RASA: ROLE-SPECIFIC ACOUSTIC–SEMANTIC ALIGNMENT FOR TELECOM FRAUD DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Distinguishing fraudulent calls from legitimate conversations requires interpreting callers' claims and requests within the surrounding conversational context. Acoustic behaviors such as disfluency, speech rate acceleration, and frequent interruptions can supplies crucial acoustic evidence when assessing deceptive intent. However, transcript-based large language models may overlook these cues, while direct audio input into large audio-language models (LALMs) does not guarantee their adequate preservation by recognition-oriented encoders. So that we propose Role-specific Acoustic–Semantic Alignment (RASA), which combines a recognition-oriented speech encoder with a reconstruction-oriented neural audio codec. Separate Fixed-Length Query Adapters aggregate their representations into compact soft-token sequences for joint reasoning within an LLM. Using transcript- and script-informed teacher supervision from ParaScam, our synthesized call dataset, role-specific alignment establishes a semantic anchor and trains the acoustic pathway to contribute acoustic evidence. The aligned model supports downstream adaptation using only audio and task labels, with no transcripts required at inference. On TeleAnti, RASA achieves 98.13% fraud-detection weighted F1 and 88.61% average weighted F1 across fraud detection, fraud-type identification, and scene classification. Under a 500-call downstream adaptation setting, RASA improves fraud-type F1 by 27.37 percentage points over its semantic-only variant, supporting the value of explicitly aligned acoustic evidence for fine-grained fraud recognition.

1 INTRODUCTION

Telecom fraud poses a serious threat to financial security and social trust, motivating reliable automated detection of fraudulent calls (Ma et al., 2025b; Wang et al., 2026d). However, costly manual annotation severely limits the availability of labeled training data, making it difficult to extract robust fraud patterns conventionally. Benefiting from the powerful analytical and reasoning capabilities of large language models (LLMs), recent approaches increasingly rely on them to infer deceptive intent from call transcripts (Shen et al., 2025; Ma et al., 2025a; Wu et al., 2026). Such reasoning is a foundation to capture the dynamic development of an interaction: requests for money transfers or identity verification may occur in legitimate calls, but the subtle context—such as attempts to prevent independent verification—provides the definitive evidence of deception.

However, calls provides additional acoustic evidence. A caller impersonating a professional representative may hesitate when challenged beyond a prepared script; in a fabricated emergency, repeated interruptions and speech rate acceleration by caller may restrict the callee's opportunity to question or verify. These behaviors acquire meaning through their semantic context, while transcripts omit their acoustic realization. Recent speech benchmarks and prosody-aware reasoning models likewise highlight the importance of interpreting non-lexical cues alongside linguistic content (Wang et al., 2026b;c). Direct audio input offers access to such evidence, yet audio encoder does not ensure its effective representation. Existing large audio-language model (LALM)-based fraud detectors adopt a Whisper-derived audio interface (Ma et al., 2025b; Wang et al., 2026d), whose recognition-orientedpretraining prioritizes semantic features without explicitly requiring the preservation of fraud-relevant acoustic evidence (Radford et al., 2023; Chu et al., 2024).

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.

To address this representation gap, we propose **Role-specific Acoustic–Semantic Alignment (RASA)**. RASA retains a Whisper-derived semantic pathway and introduces a reconstruction-oriented WavTokenizer pathway to supply crucial acoustic representations (Ji et al., 2024). Separate **Fixed-Length Query Adapters (FLQ-Adapters)** aggregate their variable-length outputs into compact soft-token sequences, controlling the audio-token budget while enabling joint semantic–acoustic reasoning within the LLM.

Learning complementary roles also requires guidance on the evidence each pathway should contribute. To provide this supervision in the absence of suitable fine-grained datasets and to overcome the scarcity of labeled data, we construct **ParaScam**, comprising 16,060 synthesized calls. We de-identify transcripts of unlabeled calls, enrich them with explicit acoustic descriptions, and synthesize corresponding calls. Each call is thus paired with a plain transcript and a delivery-enriched script. A text LLM teacher analyzes these two views to generate content-based and acoustically informed targets, respectively.

These analysis targets guide FLQ-Adapters to retain evidence useful for fraud reasoning. To guide the two pathways toward complementary roles, our role-specific alignment applies these supervision targets sequentially. It first trains the semantic adapter against transcript-conditioned analyses to establish a semantic anchor. With this anchor fixed, the acoustic adapter then learns from script-informed analyses to contribute additional acoustic evidence. For downstream adaptation, both adapters remain frozen while the LLM is adapted using only audio and task labels.

Evaluated on TeleAnti, RASA achieves 98.13% fraud-detection weighted F1 and 88.61% average weighted F1 across fraud detection, fraud-type identification, and scene classification. With 500 labeled training calls for downstream adaptation, it achieves 84.86% average weighted F1. Under this setting, RASA improves fraud-type F1 by 27.37 percentage points over its semantic-only variant.

Our main contributions are threefold:

1. We introduce **RASA**, combining complementary speech representations via fixed-length query adapters for joint semantic-acoustic reasoning.
2. We construct the **ParaScam** dataset and propose a role-specific alignment strategy to extract semantic anchors and acoustic evidence, overcoming the scarcity of fine-grained manual labels.
3. We evaluate RASA under full-data and limited-label adaptation settings, with ablations showing the contribution of acoustic alignment to fine-grained fraud-type recognition.

2 Related Work

2.1 Telecom Fraud Call Detection

Transcript-based LLM detectors analyze call content, but face challenges involving recall, hallucinations, and dataset bias (Shen et al., 2025). Limited annotations and changing deceptive narratives create further difficulties. Ma et al. (2025a) address these conditions through expert rules and hierarchical few-shot prompting, combining scam discrimination with reflection against normal-call patterns. Their framework also filters easy normal calls with a lightweight model to reduce large-LLM inference costs. Wu et al. (2026) instead organize detection around stable semantic primitives, identifying their presence in transcripts before combining them into a verdict. These studies motivate evaluation under scarce target-task labels, which we simulate by limiting TeleAnti task-specific adaptation to 500 labeled calls.

Audio-based fraud detection extends the evidence available to these systems. Ma et al. (2025b) introduce TeleAntiFraud-28k and fine-tune Qwen2-Audio for fraud detection and associated scene and fraud-type predictions. SAFE-QAQ builds on this model with reinforcement learning for structured reasoning and risk assessment during ongoing calls (Wang et al., 2026d). RASA addresses the fundamental audio representations used for these decisions: rather than relying on a single entangled encoder, it retains a robust semantic anchor and explicitly aligns crucial acoustic evidence prior to downstream task adaptation.

2.2 Large Audio-Language Models and Speech Alignment

LALMs connect audio representations to language models for speech understanding and instruction following. Beyond transcription, LTU and Audio Flamingo extend audio-language interfaces toward open-ended sound understanding and dialogue (Gong et al., 2024; Kong et al., 2024). Qwen2-Audio initializes its audio encoder from Whisper and trains on diverse audio tasks (Chu et al., 2024). Dual-encoder designs already broaden this interface:SALMONN combines Whisper and BEATs through a window-level Q-Former (Tang et al., 2024; Chen et al., 2023), while WavLLM combines Whisper and WavLM to capture semantic and speaker information (Hu et al., 2024). RASA builds on this dual-encoder paradigm through independently parameterized FLQ-Adapters and a role-specific learning strategy that assigns distinct supervision to the semantic and acoustic pathways.

