; Case 4 is a failure case where re-listening introduces an error despite accurate segment localization. “Correct answer” denotes the benchmark’s reference label for each multiple-choice question.

Case 1: Speaker Comparison (success).

Audio. Two speech segments: speaker A at 0–4.0 s, speaker B at 4.3–8.3 s.

Question. Listen to both speech segments. Are they from the same speaker or different speakers?

Options. (A) Yes, same speaker; (B) No, different speakers; (C) Cannot determine; (D) Only one speaker is present.

Correct answer: A.

First-pass output: B.

Second-pass output: A.

Table 10: MMAU-mini full subtask breakdown. Accuracy (%) for Base Qwen, SFT-only, and Overtone. N denotes samples per subtask.

Domain	Subtask	N	Base	SFT	Overtone
Speech	Multi Speaker Role Mapping	27	81.5	85.2	96.3
	Event-Based Knowledge Retrieval	33	48.5	90.9	87.9
	Key Highlight Extraction	21	90.5	90.5	95.2
	Conversational Fact Retrieval	22	63.6	95.5	95.5
	Phonological Sequence Decoding	49	61.2	75.5	87.8
	Dissonant Emotion Interpretation	35	54.3	74.3	88.6
	Phonemic Stress Pattern Analysis	53	37.7	56.6	84.9
	Counting	29	44.8	55.2	82.8
	Emotion State Summarisation	44	43.2	36.4	63.6
	Emotion Flip Detection	20	50.0	30.0	75.0
<i>Speech subtotal</i>		333	54.7	67.3	84.7
Sound	Sound-Based Event Recognition	46	67.4	87.0	89.1
	Event-Based Sound Reasoning	48	68.8	81.3	87.5
	Eco-Acoustic Knowledge	47	48.9	76.6	74.5
	Acoustic Source Inference	48	68.8	87.5	93.8
	Acoustic Scene Reasoning	48	50.0	77.1	77.1
	Ambient Sound Interpretation	48	64.6	79.2	77.1
	Temporal Event Reasoning	48	62.5	64.6	72.9
<i>Sound subtotal</i>		333	61.6	79.0	81.7
Music	Socio-cultural Interpretation	20	70.0	75.0	80.0
	Emotional Tone Interpretation	33	63.6	87.9	93.9
	Lyrical Reasoning	10	60.0	80.0	90.0
	Instrumentation	35	60.0	71.4	80.0
	Melodic Structure Interpretation	33	57.6	66.7	78.8
	Rhythm and Tempo Understanding	46	32.6	56.5	80.4
	Musical Genre Reasoning	34	67.6	76.5	85.3
	Musical Texture Interpretation	34	52.9	67.6	88.2
	Harmony and Chord Progressions	33	63.6	66.7	81.8
	Temporal Reasoning	56	37.5	35.7	50.0
<i>Music subtotal</i>		334	53.6	64.7	78.1
Overall		1000	56.5	70.3	81.5

First-pass CoT. “I need to compare the vocal characteristics. . . <seg>0.00, 8.00</seg> . . . have distinctly different vocal qualities.”

Second-pass CoT. “. . . share similar vocal timbre and speaking style.”

Localization. Predicted [0.0, 8.0], oracle [0.0, 8.0], IoU 1.00.

Analysis. A direct illustration of Overtone’s core mechanism. The first pass misjudges speaker identity using the LLM’s deep-layer representation, where speaker information has been compressed. Re-listening with WavLM features supplies the timbre evidence needed to correct the comparison.

Case 2: Pitch Comparison (success).

Audio. Two consecutive speech segments spanning 0–7.3 s; the task is to compare the pitch of the second segment relative to the first.

Question. How does the pitch of the second speech segment compare to the first?

Options. (A) Same pitch; (B) Higher pitch; (C) Cannot determine; (D) Lower pitch.

Correct answer: D.

First-pass output: B.

Second-pass output: D.

First-pass CoT. “. . . <seg>0.00, 7.30</seg> . . . the second segment has a higher fundamental frequency.”

Second-pass CoT. “. . . the second segment has a lower fundamental frequency.”

Localization. Predicted [0.0, 7.3], oracle [0.0, 7.3], IoU 1.00.

Analysis. A direct comparison task on pitch—one of the most heavily compressed acoustic attributes in our probing study. The first pass inverts the relative F0 ordering. With WavLM features routed through the bypass during the second pass, the model accesses pitch contours preserved by the self-supervised speech encoder and correctly identifies the descending direction.

Table 11: Overtone’s per-subtask accuracy on the full MMAU benchmark. N denotes total samples; E/M/H denote difficulty stratification (Easy / Medium / Hard).

