SEPARATING WHAT FROM WHEN: STAGE-SPECIALISED CREDIT ASSIGNMENT FOR AUDITABLE FOREGROUND SOUND EVIDENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Audio–language models can recognise a foreground sound while failing to show when it occurs. A tool-assisted system can also change a correct class label even when its timing tool supplies no new class evidence. We study whether these two outputs can be attributed to separate stages. Our protocol has three auditable steps: the model proposes a class, a frozen label-free tool returns an interval, and a finalisation rule keeps the proposal unless the new observation contains class evidence. We adapt only the proposal stage, while holding the temporal tool and answer contract fixed.

On source-disjoint evaluation splits, stage specialisation raises answer exact match from 0.639 to 0.736 on the primary split without changing any predicted interval. The same direction recurs on a later split and on new recordings of adapted classes, but reverses for an unseen class. With the semantic stage fixed, a frozen spectral-contrast localiser raises temporal IoU from 0.293 to 0.464 over an energy-based tool. External audio–language models and the unadapted backbone use the same tool and answer contract, so their comparison measures semantic recognition under this protocol rather than native end-to-end performance. Independent retraining preserves the semantic effect, while learned routing does not improve localisation. The evidence supports stage-specific attribution within the evaluated protocol and locates its limits in low-SNR localisation and unseen-class transfer; it does not establish general state-of-the-art performance.

1 INTRODUCTION

Ask an audio–language model (LALM) which short event interrupts a continuous background recording and it may identify the event. Ask it *when* the event happened and the predicted interval is often wrong. The two failures are entangled: a pipeline that calls an acoustic tool and then lets the model write its final answer will sometimes rewrite a label that was already right, using nothing but the timing the tool returned. A single aggregate score cannot say which stage helped and which stage hurt.

The distinction is important because the two fields depend on different evidence. Class recognition asks which sound best explains the foreground event, often against a much longer background source. Temporal localisation asks which short interval supports that decision. The same waveform can therefore produce a correct class with a poor interval, or a useful interval that does not repair an incorrect class. When both fields are generated by one unconstrained answer step, a change in either field can be attributed to the wrong component. It becomes unclear whether a tool improved perception, changed the language model's wording, or introduced a new error after an initially correct proposal.

This problem is not solved by recording more intermediate text. A trace can show that a model called a tool, but it does not by itself show that the tool caused the final class or interval. The same issue appears in comparisons between tool-augmented systems: if each model uses a different acoustic front-end, a better final score mixes semantic recognition with the quality of the front-end. We instead ask a narrower causal question. If one stage is changed while the remaining stages, tool

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.

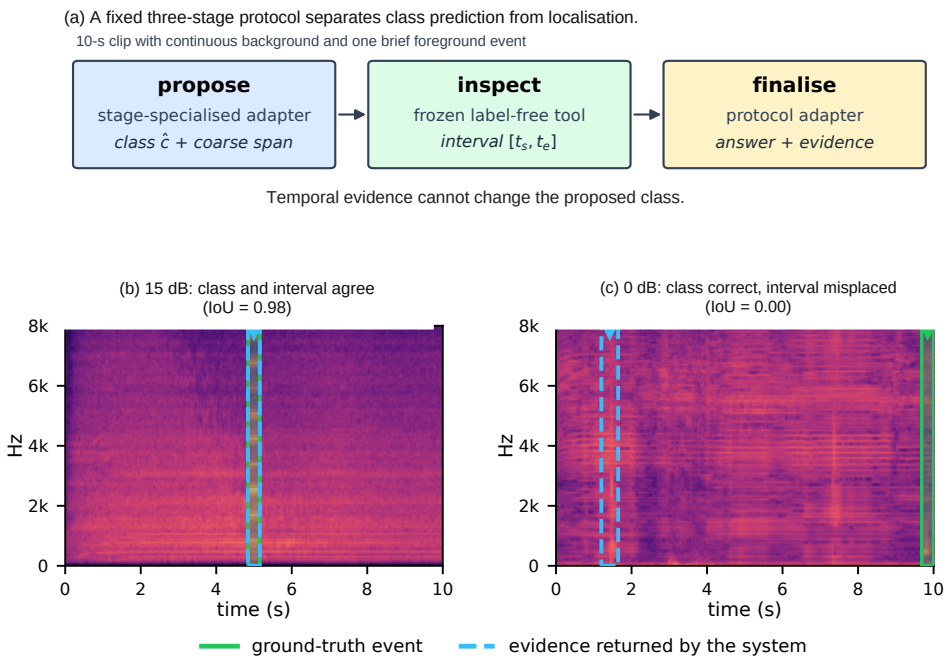

Figure 1: (a) Fixed three-stage protocol for separating class prediction from temporal localisation. The class proposal passes through inspection by a frozen, label-free tool before finalisation. (b) Mel spectrogram of a 15 dB evaluated clip with a correct class and interval. (c) Mel spectrogram of a 0 dB evaluated clip with a correct class but misplaced interval. In (b) and (c), the horizontal axis shows time in seconds and the vertical axis shows frequency in hertz; colour intensity encodes spectrogram energy. The solid green interval marks the annotated event, and the dashed cyan interval marks the system's temporal evidence.

interface, and answer contract stay fixed, which field changes? The resulting claim is smaller than a general agent claim, but it is directly testable.