Alignment objectives also determine how speech representations are interpreted. BLSP trains a speech adapter using transcript-conditioned text continuations to align speech- and text-conditioned LLM behavior (Wang et al., 2023a). BLSP-Emo first performs semantic alignment and then uses emotion-aware continuations for emotion alignment (Wang et al., 2024). Broader evaluations, including MMAU and Dynamic-SUPERB Phase-2, reveal substantial variation in audio-language models' capabilities across tasks (Sakshi et al., 2025; Huang et al., 2025). Recent work further studies speech–text reasoning gaps and process-level supervision for audio reasoning (Wang et al., 2026a; Fan et al., 2026). RASA instead investigates task-oriented representation alignment for complementary semantic and acoustic evidence. RASA shares the principle of establishing semantic competence before learning supplementary cues. Specifically, RASA freezes the semantic adapter after transcript-based alignment and trains the reconstruction-oriented acoustic adapter with ParaScam's script-informed targets, conditioning acoustic alignment on a fixed semantic pathway.

2.3 Acoustic Representation Learning

Existing approaches expose non-lexical speech information through specialized representations and supervision. emotion2vec learns emotion representations through self-supervised distillation at frame and utterance levels (Ma et al., 2024); ProsodyLM pairs transcripts with word-level prosody tokens to model content–prosody dependencies (Qian et al., 2025). The extent to which general audio models use such information also requires evaluation: on the LISTEN emotion benchmark, Chen et al. (2026) find that evaluated LALMs inherently favor lexical content over acoustic emotion cues, corroborating the need for explicit acoustic alignment.

Neural audio codecs learn compact representations through waveform reconstruction, as exemplified by EnCodec and DAC (Défossez et al., 2023; Kumar et al., 2023). SpeechTokenizer organizes semantic and acoustic information across quantization layers, while Language-Codec and ALMTokenizer further investigate the suitability of codec representations for speech and audio language modeling (Zhang et al., 2024; Ji et al., 2025a; Yang et al., 2025). WavTokenizer provides compact acoustic tokens with strong reconstruction quality and retained semantic information (Ji et al., 2025b). RASA uses its quantized codebook embeddings as a supplementary representation and learns their task-specific contribution relative to a fixed semantic pathway.

3 Task Description

In real-world scenarios, a massive number of calls take place every day. Our task is to identify the fraudulent calls among them. However, due to the high cost of manual annotation, only a limited number of calls are labeled. Therefore, we describe the Telecom Fraud Detection task as follows: given a limited training set $\mathcal{D}_{\text{train}} = \{(a_i, y_i)\}_{i=1}^{N}$, where a_i denotes a raw call waveform and y_i comprises the binary fraud label, fraud type when applicable, and communication scene, the objective is to jointly predict the corresponding labels for an unseen call a.

Thanks to the powerful language understanding and reasoning capabilities of large language models, existing Telecom Fraud Detection methods typically employ a two-stage pipeline: first, the call is transcribed into text via ASR (Automatic Speech Recognition), and then a large language model is used to perform in-depth analysis and reasoning on the transcribed text. While the transcribed text does reflect the complete semantic content of the call, it also contains abundant acoustic information beyond semantics, such as fluency, speech rate acceleration, and frequent interruptions. As explained earlier, these acoustic features also contain substantial evidence related to fraud. Therefore, fully

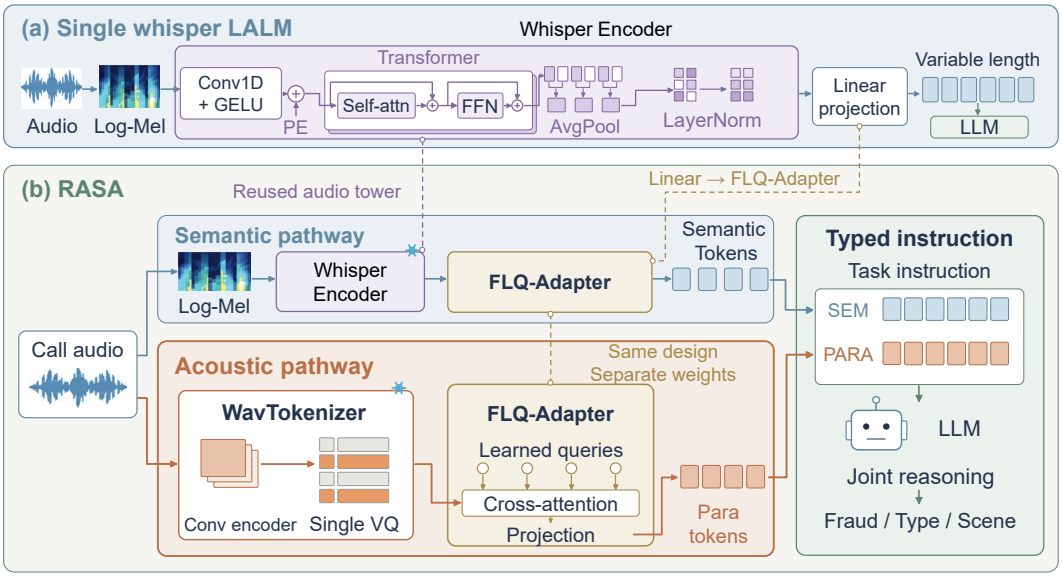


Figure 1: RASA architecture. Separate query adapters map Whisper-derived semantic features and WavTokenizer-derived supplementary features into typed soft tokens for joint fraud reasoning.

utilizing the acoustic information beyond the transcribed text of the call is crucial for improving the accuracy of fraud call detection.

Recent studies have attempted to utilize acoustic features from calls via Large Audio Language Models (LALMs), which can directly use audio as input, avoiding the loss of acoustic information caused by transcribing the call into text. The application of LALMs in Telecom Fraud Detection tasks can be described as illustrated in Figure 1(a), The call audio a is first extracted into a log-Mel spectrum and then processed into a whisper encoder to obtain the hidden state H_{whisper}:

$$H_{\text{whisper}} = E_{\text{whisper}}(\text{LogMel}(a)). \tag{1}$$

Subsequently, a linear adapter converts H_{whisper} into $Z^{\text{base}}_{\text{whisper}}$, which are inserted into the designated placeholders within the input sequence, allowing the LLM to perform in-depth analysis and inference.

Although the whisper pathway allows the LLM to process audio feature directly, it does not guarantee that crucial acoustic evidence is adequately preserved. As Whisper is optimized for speech recognition and translation through autoregressive token prediction:

$$\mathcal{L}_{\text{Whisper}} = -\sum_{\tau} \log p(c_\tau \mid c_{<\tau}, H_{\text{whisper}}). \tag{2}$$

where c_τ denotes a target token predicted by the Whisper decoder. This objective causes the Whisper encoder to naturally focus more on the semantic information in the audio while marginalizing the rich acoustic information. Consequently, LALMs cannot fully leverage the rich fraud-related features contained in the acoustic information in existing telecom fraud detection methods.