Domain	Subtask	N	E / M / H	Correct	Acc. (%)
Speech	Multi Speaker Role Mapping	233	0 / 233 / 0	231	99.1
	Event-Based Knowledge Retrieval	283	0 / 0 / 283	267	94.3
	Key Highlight Extraction	189	0 / 189 / 0	183	96.8
	Conversational Fact Retrieval	199	0 / 199 / 0	183	92.0
	Phonological Sequence Decoding	465	465 / 0 / 0	426	91.6
	Dissonant Emotion Interpretation	304	0 / 304 / 0	289	95.1
	Phonemic Stress Pattern Analysis	580	364 / 15 / 201	373	64.3
	Counting	249	0 / 249 / 0	180	72.3
	Emotion State Summarisation	368	0 / 0 / 368	249	67.7
	Emotion Flip Detection	130	0 / 0 / 130	48	36.9
<i>Speech subtotal</i>		3000	829 / 1189 / 982	2429	81.0
Sound	Sound-Based Event Recognition	453	0 / 453 / 0	428	94.5
	Event-Based Sound Reasoning	436	0 / 436 / 0	346	79.4
	Eco-Acoustic Knowledge	442	0 / 442 / 0	349	79.0
	Acoustic Source Inference	426	0 / 426 / 0	333	78.2
	Acoustic Scene Reasoning	453	0 / 453 / 0	345	76.2
	Ambient Sound Interpretation	390	0 / 0 / 390	252	64.6
	Temporal Event Reasoning	400	133 / 0 / 267	283	70.8
<i>Sound subtotal</i>		3000	133 / 2210 / 657	2336	77.9
Music	Socio-cultural Interpretation	32	14 / 17 / 1	29	90.6
	Emotional Tone Interpretation	269	97 / 131 / 41	215	79.9
	Lyrical Reasoning	90	33 / 37 / 20	73	81.1
	Instrumentation	313	221 / 88 / 4	246	78.6
	Melodic Structure Interpretation	375	157 / 157 / 61	280	74.7
	Rhythm and Tempo Understanding	392	138 / 242 / 12	323	82.4
	Musical Genre Reasoning	316	36 / 250 / 30	235	74.4
	Musical Texture Interpretation	350	138 / 194 / 18	317	90.6
	Harmony and Chord Progressions	367	120 / 235 / 12	286	77.9
	Temporal Reasoning	496	97 / 283 / 116	220	44.4
<i>Music subtotal</i>		3000	1051 / 1634 / 315	2224	74.1
Grand Total		9000	2013 / 5033 / 1954	6989	77.7

Case 3: Environmental Temporal Gap (success).

Audio. Bird/farm ambience; chicken clucking from 0 s to ~ 4.5 s; relative quiet with background birds; rooster crowing begins at ~ 8.35 s.

Question. Which statement accurately describes the sound events between 4.0 and 8.5 seconds?

Options. (A) Clucking transitions directly into crowing; (B) Relative quiet with only background bird chirping; (C) Rooster crows continuously; (D) Completely silent.

Correct answer: B.

First-pass output: A.

Second-pass output: B.

First-pass CoT. "... <seg>4.00, 8.50</seg> ... clucking ends around 4.5 s, followed by a brief transition, then crowing. ..."

Second-pass CoT. "... a distinct period of relative quiet occurs between approximately 5.5 and 7.5 seconds. ..."

Localization. Predicted [4.0, 8.5], oracle [3.8, 8.7], IoU 0.92.

Analysis. The bypass is not limited to speech tasks. Overtone identifies the relevant temporal window and the second pass recovers fine temporal structure within the environmental scene—a quiet gap between two distinct events that the first pass over-compressed into a direct transition.

Case 4: Music Tempo (failure).

Audio. Music throughout 0–10 s; large room/hall ambience.

Question. How does the tempo of the music change from the beginning to the end of the clip?

Options. (A) Remains constant; (B) Gradually speeds up; (C) Fluctuates unpredictably; (D) Gradually slows down.

Correct answer: A.

First-pass output: A.

Second-pass output: B.

First-pass CoT. "... <seg>0.00,3.00</seg> ... rhythm and pace remain consistent throughout..."

Second-pass CoT. "... by the 3-second mark, the pace... increased... final moments being the fastest..."

Localization. Predicted [0.0, 3.0], oracle [0.0, 3.0], IoU 1.00.

Analysis. An informative failure mode. The segment is localized perfectly, but the question requires a *global* comparison between the clip's beginning and end. Concentrated re-listening on the opening 3 seconds causes the model to over-interpret local rhythmic variation as a global tempo trend. This case clarifies a boundary of the approach: segment-localized re-listening is most beneficial when the diagnostic evidence is local, and can be counterproductive for questions requiring global temporal integration.

Note on segment localization. Case 3 (IoU 0.92 on a 10 s clip with a relevant window of approximately 5 s) demonstrates that the learned trigger localizes precisely on longer audio with multiple distinct events—substantially better than would be achievable by trivially selecting the full clip. On shorter clips with a single salient region, oracle annotations and learned predictions both naturally span the full duration.