This distinction also changes how training feedback should be assigned. A correct class paired with a wrong interval should not produce the same update as a wrong class paired with a useful interval. In the first case, the semantic proposal already contains the needed information and localisation is the likely bottleneck. In the second, improving the localiser cannot repair the semantic decision unless the protocol explicitly allows temporal evidence to change the class. Mixing both cases into one reward or loss hides the source of the error. Our protocol therefore treats an unchanged output field as a control observation rather than as missing information.

We treat *what* a sound is and *when* it occurs as separate fields and intervene on one stage at a time. The test bed uses ten-second soundscapes with continuous backgrounds and one short foreground event. Each system follows the same three steps: PROPOSE, INSPECT, and FINALISE. A stage-specialised adapter predicts the class, a label-free tool returns an interval, and a finalisation rule retains the proposed class unless new class evidence is available. This controlled loop tests stage-specific attribution; it is not a general autonomous agent.

The protocol supports two complementary interventions. In the semantic intervention, only the adapter active during PROPOSE is replaced. The inspection tool, protocol adapter, decoding contract, and finalisation rule remain unchanged. If the class changes while the interval does not, the observed gain is localised to the proposal stage. In the temporal intervention, the model and protocol are held fixed while the localiser is replaced. Changes in temporal IoU or boundary accuracy then belong to the temporal stage rather than to language-model adaptation. A shared localiser also provides a controlled interface for external audio–language models, whose native output formats and temporal capabilities are otherwise different.

The separation is useful even when the final task requires a single answer. It gives two checks that an aggregate score cannot provide. A semantic intervention should change class metrics while leaving temporal outputs fixed. A temporal intervention should change interval metrics while preserving the semantic proposal under the same language model. These checks distinguish an improved component from a later formatting or rewriting effect. They also make external comparisons more conservative: the shared interface measures behaviour under a common protocol, not the full native capability of each model.

This design makes negative results informative. A router may learn that a clip is noisy without learning which tool is better for that particular clip. A prompt may induce abstention without making the answer depend on the audio evidence. An adapter may improve classes represented during adaptation and fail on an unseen class. Each case separates a behavioural change from a supported mechanism claim, which is the level of analysis needed when a single end-to-end score would hide the source of improvement.

On the primary sealed set, stage specialisation raises exact match from 0.639 to 0.736 without changing any predicted interval. The gain recurs on a later split, while evaluation on real ESC-50 foregrounds shows a clear transfer boundary: performance improves on the adapted classes and reverses on Cat. With the model fixed, the spectral-contrast localiser also improves temporal IoU over the energy tool. A learned router develops a low-SNR preference but does not outperform the fixed tool.

The evaluation follows this attribution logic. The primary split tests whether the semantic change is local to the proposal stage. The later split tests whether the sign of the paired effect survives new foreground sources. The cross-corpus split changes the recording source and tests whether transfer depends on the classes represented during adaptation. The localisation and routing analyses then ask whether a better temporal tool is sufficient, and whether a learned choice between tools can recover the remaining headroom. Together these checks turn the main claim into a sequence of interventions rather than a single best score.

The protocol does not require the temporal tool to understand the class, and that restriction is intentional. A label-free tool can provide evidence about where acoustic activity occurs without becoming a second semantic classifier. If the proposal changes after such a tool is called, the change cannot be credited to a hidden class prediction inside the tool. Conversely, freezing the proposal while replacing the localiser prevents a better interval from being mistaken for better recognition. This division does not prove that the stages are independent in every audio task. It identifies the conditions under which a measured change can be assigned to one stage: the interface is fixed, the other stages are held constant, and the unchanged output field is checked explicitly. These conditions matter for the interpretation of both positive and negative results. A gain that survives the semantic intervention is useful evidence about proposal quality, while a gain that appears only after changing several stages remains a property of the whole pipeline.

The empirical story is therefore about separability, not a universal ranking of audio–language models. The semantic comparison uses one shared temporal tool and reports answer metrics; it is not a replacement for native end-to-end evaluation. The cross-corpus study changes the foreground recordings and exposes a class-dependent transfer boundary, but the mixtures and backgrounds remain controlled. The temporal comparison isolates the localiser, yet low-SNR clips and long real events remain difficult. These distinctions are kept in the main claims so that the reported gains are not read as evidence for broader coverage than the evaluation provides.

Contributions.

1. **Field-level attribution.** A source-disjoint protocol evaluates class prediction and temporal localisation separately. Paired interventions hold the remaining stages fixed, so the unchanged field is part of the evidence rather than an unreported side effect (§4, §4).
2. **A semantic-stage intervention.** The proposal-only adapter improves exact match on two sealed sets without changing any evaluated interval. Real foreground recordings show that the gain transfers to adapted classes but reverses on an unseen class (§3, §4).
3. **A separated evidence account.** Shared-tool and fixed-model comparisons isolate semantic and localisation effects. The analysis reports both positive effects and failure cases for low-

SNR localisation, learned routing, prompt-based abstention, and unseen-class transfer (§4, §4).