4 Method

4.1 Integrating Complementary Audio Representations

To better leverage the acoustic information in call audio, we explore neural audio codecs as an alternative to the Whisper encoder. As illustrated in Figure 1(b), the input call is first encoded into continuous latent features and then quantized into discrete acoustic representations:

$$H_{\text{wav}} = E_{\text{wav}}(a). \tag{3}$$

The motivation for this pathway lies in its reconstruction-oriented pretraining. While Whisper's token-prediction objective emphasizes semantic content without explicitly supervising fraud-relevant acoustic information, WavTokenizer is trained to reconstruct the input waveform:

$$\mathcal{L}_{\text{Wav}} \approx \mathcal{L}_{\text{recon}}(a, \hat{a}) + \lambda \mathcal{L}_{\text{adv}}, \tag{4}$$

where $\hat{a} = D_{\text{wav}}(H_{\text{wav}})$ is the waveform reconstructed by the codec's decoder, $\mathcal{L}_{\text{recon}}$ denotes the time- and frequency-domain reconstruction losses, and $\mathcal{L}_{\text{adv}}$ represents the adversarial penalty. This objective encourages the retention of acoustic features, which contain abundant fraud-related evidence that LLMs can leverage for in-depth analysis and reasoning.

Although substituting Whisper with WavTokenizer preserves abundant acoustic information, it introduces a critical trade-off: neural audio codecs excel at audio reconstruction but perform significantly worse at recover semantic content (Ji et al., 2024). This divergence is mathematically rooted in their exclusive training objectives. Autoregressive token prediction ($\mathcal{L}_{\text{Whisper}}$) forces H_{whisper} away from acoustic features, while waveform reconstruction ($\mathcal{L}_{\text{Wav}}$) forces H_{wav} preserving acoustic evidence at the expense of semantic information.To capture both modalities without compromise, we adopt the dual-channel architecture illustrated in Figure 1.

The relevance of acoustic evidence to fraud detection depends on the surrounding semantic context. We therefore organize the two pathways as a semantic anchor and a crucial source of acoustic evidence. Their outputs, Z_{whisper} and Z_{wav}, are inserted at distinct placeholders, `<SEM>` and `<PARA>`, respectively (Figure 3). This organization identifies the two sources and makes their representations jointly available to the LLM for contextual reasoning.

FLQ-Adapters. Since we chose to assemble Z_{whisper} and Z_{wav} in the LLM embedding space, we have to account for their duration-dependent sequence lengths. Because for dual pathway Z with length L, combining them increases the self-attention cost of LLM from $\mathcal{O}(L^2)$ to $\mathcal{O}((2L)^2)$, increasing the computational burden for long calls. However, fraud detection often hinges on a few suspicious semantic segments and their associated acoustic evidence. Following query-based latent bottlenecks and resampling interfaces (Jaegle et al., 2021; Alayrac et al., 2022; Li et al., 2023), we therefore use a Fixed-Length Query Adapter (FLQ-Adapter) to aggregate each encoder sequence into a fixed number of LLM-compatible tokens. For each pathway, we introduce M learnable queries Q that attend to the encoder output H:

$$Z = \text{MLP}(R) + e, \qquad R = \text{LN}(\text{MHA}(Q, H, H)), \tag{5}$$

where the learnable queries Q aggregate H through multi-head cross-attention. Layer normalization and a two-layer MLP produce LLM-compatible embeddings, to which a learned source embedding e is added at every query position to explicitly distinguish the two streams in the LLM embedding space. The independently parameterized adapters yield Z_{whisper} and Z_{wav}. Detailed adapter configurations and structural dimensions are provided in Appendix A.

4.2 Role-specific Alignment

While the pathway-specific source embeddings e explicitly distinguish the streams in the LLM's embedding space, this additive identifier alone cannot guarantee that the cross-attention mechanisms will naturally extract decoupled representations R_{whisper} and R_{wav}. We therefore introduce the sequential alignment procedure in Figure 2: we first establish a pure semantic foundation, and subsequently align the acoustic features conditionally.

ParaScam Construction. Implementing this role-specific alignment requires specialized supervision: a pure semantic target, a joint semantic-acoustic target, and the corresponding audio. Since existing datasets lack such fine-grained dataset, we construct ParaScam by leveraging the massive volumes of unlabeled calls a in real-world. To strictly preserve privacy, we first de-identify their ASR transcripts to obtain a plain transcript t, perfectly isolating the semantic content.

To completely eliminate original speaker identities, we must replace a with synthesized calls $\widetilde{a}$. As t no longer contains any acoustic information, we ultilize an LLM to augment each plain transcript t with descriptions of intended voice delivery (e.g., anxious interruptions, disfluencies) into t to guide the speech synthesis. A TTS model then synthesizes the corresponding waveform $\tilde{a}$ conditioned on

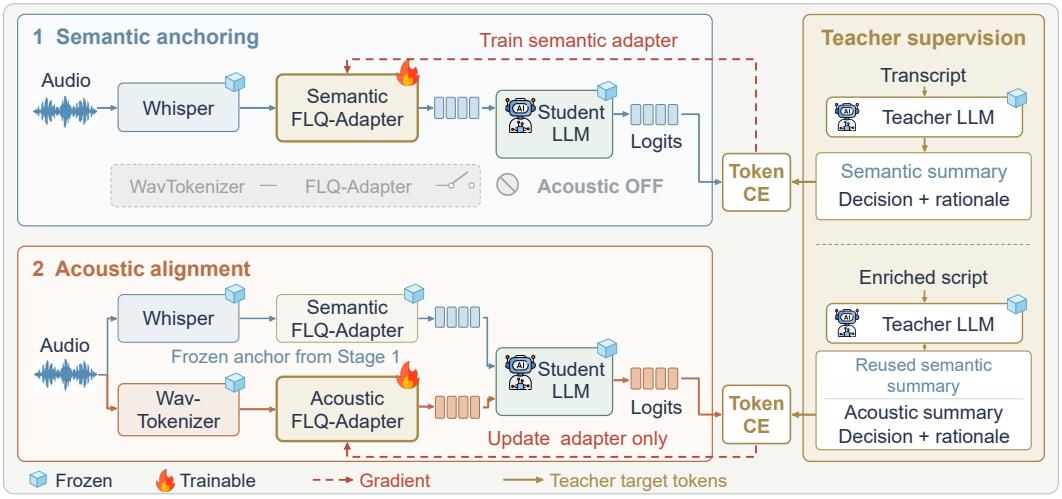


Figure 2: RASA role-specific training pipeline. The framework first establishes a semantic anchor using transcript-conditioned targets, then conditionally aligns the supplementary acoustic features using enriched scripts before task-specific adaptation.

these descriptions. Because this augmented text explicitly textualizes the dynamic acoustic behaviors, we directly adopt it as our enriched script $\widetilde{t}$ to serve as the joint semantic-acoustic supervision signal. This pipeline elegantly yields the exact $(\widetilde{a}, t, \widetilde{t})$ triplet. Construction details are deferred to Appendix B.