The evidence is deliberately bounded. The study covers three classes per set, 99 sources and one backbone. Backgrounds and mixtures are synthetic, including for the real ESC-50 foreground recordings. The learned router does not improve on the fixed tool, and abstention follows the prompt rather than the audio (§4).

2 Related Work

Instruction-tuned audio–language models. Coupling an audio encoder to a language model produces systems that answer open-ended questions about sound (Gong et al., 2023; Tang et al., 2024; Chu et al., 2023; 2024; Kong et al., 2024; 2025; Ghosh et al., 2026; Xu et al., 2025), evaluated on multiple-choice or free-text benchmarks (Yang et al., 2024; Sakshi et al., 2024; Lee et al., 2025) that score *what* was heard; none requires a verifiable *when*, so an answer that is right for the wrong moment scores the same as one supported by the correct segment. We keep the backbone family and change the contract.

Evidence-grounded audio reasoning and auditing. A recent line makes the evidence behind an audio answer a first-class object (Xie et al., 2026; Wang et al., 2026; Yu et al., 2026; Chen et al., 2026). We do not claim to introduce evidence-grounded audio training; our difference is methodological. Those systems improve an end-to-end score, whereas an action trace is not by itself evidence that each observation caused the final decision; we separate the semantic and temporal stages, hold one fixed while intervening on the other, and attach the same tool to every competing system, so that each gain traces to one stage and cross-model comparisons are not contaminated by localisation machinery.

Reinforcement learning for audio models. Group-relative policy optimisation (Shao et al., 2024) and related recipes (Schulman et al., 2017) have been applied to audio reasoning through curricula, logic rewards and adaptive thinking (Wen et al., 2025; Diao et al., 2025; Wu et al., 2025; Song et al., 2026; Ma et al., 2025); ours is a two-action router over tools, reported as a negative result on which the main claims do not depend.

Sound event detection and soundscape synthesis. Strongly labelled corpora and synthesis tooling (Gemmeke et al., 2017; Turpault et al., 2019; Salamon et al., 2017; Fonseca et al., 2021; Piczak, 2015) supply the onsets and offsets that make a temporal claim checkable, and pre-trained detectors such as BEATs (Chen et al., 2023) are strong when the target event matches their label space. We use one as a tool rather than a competitor, because a short foreground event in continuous background is a regime in which a general detector reports the background instead (§4); no tool here reads a label or a gold interval, which is what allows the same tool to be attached to every system compared.

3 Method

The design has three parts and one contract. The main system uses a frozen Qwen2-Audio-7B-Instruct backbone (Chu et al., 2024) and low-rank adapters (Hu et al., 2022); an optional routing diagnostic adds a learned policy. The class and interval are produced by different stages under separate contracts, allowing their changes to be measured independently.

3.1 Stage-specialised semantic adaptation

A single adapter trained to do the whole protocol has to satisfy two objectives at once: name the class and emit a well-formed action sequence with timing. In preliminary joint training, increasing pressure on class accuracy degraded action format and temporal output. On a held-out confirmation split, proposal accuracy also dropped when the source distribution changed (see the supplementary material). These observations motivate separating the semantic and protocol objectives; they do not by themselves establish proposal drift as the sole semantic bottleneck. We instead use two adapters. The *protocol adapter* is trained on complete three-step trajectories with counterfactual supervision

and controls tool syntax, state transitions and the final serialisation. The *stage-specialised proposal adapter* is trained on PROPOSE states only, with an activity-inspection curriculum and a coarse full-interval target, and is active only for the first action. The controller then passes the proposal to the acoustic tool and restores the protocol adapter for the final response. Its objective prioritises class identity while retaining a coarse interval target for the protocol interface. The semantic comparison changes only this adapter; the protocol adapter and tool that determine the final interval remain fixed. Predicted intervals are unchanged on every evaluated clip, localising the measured answer change to the semantic decision under this protocol.

3.2 FROZEN LABEL-FREE LOCALISATION

The fixed spectral-contrast localiser reads only the waveform and sample rate; it uses no class label, question, proposal, gold interval, or learned parameters. The same tool can therefore be attached to each competing model. It frames the clip with $25\,\mathrm{ms}$ Hann windows and a $10\,\mathrm{ms}$ hop, then computes the log-magnitude spectrum and log RMS for each frame. To reduce spectral-gain effects, it subtracts each frame's median across frequency and estimates the continuous background with the temporal median of the normalised spectra. For frame t, it scores positive deviations from that background in energy, spectral shape and spectral flux,

$$d_t = \mathbf{w}^\top \mathbf{z}_t^+, \tag{1}$$

where $\mathbf{z}_t^+$ is the feature vector of median-scaled positive deviations in energy, spectral shape, and spectral flux. The fixed, nonnegative coefficient vector $\mathbf{w}$ is kept implicit here; its values and the score-processing settings are reported in the supplementary material. After a seven-frame moving average, frames above 0.6 of the range between the median and the 0.98 quantile of d_t are active. We fill short gaps, drop short runs (see the supplementary material), and return the highest-scoring run, capped at $1.0\,\mathrm{s}$, with a confidence. The *energy* tool it replaces thresholds log RMS alone at its 0.7 quantile. This comparison isolates the contribution of spectral shape beyond amplitude.

A second tool is *proposal-conditioned*: it maps the proposed class to an AudioSet channel of a pre-trained event detector (Chen et al., 2023), centres the interval on that channel's frame-score peak, and sets its length from a class-duration prior estimated on training and development manifests. It is weaker overall (0.317 versus 0.464 IoU) but wins on a subset of $0\,\mathrm{dB}$ clips, and a wrong proposal points it at the wrong channel; the two tools fail for different reasons, which motivates §3.4.

3.3 THE NO-EVIDENCE-NO-REWRITE CONTRACT

Both tools return an interval and confidence, but no class evidence. The final stage therefore needs a rule that uses the new temporal observation without changing the semantic decision. Let the proposal be $(c_\mathrm{p}, \tau_\mathrm{p})$ and the tool return interval τ_t. The class-retaining finalisation rule maps these to the final output:

$$\mathcal{C}_\mathrm{nr} : \big(\mathcal{Y}\times\mathcal{T}\big)\times\mathcal{T} \to \mathcal{Y}\times\mathcal{T}, \qquad \mathcal{C}_\mathrm{nr}((c_\mathrm{p},\tau_\mathrm{p}),\tau_\mathrm{t}) = (c_\mathrm{f},\tau_\mathrm{f}) = (c_\mathrm{p},\tau_\mathrm{t}). \tag{2}$$

The first component is invariant to the tool observation, while the second component is replaced by the returned interval. Thus $c_\mathrm{f} = c_\mathrm{p}$ and $\tau_\mathrm{f} = \tau_\mathrm{t}$ hold for every tool output. The contract is part of the evaluated protocol and was fixed before test labels were examined.

3.4 OPTIONAL LEARNED TOOL ROUTING

Given two tools with complementary failure modes, a natural question is whether the choice between them can be learned without labels. The router sees four label-free statistics of the waveform: the median and 0.8 quantile of log RMS, the mean spectral-shape novelty, and the 0.99 quantile of d_t. It receives nothing about SNR, position, proposal correctness, or gold intervals. It chooses $a = 0$ (spectral tool) or $a = 1$ (proposal-conditioned tool). The training reward combines interval overlap, boundary accuracy, a thresholded localisation bonus, and a cost for the proposal-conditioned tool:

$$R(a) = \mathrm{IoU}_a + 0.5\,\mathrm{BF1}_a + 0.5\,\mathbf{1}\{\mathrm{IoU}_a \geq 0.5\} - 0.08\,\mathbf{1}\{a = 1\} \tag{3}$$

where $\mathbf{1}[\cdot]$ is the indicator function and action $a = 1$ selects the heavier proposal-conditioned detector. A linear Bernoulli policy samples multiple actions per state. Rewards are centred and scaled within each group (Shao et al., 2024), followed by one Adam update under a controlled optimisation

protocol; three independent retrainings assess repeatability (see the supplementary material). This group-relative policy-gradient diagnostic does not use a clipped objective. As reported in §4, the learned policy does not improve over the fixed spectral tool.

Training and evaluation sources are compared by normalised foreground-source identity, and the router trainer aborts if any source cluster appears on both sides. The proposal adapter is selected on development data and the tool configuration frozen before any sealed evaluation; sealed labels are read only to compute metrics, and model outputs, tool intervals and routing actions are stored per clip so that every aggregate in the paper can be reconstructed.

4 EXPERIMENTS

Overview. An instance is a 10-second waveform x containing continuous background and exactly one short foreground event, together with a question naming the candidate label set. A system must emit a parsable triple $\langle \texttt{answer} \rangle\, c$; $\langle \texttt{evidence} \rangle\, [\, t_\mathrm{s}, t_\mathrm{e}\,]$; $\langle \texttt{reason} \rangle\, r$, where c is the foreground class, $[t_\mathrm{s}, t_\mathrm{e}]$ is the interval offered as support, and r is a short rationale. Gold annotation gives the foreground class $c^\star$ and the true onset/offset $[t_\mathrm{s}^\star, t_\mathrm{e}^\star]$ of the foreground event, so both halves of the output are checkable independently: the class by exact match after normalisation, the interval by temporal IoU and by boundary F1 with a 0.5 s tolerance on each edge. Joint success requires the class to be correct *and* $\mathrm{IoU} \geq 0.5$ on the same clip; it is the metric a downstream consumer of the evidence cares about, and it is lower than either component for every evaluated system. Because the foreground is short, with gold events averaging 0.65 s inside a 10 s recording, predicting the whole clip is not a shortcut: the constant interval $[0, 10]$ scores 0.065 IoU on the sealed set.