Equipped with the ParaScam dataset, we initiate the two-stage alignment process by establishing the semantic anchor. Because crucial acoustic evidence relies on semantic context, we first establish a pure content-based foundation. Using teacher-generated analyses as training targets follows the broader practice of transferring task knowledge through generated rationales (Ho et al., 2023; Hsieh et al., 2023), so a frozen text-only teacher $\mathcal{T}$ analyzes the plain transcript t under instruction I_{sem}, without access to the waveform or the enriched script's acoustic descriptions:

$$Y^{\text{sem}} = \mathcal{T}(t; I_{\text{sem}}), \tag{6}$$

To align the semantic pathway with the downstream reasoning task, I_{sem} requests a semantic summary, fraud prediction, and supporting rationale. The student processes the corresponding synthesized waveform $\tilde{a}$ with the acoustic pathway disabled. Under an instruction containing only the <SEM> placeholder, we optimize the semantic FLQ-Adapter:

$$\mathcal{L}_{\text{sem}} = -\sum_{j=1}^{|Y^{\text{sem}}|} \log p_{\theta_0}(v_j \mid v_{<j}, I_{\text{sem}}, Z_{\text{whisper}}), \tag{7}$$

Here, v_j denotes the j-th target token and θ_0 denotes the frozen LLM parameters. This objective trains the semantic FLQ-Adapter to produce audio embeddings that support generation of the transcript-conditioned teacher response.

With the semantic foundation established, the remaining task is to extract the acoustic evidence from $\widetilde{a}$. Because crucial acoustic information inherently relies on semantic foundation, we do not train the acoustic pathway in isolation. This dependency is exactly mirrored in our data construction: the enriched script $\widetilde{t}$ is formulated simply by augmenting the baseline transcript t with additional acoustic information. Consequently, we generate the joint semantic-acoustic target by prompting the frozen Teacher $\mathcal{T}$ with $\widetilde{t}$:

$$Y^{\text{para}} = \mathcal{T}(\widetilde{t}; I_{\text{para}}), \tag{8}$$

where I_{para} instructs the Teacher to incorporate the injected acoustic behaviors into its analytical reasoning. To align the acoustic FLQ-Adapter with this enriched target, the Student processes $\widetilde{a}$ with both audio pathways active. Under an instruction I_{aco} containing both the <SEM> and <PARA>

placeholders, the objective is defined as:

$$\mathcal{L}_{\text{para}} = -\sum_{j=1}^{|Y^{\text{para}}|} \log p_{\theta_0}(v_j \mid v_{<j}, I_{\text{para}}, Z_{\text{whisper}}, Z_{\text{wav}}) . \tag{9}$$

During this stage, we freeze the semantic adapter and update only the acoustic adapter. The acoustic pathway is thus trained to support script-informed predictions in the presence of the fixed semantic representation Z_{whisper}. This supervision is intended to encourage the ultilization of acoustic evidence alongside the semantic anchor. Target-generation prompts and optimization settings are provided in Appendix C.

With both the semantic and acoustic adapters successfully decoupled via the ParaScam triplet, they are now fully primed to analyze authentic audio. As we transition to Telecom Fraud Detection on the real-world dataset $\mathcal{D}_{\text{train}}$, we intentionally freeze both of these audio pathways, because continuing to update them jointly on a singular downstream target (a, y) would risk catastrophic forgetting, causing the carefully disentangled features to collapse back into entangled redundancy.

After role-specific alignment, we adapt RASA to the downstream dataset D_{train}. Both encoders and FLQ-Adapters remain frozen, preserving the learned audio-to-embedding mappings during task adaptation. We apply Low-Rank Adaptation (LoRA) (Hu et al., 2022) to the LLM backbone to learn the target prediction schema using only audio and task labels. For a labeled call (a, y), let Y^{task} serialize the annotated targets and I_{task} specify the prediction requirements. We optimize the LoRA parameters λ through:

$$\mathcal{L}_{\text{task}} = -\sum_{j=1}^{|Y^{\text{task}}|} \log p_{\theta_0+\Delta\theta_\lambda}(v_j \mid v_{<j}, I_{\text{task}}, Z_{\text{whisper}}, Z_{\text{wav}}) , \tag{10}$$

where $\Delta\theta_\lambda$ denotes the low-rank backbone updates. Only λ is trainable; the base LLM parameters and both audio pathways remain frozen.

5 Experiments

5.1 Experimental Setup

Datasets and metrics. To evaluate RASA, we conduct experiments on TeleAntiFraud-28k (TeleAnti), a comprehensive audio-text telecom fraud dataset (Ma et al., 2025b). The benchmark assesses three core prediction tasks: call-scene classification, binary fraud detection, and fraud-type identification. We report class-support-weighted F1 scores for all three tasks and their arithmetic mean. For the ablation study, we focus exclusively on binary fraud detection and fraud-type identification—the most critical fraud-reasoning objectives—and additionally report fraud-positive precision and recall. While the original benchmark includes an LLM-as-a-judge "Quality-of-Thought" metric, we omit this composite score to ensure a strictly objective, deterministic evaluation of the model's predictive accuracy. Detailed split statistics are deferred to Appendix D.

Baselines. Following the processing paradigms analyzed in Section 3, we categorize the baselines into three distinct groups: *(1) ASR–LLM baselines.* These systems transcribe audio before prompting an instruction-tuned text LLM (Qwen2-7B-Base, Llama-3.1-8B-Instruct, GLM-4-9B-Chat). *(2) Direct LALM baselines.* These zero-shot systems process audio directly without downstream target-label adaptation (Qwen2-Audio, StepAudio2-Think, GLM-4-Voice). *(3) Target-adapted baselines.* This group comprises best audio-language models specifically fine-tuned or optimized via reinforcement learning on the target dataset (AntiFraud-Qwen2-Audio (Ma et al., 2025b), and the SAFE-RL, SAFE-RS, and SAFE-LS variants (Wang et al., 2026d)). Results marked with † are evaluated locally; all other values are sourced from their original publications.

Implementation details. RASA employs Whisper as the frozen semantic encoder, WavTokenizer as the frozen acoustic encoder, and Qwen2-7B-Instruct as the reasoning backbone. Following the role-specific alignment on our ParaScam dataset (Section 4.2), the model undergoes task-specific adaptation on TeleAnti. During this final stage, both the semantic and acoustic FLQ-Adapters are

Table 1: Results on TeleAnti (%). All F1 scores use class-support weighting. Avg. F1 is the average of Fraud, Fraud Type, and Scene weighted F1 scores.

Paradigm	Model	Fraud F1	Fraud Type F1	Scene F1	Avg. F1
Without target-label adaptation					
ASR+LLM	Qwen2-7B†	79.88	32.99	63.04	58.64
	Llama-3.1-8B-Instruct†	95.37	53.99	60.38	69.91
	GLM-4-9B-Chat	46.91	82.22	75.10	68.08
Direct LALM	Qwen2-Audio-7B-Instruct	58.51	20.48	70.22	49.74
	StepAudio2-Think†	66.12	11.67	58.19	45.33
	GLM-4-Voice-9B	26.83	38.33	0.00	21.72
With target-label adaptation					
Adapted LALM	AntiFraud-Qwen2-Audio	84.78	82.91	81.31	83.00
SAFE-QAQ	SAFE-RL	90.20	87.25	81.57	86.34
	SAFE-RS	89.61	86.39	81.60	85.87
	SAFE-LS	89.61	**88.23**	84.64	87.49
RASA	LoRA, adapters frozen† (ours)	**98.13**	80.17	**87.52**	**88,61**
RASA-500	LoRA, adapters frozen† (ours)	96.91	81.00	76.68	84.86

† Performance evaluated by ourselves.

strictly frozen, and only the LoRA parameters within the LLM backbone are updated. A single joint instruction prompts the model to generate the scene, binary fraud decision, and fraud type within a unified JSON object. At inference, the model relies exclusively on the raw call audio and the unified task instruction. The complete prompt and further hyperparameters are detailed in Appendix E.