Every system in this paper runs the same three steps (Figure 1a). In PROPOSE the model reads the waveform and emits a candidate class with a coarse interval. In INSPECT the environment, not the model, executes an acoustic tool that receives the waveform and its sample rate and nothing else: no label, no gold interval, no question text. In FINALISE the model turns the proposal and the tool observation into the final answer, evidence and rationale. Here, "agent" denotes a fixed, auditable tool protocol rather than open-ended planning. Applying the same loop and tool to every system controls localisation in the semantic comparison. Differences between external systems also reflect their model families and prompts.

The primary sealed set uses foreground sources excluded from training, calibration, and all auxiliary evaluation splits; the separate diagnostic protocols are summarised in the supplementary material. The source-overlap count with prior splits is zero. Soundscapes are synthesised so that onset, offset and SNR are known exactly rather than estimated (Turpault et al., 2019; Salamon et al., 2017): 12 independent sources, four each of Alarm bell ringing, Dog and Speech, rendered on a $\{0, 15\}$ dB $\times$ $\{$early, mid, late$\}$ grid into 72 clips (12 per condition); a fourth class lacked enough unused sources, and re-using sources would have broken the disjointness the design exists to protect. The manifest and audio package were frozen before any system in §4 was run against them, and the archived evaluation record retains every per-clip prediction.

After every system had been scored on the primary sealed set, we built a second sealed split from all foreground sources not used by training, calibration, the primary sealed set, or any auxiliary diagnostic split: 18 alarm, 6 dog and 3 speech sources, hence 27 source clusters and 162 clips on the same grid, rendered by the same script from the same background pool (gold events average 0.70 s). It is class-imbalanced by construction, which is why we report it as a replication of the paired comparisons rather than as a second benchmark; nothing was changed between the splits.

To test the same comparison on foreground sounds that are real recordings, we took 20 recordings each of Alarm, Cat and Dog from ESC-50 (Piczak, 2015), one event per recording (the shortest span holding 85% of the clip's energy, accepted when 0.15–2.0 s long), and rendered them with the unchanged script over the same DESED background pool on the same grid: 60 source clusters, 360 clips, gold events averaging 1.10 s. The set was registered as a confirmation set before any system was run on it and selected nothing. Cat is a class that the single adapter's training data contain but the proposal adapter's 48 adaptation clips do not; §4 reports it separately for that reason.

Baselines and metrics. All systems follow the same protocol on the sealed manifests. Semantic comparisons share one frozen spectral tool, so temporal predictions are identical and answer scores

Table 1: Semantic comparison on the sealed source-disjoint set ($n = 72$, 12 source clusters). All rows use the same frozen temporal tool and answer contract. Δ is the paired difference in answer exact match relative to the proposed system.

Semantic stage	Answer EM	Joint	ΔEM vs. ours [95% CI]
Stage-specialised (ours)	**0.736**	**0.319**	*Ref.*
Single adapter (ours, no stage split)	0.639	0.292	$+0.097$ [$+0.014$, $+0.194$]
Qwen2.5-Omni-7B	0.319	0.181	$+0.417$
Audio Flamingo Next	0.292	0.097	$+0.444$
Audio Flamingo 3 (7B)	0.153	0.069	$+0.583$
Qwen2-Audio-7B-Instruct (backbone)	0.056	0.014	$+0.681$

Table 2: Localisation under a fixed semantic stage on the sealed set. The table compares temporal tools and a per-clip oracle.

Localiser	Temporal IoU	Boundary F1	IoU ≥ 0.5
Frozen spectral contrast (ours)	**0.464**	**0.625**	**0.444**
Proposal-conditioned detector tool	0.317	0.472	0.333
Energy activity tool (previous default)	0.293	0.375	0.222
Per-clip oracle over the two tools	0.571	0.750	0.597
Constant full-clip interval	0.065	0.000	0.000

isolate the semantic stage. We compare our stage-specialised adapter with the same backbone using one protocol adapter, three external zero-shot audio–language models (Xu et al., 2025; Kong et al., 2025; Ghosh et al., 2026), and the unadapted Qwen2-Audio backbone. For localisation, we compare the proposal-conditioned detector, the energy tool, and a per-clip oracle over both; direct use of a pre-trained detector (Chen et al., 2023) appears in §4.

We use greedy decoding, at most 64 tokens per action, three protocol actions, ten-second audio, and fixed seeds. External baselines keep their released weights but receive the same rubric prompt and temporal-evidence contract. Metrics are exact-match answer accuracy, temporal IoU, boundary F1 (both edges within 0.5 s), IoU ≥ 0.5 rate, joint success, and format validity.

Data separation and uncertainty. The primary set has 72 clips from 12 independent sources. We average each source's six renderings and bootstrap source clusters (10,000 replicates) for all 95% intervals; comparisons use paired source-level differences. The same procedure covers 27 replication sources, 60 cross-corpus sources, and 99 pooled sources.

We froze each manifest and audio package before evaluation, checked the router's training data for source overlap, and did not rerun configurations after inspecting sealed-set scores. Archived records retain configurations, seeds, parser settings, and per-clip outputs.

Semantic comparison. Table 1 isolates the semantic stage: every system uses the same frozen spectral tool, so their temporal predictions are identical. Stage specialisation raises exact match from 0.639 to 0.736; all 72 predicted intervals remain unchanged. Eight of the nine changed answers move from incorrect to correct, placing the gain in the proposal stage.