5.2 Experimental Results

We structure our experimental analysis to mirror the evolutionary trajectory of telecom fraud detection systems. We first evaluate transcript-based models, which isolate semantic content while discarding acoustic evidence. As shown in Table 1, these models provide a strong baseline for binary fraud logic (e.g., Llama-3.1-8B-Instruct achieves 95.37% Fraud F1). However, discarding vocal delivery severely limits fine-grained reasoning: Qwen2-7B manages a mere 32.99% on Fraud Type, and the paradigm exhibits a hard ceiling on Average F1 (69.91%), confirming that transcripts alone are insufficient.

To recover acoustic evidence, recent approaches feed waveforms directly into LALMs, but zero-shot evaluation exposes a severe dual vulnerability. Direct LALMs perform significantly worse than text baselines (e.g., Qwen2-Audio-7B-Instruct at 49.74% Avg. F1). On the semantic front, without explicit transcription, they fail to reconstruct semantic content as reliably as ASR pipelines. Simultaneously, on the acoustic front, their reliance on recognition-oriented Whisper encoders inherently marginalizes acoustic evidence. Thus, zero-shot LALMs fail to leverage either modality effectively.

Target-adapted LALMs attempt to rebuild this semantic competence either through extensive fine-tuning (e.g., AntiFraud-Qwen2-Audio at 83.00% Avg. F1) or computationally expensive reinforcement learning (e.g., SAFE-LS at 87.49% Avg. F1). While they improve significantly over their zero-shot counterparts, they still hit a strict performance ceiling. Their single-encoder architectures inherently bottleneck the extraction of crucial acoustic evidence, limiting further breakthroughs regardless of the optimization method.

Rather than relying on expensive RL, RASA shatters this bottleneck structurally by simply adding a parallel reconstruction-oriented WavTokenizer pathway. Trained on the full training dataset, this dual-stream architecture achieves effective results: an exceptional 98.13% Fraud F1, 87.52% Scene F1, and the highest Average F1 of 88.61%. Crucially, RASA demonstrates extreme sample efficiency. To simulate the rigorous label scarcity of real-world $\mathcal{D}_{\text{train}}$, we constrain RASA to a minimal subset of only 500 labeled examples. Remarkably, RASA-500 achieves an Average F1 of 84.86%, tightly approaching our full best performance and explicitly outperforming the target-

Table 2: Ablation configurations and results under the TeleAnti train-500 protocol. "Trained" indicates adapter training on ParaScam before downstream LoRA adaptation. Full RASA uses sequential alignment, whereas Joint alignment trains both adapters simultaneously. All available adapters remain frozen during downstream adaptation. F1 scores are class-support-weighted; Precision and Recall use fraud as the positive class. All values are percentages.

	Semantic		Acoustic		Performance			
Configuration	Used	Trained	Used	Trained	Fraud F1	Precision	Recall	Type F1
Full RASA	✓	✓	✓	✓	96.91	94.21	99.56	81.00
Single-VQencoder	✗	✗	✓	✓	99.69	99.43	99.91	34.33
Semantic only	✓	✓	✗	✗	96.17	98.83	93.00	53.63
Joint alignment	✓	✓	✓	✓	93.98	94.24	92.91	78.55
No supp. alignment	✓	✓	✓	✗	97.49	95.79	99.03	60.66

adapted AntiFraud-Qwen2-Audio (83.00%) that was fully fine-tuned on the entire dataset. This conclusively proves that making full use of crucial acoustic evidence is of fundamental importance for enabling LLMs to conduct in-depth analysis and reasoning on telecom fraud calls.

5.3 Ablation Studies

Design. We examine the contributions of the acoustic pathway and role-specific adapter training under the TeleAnti train-500 protocol (Appendix E). Table 2 defines the configurations and reports their corresponding performance. In every configuration, the available adapters are frozen during task specialization and only Qwen2 LoRA parameters are trained.

Dual pathway contributes to fraud-type recognition. Full RASA achieves 81.00% Type F1, compared with 53.63% for Semantic only and 34.33% for Single-VQencoder. These results support the value of combining the two aligned pathways for fine-grained classification under the train-500 protocol. The benefit differs across tasks: Single-VQencoder achieves the highest binary Fraud F1 at 99.69%, compared with 96.91% for Full RASA. Thus, the dual-pathway configuration primarily improves fraud-type recognition in this comparison.

Sequential versus joint alignment. We compare RASA's sequential alignment with a joint alternative that trains both adapters simultaneously on ParaScam. Sequential alignment improves Fraud F1 from 93.98% to 96.91% and Type F1 from 78.55% to 81.00%. It also increases fraud recall from 92.91% to 99.56%, while maintaining nearly identical precision (94.21% versus 94.24%). These results support the benefit of establishing a semantic anchor before training the acoustic adapter under the train-500 protocol.

Contribution of acoustic alignment. With the semantic adapter pretrained but the acoustic adapter left unaligned, *No supp. alignment* achieves 60.66% Type F1. Training the acoustic adapter with script-informed targets increases Type F1 to 81.00%, an improvement of 20.34 percentage points. This result supports the contribution of the acoustic alignment stage to fine-grained classification within the proposed training pipeline.

6 Conclusion

This paper introduces RASA, a dual-pathway framework for telecom fraud detection that combines Whisper-derived semantic representations with reconstruction-oriented acoustic representations. Using ParaScam, role-specific alignment first trains a semantic adapter with transcript-conditioned targets and then trains an acoustic adapter with script-informed targets while keeping the semantic pathway fixed. Downstream LoRA adaptation requires only audio and task labels. Experiments on the evaluated TeleAnti split show strong predictive performance, with ablations demonstrating gains in fraud-type recognition from acoustic alignment under limited-label adaptation. Future work will evaluate robustness across diverse acoustic conditions and further investigate the evidence captured by each pathway.

AI USE STATEMENT

In this work, we utilized generative AI models as core components of the ParaScam dataset construction and role-specific training pipeline. Specifically, a text-based large language model was employed to augment plain transcripts with explicit acoustic annotations, and to generate the transcript-conditioned and acoustically informed instructional targets. RASA itself employs a frozen large language model as its reasoning backbone. The generative models and their exact roles in the data-construction and training procedures are rigorously detailed in the main paper and appendix. AI-generated textual data were subjected to the structural and consistency checks defined in the automated pipeline and were manually reviewed by the authors. This work does not involve formal mathematical proofs, and generative AI was not used to produce proofs. The authors take full responsibility for the final content of this paper, including the data.

ETHICS STATEMENT

Telecom fraud detection is fundamentally a safety-oriented application; however, the data and methods studied herein raise important privacy, fairness, and dual-use considerations. Our **ParaScam** pipeline utilizes generative LLMs and TTS synthesis to construct realistic fraud calls, significantly reducing the necessity of collecting and exposing live, sensitive fraudulent interactions involving real victims. Nonetheless, synthetic scripts and voices may inadvertently reproduce demographic biases present in their source models, and the synthesis pipeline itself could theoretically be misused by malicious actors to imitate deceptive calls. Any public release of this dataset or framework will comply strictly with the licenses of the source models, inherently strip personally identifying information, and restrict the distribution of materials that could unnecessarily facilitate abuse. Furthermore, detection errors (false positives) may disproportionately affect specific speakers, accents, or recording conditions underrepresented in the evaluation data. Consequently, RASA should not be deployed as the sole basis for punitive or account-level automated interventions. Real-world deployment requires rigorous domain-specific validation, conservative confidence thresholds, human-in-the-loop oversight, and an accessible recourse process for affected users. No human participants were recruited, and no model was actively deployed against real users in the course of this study.

REPRODUCIBILITY STATEMENT

The main paper details the RASA architecture, the distinct supervision assigned to its semantic and acoustic pathways, and the sequential training objectives. The appendix provides exhaustive configurations, including the ParaScam construction protocols, baseline identifiers, exact optimization hyperparameters, and metric definitions. To facilitate reproducibility, our supplementary material includes the complete anonymized training/evaluation scripts, and configuration YAMLs. Due to strict attachment size limits and data dissemination policies, the raw datasets and trained checkpoints are excluded from the initial submission; however, our codebase supports exact reproduction once the specified resources are supplied locally. Access to all third-party models and datasets remains subject to their original licenses.

REFERENCES

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, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikoł aj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: a Visual Language Model for Few-Shot Learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 23716–23736. Curran Associates, Inc., 2022. doi: 10.52202/068431-1723. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/960a172bc7fbf0177ccccbb411a7d800-Abstract-Conference.html.

Jingyi Chen, Zhimeng Guo, Jiyun Chun, Pichao Wang, Andrew Perrault, and Micha Elsner. Do audio llms really listen, or just transcribe? measuring lexical vs. acoustic emotion cues reliance. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5848–5877, 2026.

Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, Wanxiang Che, Xiangzhan Yu, and Furu Wei. BEATs: Audio Pre-Training with Acoustic Tokenizers. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 5178–5193. PMLR, 2023. URL https://proceedings.mlr.press/v202/chen23ag.html.

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

Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High Fidelity Neural Audio Compression. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=ivCd8z8zR2.

Zhihao Du, Changfeng Gao, Yuxuan Wang, Fan Yu, Tianyu Zhao, Hao Wang, Xiang Lv, Hui Wang, Xian Shi, Keyu An, et al. Cosyvoice 3: Towards in-the-wild speech generation via scaling-up and post-training. *arXiv preprint arXiv:2505.17589*, 2025.

Jiajun Fan, Roger Ren, Jingyuan Li, Rahul Pandey, Prashanth Gurunath Shivakumar, Ivan Bulyko, Ankur Gandhe, Ge Liu, and Yile Gu. Incentivizing consistent, effective and scalable reasoning capability in audio llms via reasoning process rewards. In *International Conference on Learning Representations*, volume 2026, pp. 2110–2173, 2026.

Yuan Gong, Hongyin Luo, Alexander Liu, Leonid Karlinsky, and James R Glass. Listen, think, and understand. In *International Conference on Learning Representations*, volume 2024, pp. 18516–18545, 2024.

Namgyu Ho, Laura Schmid, and Se-Young Yun. Large Language Models Are Reasoning Teachers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14852–14882. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.830. URL https://aclanthology.org/2023.acl-long.830/.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. Distilling Step-by-Step! Outperforming Larger Language Models with Less Training Data and Smaller Model Sizes. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 8003–8017. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-acl.507. URL https://aclanthology.org/2023.findings-acl.507/.

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*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Shujie Hu, Long Zhou, Shujie Liu, Sanyuan Chen, Lingwei Meng, Hongkun Hao, Jing Pan, Xunying Liu, Jinyu Li, Sunit Sivasankaran, et al. Wavllm: Towards robust and adaptive speech large language model. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 4552–4572, 2024.

Chien-yu Huang, Wei-Chih Chen, Shu-wen Yang, Andy T Liu, Chen-An Li, Yu-Xiang Lin, Wei-Cheng Tseng, Anuj Diwan, Yi-Jen Shih, Jiatong Shi, et al. Dynamic-superb phase-2: A collaboratively expanding benchmark for measuring the capabilities of spoken language models with 180 tasks. In *International Conference on Learning Representations*, volume 2025, pp. 82908–82973, 2025.

Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, et al. Perceiver io: A general architecture for structured inputs & outputs. *arXiv preprint arXiv:2107.14795*, 2021.

Shengpeng Ji, Ziyue Jiang, Wen Wang, Yifu Chen, Minghui Fang, Jialong Zuo, Qian Yang, Xize Cheng, Zehan Wang, Ruiqi Li, et al. Wavtokenizer: an efficient acoustic discrete codec tokenizer for audio language modeling. *arXiv preprint arXiv:2408.16532*, 2024.

Shengpeng Ji, Minghui Fang, Jialong Zuo, Ziyue Jiang, Dingdong Wang, Hanting Wang, Hai Huang, and Zhou Zhao. Language-Codec: Bridging Discrete Codec Representations and Speech Language Models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13332–13345. Association for Computational Linguistics, 2025a. doi: 10.18653/v1/2025.acl-long.654. URL https://aclanthology.org/2025.acl-long.654/.

Shengpeng Ji, Ziyue Jiang, Wen Wang, Yifu Chen, Minghui Fang, Jialong Zuo, Qian Yang, Xize Cheng, zehan wang, Ruiqi Li, Ziang Zhang, Xiaoda Yang, Rongjie Huang, YIDI JIANG, Qian Chen, Siqi Zheng, and Zhou Zhao. WavTokenizer: an Efficient Acoustic Discrete Codec Tokenizer for Audio Language Modeling. In *International Conference on Learning Representations*, 2025b. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/ea1f5f0878d43ff4fb8bf64ef4a2326c-Abstract-Conference.html.

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 *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 25125–25148. PMLR, 2024. URL https://proceedings.mlr.press/v235/kong24a.html.

Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar. High-Fidelity Audio Compression with Improved RVQGAN. In *Advances in Neural Information Processing Systems*, volume 36, pp. 27980–27993. Curran Associates, Inc., 2023. doi: 10.52202/075280-1214. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/58d0e78cf042af5876e12661087bea12-Abstract-Conference.html.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 19730–19742. PMLR, 2023. URL https://proceedings.mlr.press/v202/li23q.html.

Haoyu Ma, Qinliang Su, Minhua Huang, and Wu Kai. Detecting continuously evolving scam calls under limited annotation: A llm-augmented expert rule framework. In *EMNLP (Findings)*, pp. 5047–5068, 2025a.

Zhiming Ma, Peidong Wang, Minhua Huang, Jinpeng Wang, Kai Wu, Xiangzhao Lv, Yachun Pang, Yin Yang, Wenjie Tang, and Yuchen Kang. Teleantifraud-28k: An audio-text slow-thinking dataset for telecom fraud detection. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 5853–5862, 2025b.

Ziyang Ma, Zhisheng Zheng, Jiaxin Ye, Jinchao Li, Zhifu Gao, Shiliang Zhang, and Xie Chen. emotion2vec: Self-supervised pre-training for speech emotion representation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 15747–15760, 2024.

Kaizhi Qian, Xulin Fan, Junrui Ni, Slava Shechtman, Mark Hasegawa-Johnson, Chuang Gan, and Yang Zhang. Prosodylm: Uncovering the emerging prosody processing capabilities in speech language models. *arXiv preprint arXiv:2507.20091*, 2025.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. Robust Speech Recognition via Large-Scale Weak Supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 28492–28518. PMLR, 2023. URL https://proceedings.mlr.press/v202/radford23a.html.