External comparisons. Under the shared tool and answer contract, the adapted proposal stage scores above each of the four zero-shot baselines. This comparison isolates semantic answers; it does not rank the baselines' native end-to-end systems. Their main error is naming the continuous background instead of the brief foreground event (§4).

Table 3: Semantic results on the later DESED split, the ESC-50 cross-corpus set, and the pooled evaluation. Values are answer exact match; bracketed values are source-cluster bootstrap intervals for paired contrasts.

	Replication split, DESED ($n = 162$)		Cross-corpus set, ESC-50 ($n = 360$)		Pooled, three sets ($n = 594$)	
Semantic stage	Answer EM	ΔEM vs. ours	Answer EM	ΔEM vs. ours	Answer EM	ΔEM vs. ours
Stage-specialised (ours)	**0.802**	*Ref.*	**0.714**	*Ref.*	**0.741**	*Ref.*
Single adapter	0.599	$+0.204$ $[+0.130, +0.284]$	0.661	$+0.053$ $[-0.017, +0.117]$	0.641	$+0.099$ $[+0.051, +0.148]$
Qwen2.5-Omni-7B	0.333	$+0.469$	0.453	$+0.261$	0.404	$+0.337$
Audio Flamingo Next	0.265	$+0.537$	0.564	$+0.150$	0.449	$+0.291$
Audio Flamingo 3	0.142	$+0.660$	0.333	$+0.381$	0.259	$+0.481$
Backbone (Qwen2-Audio)	0.049	$+0.753$	0.094	$+0.619$	0.077	$+0.663$
Ours − single adapter on ESC-50, by class:	Alarm $+0.167$ $[+0.092, +0.258]$; Dog $+0.142$ $[+0.067, +0.225]$; Cat -0.150 $[-0.275, -0.042]$; Alarm and Dog together $+0.154$ $[+0.100, +0.217]$.					

Replication and transfer. The later DESED split repeats the gain: exact match is 0.802, with a paired improvement of $+0.204$ over the single adapter. On real ESC-50 foreground recordings, the gain holds for adapted Alarm and Dog but reverses for Cat, which was absent from adaptation. The overall ESC-50 contrast is uncertain, so the evidence supports transfer to new recordings of adapted classes, not to unseen classes. Across all three sets, pooled exact match is 0.741. Two independent retrainings retain a positive paired estimate on every set; they support the direction of the effect, not a seed-averaged absolute score.

Temporal localisation. With the semantic stage fixed, the spectral-contrast localiser raises temporal IoU from 0.293 to 0.464 and boundary F1 from 0.375 to 0.625 (Table 2). The per-clip oracle reaches 0.571 IoU, leaving a tool-selection problem for future work.

Errors separate into acoustic and semantic limits. Localisation weakens more at 0 dB than class recognition, and on ESC-50 the tool often selects one short, high-contrast segment within a longer event. Raising its duration cap recovers little of the lost overlap. Alarm-to-Dishes confusion remains the main semantic error on the sealed sets.

Routing. The learned router calls the semantic tool on ten sealed clips, mostly at 0 dB, but does not improve temporal IoU, boundary F1, or joint success. An SNR gate also fails to produce a consistent gain across sets. The evaluated global waveform statistics do not identify which clips benefit from switching tools, so we retain the fixed localiser. Relative to the fixed tool, the router changes IoU by -0.010 with a 95% interval of $[-0.102, +0.059]$, boundary F1 by $+0.014$ with $[-0.097, +0.111]$, and joint success by $+0.014$ with $[-0.069, +0.097]$. The per-clip oracle reaches 0.571 IoU, which leaves headroom but does not identify a deployable routing rule.

Failure analysis and boundaries. The low exact match of the four zero-shot LALMs in Tables 1 and 3 is not a formatting failure: on the replication split every one of their 162 outputs yields a readable class label. What they name is the background. The most frequent answer of every zero-shot model on that split is "Frying", the continuous background: the un-adapted backbone labels 117 of 162 clips as a background source, and the other three answer "Frying" on 49–52 clips each (Qwen2.5-Omni declines to answer on another 49). A pre-trained sound event detector (Chen et al., 2023) applied directly scores 0.000 for the same reason, yet reaches 0.317 IoU used *as a tool* (Table 2). The models answer "what is this recording of?" rather than "which short event interrupts it?", and most of the semantic gap in the tables is that foreground/background distinction, which is what the proposal adapter learns from 48 clips. On the cross-corpus set, where Dog and Cat are stock pre-training classes, the background-label rate of the three instruction-tuned zero-shot models falls from 0.36–0.45 to 0.22–0.30 and their scores rise accordingly. Alarm, the class most easily confused with an appliance background, stays near zero for three of the four models (the supplementary material).

Three sealed sets total 594 clips from 99 sources in two corpora, three classes each, and the proposal adapter was adapted on 48 clips. We therefore emphasize paired differences rather than treating an absolute point estimate as a benchmark result. The main tables use the designated proposal-adapter

evaluation configuration; an independent retraining study checks the direction of the paired gain across all nine seed–set combinations. Those estimates have the same sign, and the adapted-class pattern repeats. The adapter's gain did not extend to Cat, a class absent from its adaptation clips: the stage split isolates *where* semantics are learned, not which classes. The supplementary material reports the corresponding variation in absolute scores.

Onsets, offsets and SNR are exact because the soundscapes are rendered rather than recorded (Turpault et al., 2019; Salamon et al., 2017), the main external-validity risk. The cross-corpus set makes the foregrounds real recordings but not the backgrounds or the mixing, and exposed a second cost of freezing: on ESC-50 events averaging 1.10 s the tool returns 0.49 s runs (its 1 s cap bounds attainable IoU at 0.841).

Every adapted system starts from Qwen2-Audio-7B-Instruct, and we have not shown that the argument transfers. Learned routing does not beat the fixed tool. Abstention follows the prompt, not the audio: with the supporting interval zeroed and no warning no system abstains (the answer changes instead), and with the warning alone every clip is abstained (the supplementary abstention analysis). We claim no general state of the art, no coverage beyond the classes evaluated, and no general autonomous planning.

What the interventions identify. The three failure patterns are related, but they do not have the same diagnostic meaning. Background answers are a semantic error: the model has selected a plausible sound from the recording rather than the short event defined by the task. The proposal-only intervention addresses this error because it changes the class decision before the tool is called. Its effect is therefore visible in exact match while the interval remains fixed. This is also why a better temporal tool cannot be expected to repair a wrong class under the no-evidence-no-rewrite contract. The tool receives no class label and does not return class evidence, so its observation is deliberately insufficient to change that decision.

Low-SNR localisation errors have a different signature. At 0 dB, the foreground and the continuous background share enough energy that the label-free tool often selects a short high-contrast fragment instead of the full event. The class can remain correct while IoU and boundary F1 fall. The duration-cap check is useful here because extending the cap recovers only a small portion of the missing overlap. This rules out a simple explanation in which the localiser fails only because its output window is too short. The remaining error is a mismatch between acoustic contrast and event extent, which is consistent with the weaker performance on longer real foreground recordings.

The routing result supplies a third piece of evidence. The two tools have different inputs and failure modes, but the available waveform summaries do not reliably predict which failure mode will occur on a particular clip. The router's low-SNR preference is therefore not enough to produce a better selection policy. This negative result prevents us from treating tool composition as an automatic solution to localisation. Taken together, the interventions support a narrower conclusion: the protocol can assign a measured semantic gain to the proposal stage and a measured temporal gain to the localiser when the interface and the other stages are fixed. It does not show that the stages are independent, that the router can exploit all complementary tools, or that the same attribution will hold for other backbones, event inventories, or recording conditions.

5 Conclusion

Treating *what* a sound is and *when* it occurs as separate stages exposes effects that a single end-to-end score would conflate. Stage specialisation improves exact match by $+0.097$ while all 72 intervals remain unchanged, supporting attribution to the semantic stage under the controlled protocol. The direction repeats on two further sealed sets and across three training seeds, with an unseen class marking the boundary of the effect. The frozen spectral tool improves temporal IoU by $+0.171$ while the model is held fixed, supporting attribution to the localisation stage. Under this controlled protocol, the result separates the semantic gain from the temporal gain while leaving 0 dB localisation and unseen classes as explicit boundaries of the method.

Ethics Statement

The evaluation data are soundscapes synthesised from publicly released sound event corpora (DESED, ESC-50) under their original licences; no new recordings of people were collected, and the Speech class consists of pre-existing corpus material used as an acoustic event rather than for speaker or content analysis. The systems studied localise sound events in audio, a capability that has surveillance applications; the specific models here operate on 10-second clips with three event classes and do not perform speaker identification or speech transcription. We report a negative result and the failure boundaries of the method explicitly so that the capability is not overstated.

Reproducibility Statement

The three sealed evaluation manifests and their rendered audio packages were fixed before any system in this paper was run against them. §4 specifies the source-disjointness property, the condition grid and how the replication split and the cross-corpus set were built, §3 the adapter configuration, the frozen tool parameters and the routing hyper-parameters, and §4 the decoding settings and the clustered bootstrap procedure used for every interval in the paper. The routing trainer verifies programmatically that no source cluster is shared between its training data and the sealed set and aborts otherwise. An anonymised code archive with the evaluation scripts, the manifest and the per-clip predictions behind every table is provided as supplementary material; each record includes the evaluation configuration, model seed and parser settings needed for reconstruction.

Use of Large Language Models

Large language models were used as writing and coding assistants: drafting and editing prose, and helping to write evaluation and analysis scripts that the authors reviewed. They were not used to generate experimental results, research ideas presented as novel, or any number reported in this paper; all reported values are produced by the scripts in the supplementary material and were checked against the raw per-clip prediction files by the authors.

References

Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, and Furu Wei. BEATs: Audio pre-training with acoustic tokenizers. In *International Conference on Machine Learning (ICML)*, 2023.