Sakshi Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. Mmau: A massive multi-task audio understanding and reasoning benchmark. In *International Conference on Learning Representations*, volume 2025, pp. 84929–84964, 2025.

Zitong Shen, Kangzhong Wang, Youqian Zhang, Grace Ngai, and Eugene Yujun Fu. Combating phone scams with llm-based detection: Where do we stand?(student abstract). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 29487–29489, 2025.

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*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/476ab8f369e489c04187ba84f68cfa68-Abstract-Conference.html.

Vishaal Udandarao, Zhiyun Lu, Xuankai Chang, Yongqiang Wang, Albin Madappally Jose, Fartash Faghri, Josh Gardner, and Chung-Cheng Chiu. Data-Centric Lessons To Improve Speech-Language Pretraining. In *International Conference on Learning Representations*, volume 2026, pp. 61136–61190, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/63b96ace3e28465aff61918e77de2a00-Abstract-Conference.html.

Chaoren Wang, Heng Lu, Xueyao Zhang, Shujie Liu, Yan Lu, Jinyu Li, and Zhizheng Wu. Closing the Modality Reasoning Gap for Speech Large Language Models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 18821–18835. Association for Computational Linguistics, 2026a. doi: 10.18653/v1/2026.acl-long.857. URL https://aclanthology.org/2026.acl-long.857/.

Chen Wang, Minpeng Liao, Zhongqiang Huang, Jinliang Lu, Junhong Wu, Yuchen Liu, Chengqing Zong, and Jiajun Zhang. Blsp: Bootstrapping language-speech pre-training via behavior alignment of continuation writing. *arXiv preprint arXiv:2309.00916*, 2023a.

Chen Wang, Minpeng Liao, Zhongqiang Huang, Junhong Wu, Chengqing Zong, and Jiajun Zhang. Blsp-emo: Towards empathetic large speech-language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 19186–19199, 2024.

Dingdong Wang, Junan Li, Jincenzi Wu, Dongchao Yang, Xueyuan Chen, Tianhua Zhang, and Helen Meng. Mmsu: A massive multi-task spoken language understanding and reasoning benchmark. In *International Conference on Learning Representations*, volume 2026, pp. 31374–31410, 2026b.

Dingdong Wang, Shujie Liu, Tianhua Zhang, Youjun Chen, Jinyu Li, and Helen Meng. Emotion-thinker: Prosody-aware reinforcement learning for explainable speech emotion reasoning. *arXiv preprint arXiv:2601.15668*, 2026c.

Peidong Wang, Zhiming Ma, Xin Dai, Yongkang Liu, Shi Feng, Xiaocui Yang, Wenxing Hu, Zhihao Wang, Mingjun Pan, Li Yuan, et al. Safe-qaq: End-to-end slow-thinking audio-text fraud detection via reinforcement learning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 26138–26157, 2026d.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-Instruct: Aligning Language Models with Self-Generated Instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13484–13508. Association for Computational Linguistics, 2023b. doi: 10.18653/v1/2023.acl-long.754. URL https://aclanthology.org/2023.acl-long.754/.

Tong Wu, Qinliang Su, Jianxing Yu, Bo Liang, and Minhua Huang. Targeting the needle, ignoring the haystack: Anchoring crucial cues for evolving scam call detection via an llm-assisted classifier. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 20497–20513, 2026.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. Qwen2 technical report, 2024. URL https://arxiv.org/abs/2407.10671.

Dongchao Yang, Songxiang Liu, Haohan Guo, Jiankun Zhao, Yuanyuan Wang, Helin Wang, Zeqian Ju, Xubo Liu, Xueyuan Chen, Xu Tan, Xixin Wu, and Helen M. Meng. ALMTokenizer: A Low-bitrate and Semantic-rich Audio Codec Tokenizer for Audio Language Modeling. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 70850–70872. PMLR, 2025. URL https://proceedings.mlr.press/v267/yang25q.html.

Xin Zhang, Dong Zhang, Shimin Li, Yaqian Zhou, and Xipeng Qiu. Speechtokenizer: Unified speech tokenizer for speech language models. In *International Conference on Learning Representations*, volume 2024, pp. 31798–31818, 2024.

A FRAMEWORK IMPLEMENTATION

This section specifies the implementation of the architecture in Section 4.1. The semantic encoder is the Whisper-derived audio tower released with Qwen2-Audio (Radford et al., 2023; Chu et al., 2024); the acoustic encoder supplies WavTokenizer's continuous quantized codebook features (Ji et al., 2024). The Qwen2 language backbone has an embedding width of $d_{\mathrm{LLM}} = 3584$ (Yang et al., 2024).

Table 3 lists the two adapters' settings. Their parameters are independent. Each adapter uses cross-attention followed by layer normalization and a `Linear--GELU--Linear` projector with biases; a learned pathway-type embedding is added after projection. The assembled text and soft-token embeddings are supplied through the language model's `inputs_embeds` interface.

Table 3: Dimensions of the two soft-token adapters.

Setting	Semantic	Supplementary
Encoder feature width (d_r)	1280	512
Learned queries	64	64
Cross-attention heads	8	8
Resampler output	64×1280	64×512
Projector	$1280 \to 4096 \to 3584$	$512 \to 4096 \to 3584$
Type embedding	learned	learned
Final soft tokens	64×3584	64×3584
Dropout	0.0	0.0

B PARASCAM DATASET CONSTRUCTION

ParaScam is explicitly designed to provide the exact $(\widetilde{a}, t, \widetilde{t})$ triplets required by RASA's role-specific learning (Section 4.2). The construction strictly follows a three-step pipeline: semantic isolation, acoustic enrichment, and privacy-safe speech synthesis. Only the fixed training partition is used to optimize the two FLQ-Adapters.

ASR transcription and de-identification. We organize each Chinese two-party call as an ordered sequence of caller and callee turns. The source audio is transcribed using an ASR pipeline, and the transcripts are de-identified to reduce the exposure of personally identifiable information. We discard a record if its role sequence cannot be parsed, any turn is empty, or its paired fields cannot be aligned. The resulting plain transcript t provides the textual view used to generate supervision for the semantic adapter.

Acoustic script enrichment. Model-generated supervision offers a way to expand training resources beyond manually written annotations (Wang et al., 2023b), while recent speech-language pretraining studies highlight the importance of data curation and synthetic-data construction (Udandarao et al., 2026). We apply these ideas to construct paired transcript and delivery-enriched script views for role-specific alignment. To provide explicit descriptions of intended vocal delivery, we use Gemini-3.1-Flash-Lite to convert each turn into an enriched utterance paired with an acoustic description. The description specifies context-dependent vocal delivery, such as emotion, speaking rate, hesitation, or urgent interruptions. The utterance receives punctuation, short fillers, and explicit non-verbal markers (e.g., `[breath]`, `[laughter]`). To ensure the semantic foundation remains unchanged, the prompt strictly prohibits alterations to core entities, requested actions or sensitive-information requests. This yields the enriched script $\widetilde{t}$.

Privacy-safe speech synthesis. Textual descriptions alone cannot train an acoustic adapter unless the intended behaviors are physically realized in the waveform. Furthermore, to completely eliminate original speaker identities, we replace the source audio by synthesizing the accepted scripts using Fun-CosyVoice3-0.5B (Du et al., 2025). Each turn is explicitly conditioned on its generated acoustic description. A deterministic selection procedure assigns one reference voice to the caller and another to the callee, keeping each voice fixed within the call. The synthesized turns are concatenated in their original order with role-dependent pauses to form the final privacy-safe waveform

$\tilde{a}$. Consequently, this procedure pairs the synthesized waveform $\tilde{a}$ with the enriched script $\tilde{t}$ that specifies its intended delivery, while the plain transcript t provides the underlying conversational content.

Additional annotations and validation. Role-specific alignment uses teacher-generated targets derived from the paired transcripts and enriched scripts. The alignment objectives do not use the manually annotated task labels described below. However, to facilitate future in-depth research on telecom fraud detection (e.g., evaluating LALMs on synthetic acoustic behaviors or exploring new acoustic reasoning tasks), we subsequently conducted manual annotation on the generated ParaScam dataset. We annotated each call with a binary fraud label, a communication scene, and a specific fraud type for fraudulent calls. We retain a call only when its synthetic waveform, plain transcript, enriched script, role sequence, and annotations are fully consistent.

C Role-Specific Training

Instruction-target generation. Both the transcript-conditioned and acoustically informed targets are generated with Qwen2-7B-Instruct. Target generation uses a maximum input length of 8,192 tokens, a maximum of 512 generated tokens, batch size 32, bfloat16, and greedy decoding (`do_sample=False`). The system instruction is "You are an anti-telecom-fraud analysis expert. Output valid JSON only." The exact implementation uses Chinese prompts; Table 4 is a faithful English rendering of their information constraints and output schemas. We parse the first JSON object after removing optional code fences, and retain an example only if every required field is present and the binary/type values agree with the fixed label vocabulary. This yields 15,971 valid transcript-conditioned targets (99.45%) and 16,025 valid acoustically informed targets (99.78%) out of 16,060 calls.

Table 4: Instruction-target prompt contracts. The transcript view contains cleaned spoken content; the enriched-script view additionally contains acoustic annotations.

Target view	Instruction (English rendering)	Required JSON fields

Transcript-conditioned	Analyze the following call transcript using only facts verifiable from the text. Do not infer tone or emotion. Determine whether it is fraud, identify one fixed fraud type when applicable, and give concise supporting evidence. Output JSON only.	`semantic_summary`
		`reason`
		`fraud_result`
		`fraud_type`

Acoustically informed	Analyze the following dialogue script. Use both the utterance content and the provided acoustic cues. Determine whether it is fraud, identify one fixed fraud type when applicable, and explain the decision using both semantic and acoustic evidence. Output JSON only.	`semantic_summary`
		`acoustic_summary`
		`reason`
		`fraud_result`
		`fraud_type`

Role-specific prompt and supervision. The LLM receives a fixed system instruction followed by a user message containing two placeholders: `<SEM>` for the semantic soft-token sequence and `<PARA>` for the acoustic soft-token sequence. During semantic initialization, only `<SEM>` is inserted and the requested fields are `semantic_summary`, `fraud_result`, `fraud_type`, and `reason`. During acoustic role learning, both placeholders are inserted and the five fields in the acoustically informed schema are requested. The markers occur exactly once in the rendered chat template and are replaced by the learned soft tokens. Loss is masked on the system message, user message, and soft-token prefix; cross-entropy is applied only to the appended target JSON.

Role-specific objectives and optimization. Semantic initialization trains only the semantic adapter against valid transcript-conditioned targets. Acoustic role learning starts from the selected semantic adapter, keeps that pathway fixed, and trains only the WavTokenizer adapter against targets

combining the transcript-conditioned `semantic_summary` with the acoustically informed summary, decision, type, and rationale. Table 5 summarizes these two representation-learning stages.

Table 5: Optimization settings for role-specific representation learning on ParaScam.

Setting	Semantic init.	Acoustic role
Training target	Transcript target	Enriched target
Training calls	9,583	9,573
Trainable parameters	Semantic adapter	Acoustic adapter
Epochs	4	4
Micro-batch / accumulation	1 / 8	1 / 8
Learning rate	$5e{-5}$	$5e{-5}$
Weight decay	0.01	0.01
Gradient clipping	1.0	1.0
Random seed	42	42

D TeleAnti Split Statistics

TeleAnti evaluation population. Our experiments use a fixed call-level subset of 16,060 TeleAnti records with complete audio and task metadata. We construct mutually disjoint 60/10/30 training, validation, and test partitions with seed 43, stratified by binary label and fraud type. Low-resource task adaptation uses a fixed label-stratified subset of 500 training calls and a fixed 256-call validation subset, while final evaluation uses the complete test partition. Table 6 summarizes these populations. Fraud-type evaluation is restricted to the 2,272 reference-labeled fraud calls in the test partition and uses the eight metadata-provided categories.

Table 6: Data partitions and low-resource subsets used in the TeleAnti experiments.

Population	Normal	Fraud	Total
Full training partition	5,091	4,545	9,636
Selected training subset	264	236	500
Full validation partition	849	757	1,606
Selected validation subset	135	121	256
Test partition	2,546	2,272	4,818

E Task-Specific Training and Inference

Task-specific adaptation. For downstream specialization, the two trained adapters are frozen and only LoRA parameters in the Qwen2 language backbone are optimized. Each of the 500 selected training calls is paired with one joint JSON target containing `scene`, `fraud_result`, and `fraud_type`; for a normal call, `fraud_type` is set to NULL. We use the same joint prompt during training and inference. The model is trained for four epochs with micro-batch size 1, gradient accumulation 8, learning rate $2e{-4}$, weight decay 0.01, gradient clipping at 1.0, and seed 43. As in role-specific training, loss is applied only to the appended target JSON.

Checkpoint selection. We select the joint RASA checkpoint using the mean of the three validation weighted F1 scores. Scene and binary F1 are computed on all 256 validation calls, while fraud-type F1 is computed on the 121 fraudulent calls. The ablations retain the task-specific selection criterion corresponding to the output being evaluated.

Inference configuration. After checkpoint selection, the model is evaluated once on the complete test partition. The semantic and acoustic soft-token groups are inserted into the marked positions in Figure 3, and all three predictions are generated in a single response. Generation is limited to 64 new tokens. We parse the three requested fields from the first valid JSON object. Fraud-type F1 is computed only over reference-labeled fraud calls, while scene and binary F1 use the complete

scam call infer prompt

System

You are an anti-telecom-fraud analysis expert.

Output valid JSON only.

User

[description of fraud label]

The following stream provides semantic information from the call:
<SEM> [semantic soft tokens] </SEM>

The following stream provides acoustic evidence information:
<PARA> [acoustic soft tokens] </PARA>

Return one JSON object with exactly the following fields and order.

If the call is normal, set fraud_type to normal. Do not add text outside the JSON.

```
{
  "semantic_summary": "<summary of call content>",
  "acoustic_evidence": "<relevant paralinguistic cues>",
  "reason": "<brief justification based on both streams>",
  "fraud_result": "<fraud or normal>",
  "fraud_type": "<type or normal>",
  "scene": "<scene>"
}
```

Figure 3: prompt used for inference. The figure is a faithful English rendering of the Chinese implementation.

test partition. The same joint output contract and metric implementation are used for RASA and all locally evaluated main-table baselines.