Yanxi Chen, Wenhui Zhu, Xiwen Chen, Zhipeng Wang, Xin Li, Peijie Qiu, Hao Wang, Xuanzhao Dong, Yujian Xiong, Anderson Schneider, Yuriy Nevmyvaka, and Yalin Wang. AHA: Aligning large audio-language models for reasoning hallucinations via counterfactual hard negatives. In *Findings of the Association for Computational Linguistics (ACL Findings)*, 2026.

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. Qwen-Audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*, 2023.

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. *arXiv preprint arXiv:2407.10759*, 2024.

Xingjian Diao, Chunhui Zhang, Keyi Kong, Weiyi Wu, Chiyu Ma, Zhongyu Ouyang, Peijun Qing, Soroush Vosoughi, and Jiang Gui. SoundMind: RL-incentivized logic reasoning for audio-language models. *arXiv preprint arXiv:2506.12935*, 2025.

Eduardo Fonseca, Xavier Favory, Jordi Pons, Frederic Font, and Xavier Serra. FSD50K: An open dataset of human-labeled sound events. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2021.

Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset

for audio events. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

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. Audio flamingo next: Next-generation open audio-language models for speech, sound, and music. *arXiv preprint arXiv:2604.10905*, 2026.

Yuan Gong, Alexander H. Liu, Hongyin Luo, Leonid Karlinsky, and James Glass. Listen, think, and understand. *arXiv preprint arXiv:2305.10790*, 2023.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

Zhifeng Kong, Arushi Goel, Rohan Badlani, Wei Ping, Rafael Valle, and Bryan Catanzaro. Audio flamingo: A novel audio language model with few-shot learning and dialogue abilities. In *International Conference on Machine Learning (ICML)*, 2024.

Zhifeng Kong, Arushi Goel, Rohan Badlani, Wei Ping, Rafael Valle, and Bryan Catanzaro. Audio flamingo 2: An audio-language model with long-audio understanding and expert reasoning abilities. *arXiv preprint arXiv:2503.03983*, 2025.

Tony Lee, Haoqin Tu, Chi Heem Wong, Zijun Wang, Siwei Yang, Yifan Mai, Yuyin Zhou, Cihang Xie, and Percy Liang. AHELM: A holistic evaluation of audio-language models. *arXiv preprint arXiv:2508.21376*, 2025.

Ziyang Ma, Zhuo Chen, Yuping Wang, Eng Siong Chng, and Xie Chen. Audio-CoT: Exploring chain-of-thought reasoning in large audio language model. *arXiv preprint arXiv:2501.07246*, 2025.

Karol J. Piczak. ESC: Dataset for environmental sound classification. In *ACM International Conference on Multimedia (MM)*, 2015.

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. *arXiv preprint arXiv:2410.19168*, 2024.

Justin Salamon, Duncan MacConnell, Mark Cartwright, Peter Li, and Juan Pablo Bello. Scaper: A library for soundscape synthesis and augmentation. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2017.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Zhiyuan Song, Weici Zhao, Yang Xiao, Suhao Yu, Cheng Zhu, and Jiatao Gu. Learning when to think while listening in large audio-language models. *arXiv preprint arXiv:2605.27190*, 2026.

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 *International Conference on Learning Representations (ICLR)*, 2024.

Nicolas Turpault, Romain Serizel, Ankit Parag Shah, and Justin Salamon. Sound event detection in domestic environments with weakly labeled data and soundscape synthesis. In *Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2019.

Jieyi Wang, Yazhe Niu, Dexuan Xu, and Zhongyu Wei. Listen, pause, and reason: Toward perception-grounded hybrid reasoning for audio understanding. In *Findings of the Association for Computational Linguistics (ACL Findings)*, 2026.

Cheng Wen, Tingwei Guo, Shuaijiang Zhao, Wei Zou, and Xiangang Li. SARI: Structured audio reasoning via curriculum-guided reinforcement learning. *arXiv preprint arXiv:2504.15900*, 2025.

Shu Wu, Chenxing Li, Wenfu Wang, Hao Zhang, Hualei Wang, Meng Yu, and Dong Yu. Audio-Thinker: Guiding audio language model when and how to think via reinforcement learning. *arXiv preprint arXiv:2508.08039*, 2025.

Xinyuan Xie, Shunian Chen, Zhiheng Liu, Yuhao Zhang, Zhiqiang Lv, Liyin Liang, and Benyou Wang. EvA: An evidence-first audio understanding paradigm for LALMs. *arXiv preprint arXiv:2603.27667*, 2026.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen2.5-Omni technical report. *arXiv preprint arXiv:2503.20215*, 2025.

Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun Lv, Zhou Zhao, Chang Zhou, and Jingren Zhou. AIR-Bench: Benchmarking large audio-language models via generative comprehension. *arXiv preprint arXiv:2402.07729*, 2024.

Fangxu Yu, Tao Feng, Dehai Min, Zinan Lin, Weijia Xu, Michael Xu, Philip S. Yu, Ge Liu, and Tianyi Zhou. Reinforcement learning with evolving rubrics as rewards for audio reasoning. *arXiv preprint arXiv:2608.02831*, 2026.