AUDIO-ZERO: LABEL-FREE SELF-EVOLUTION FOR FINE-GRAINED AUDIO REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Large Audio Language Models (LALMs) have made rapid progress on acoustic understanding, yet they still struggle with fine-grained audio reasoning (e.g., recognizing event order, repetitions and duration). Existing post-training methods heavily rely on expensive external labels or provide only coarse semantic signals. To bridge this gap, we introduce Audio-Zero, the first label-free self-evolution framework in the field of LALMs that improves fine-grained auditory perception and reasoning. Audio-Zero constructs an auditory self-play game from unlabeled audio contrast pairs: most players hear a reference audio, while one odd listener hears a subtle variant. The model first generates clues describing what it hears and then identifies the odd listener by reasoning over inconsistencies among clues. Since the odd listener is known by construction, the game provides verifiable rewards without any annotated answers. Experiments with Qwen2-Audio-7B-Instruct and Qwen2.5-Omni-7B on TREA, MMAU Test-mini and MMAR show that Audio-Zero improves fine-grained audio reasoning while preserving broad audio understanding. Evolutionary and diagnostic analyses further reveal that increasingly fine-grained auditory descriptions emerge naturally from game pressure.

1 INTRODUCTION

Large Audio Language Models (LALMs) have demonstrated strong capability in understanding diverse audio inputs, positioning them as a promising general-purpose paradigm for audio understanding and reasoning (Chu et al., 2024; Xu et al., 2025c). However, high-quality supervision for LALMs is scarce and hard to scale (Deshmukh et al., 2023; Tang et al., 2024), as audio's continuous, temporally extended nature disperses key acoustic cues across long sequences (Mei et al., 2024), making fine-grained annotation inherently challenging. As a result, popular audio-caption and audio-QA datasets (Kumar et al., 2025; Ma et al., 2025b) rely on loosely-matched, web-scraped descriptions or coarse synthetic text, which capture only general semantics while overlooking detailed temporal-acoustic structure like transient sound events, precise event durations or subtle acoustic pitch changes. This bottleneck motivates a central question: *can LALMs self-improve their fine-grained auditory perception and reasoning directly from unlabeled audio, without relying on ground-truth labels, pre-defined QA pairs, or explicit reasoning traces?*

Self-evolving paradigms, which improve LLM capability without dense human supervision, have been extensively studied in the text and vision domains. Generally, these paradigms establish autonomous feedback loops by exploiting intrinsic environment constraints or task-inherent rules, effectively transforming self-generated outputs into verifiable optimization signals. For instance, text-based methods select self-generated reasoning paths based on symbolic correctness (Zelikman et al., 2022) or bypass human annotations through zero-sum game dynamics (Chen et al., 2024; Liu et al., 2025a; Huang et al., 2025; Zhao et al., 2026). Similarly, Large Vision-Language Models leverage spatially structured, static visual evidence and cross-view consistency constraints to construct verifiable visual rewards without relying on external manual guidance (He et al., 2025; Wang et al., 2025a).

However, these paradigms fail to directly transfer to the auditory domain. While text relies on discrete symbolic matching and vision allows global, static snapshot scanning, audio is inherently a continuous temporal stream. Crucial acoustic evidence, such as a transient sound event or a subtle ordering difference, unfolds dynamically over time and blends seamlessly into the complex background acoustic scene. Consequently, continuous acoustic cues cannot be verified via exact

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.

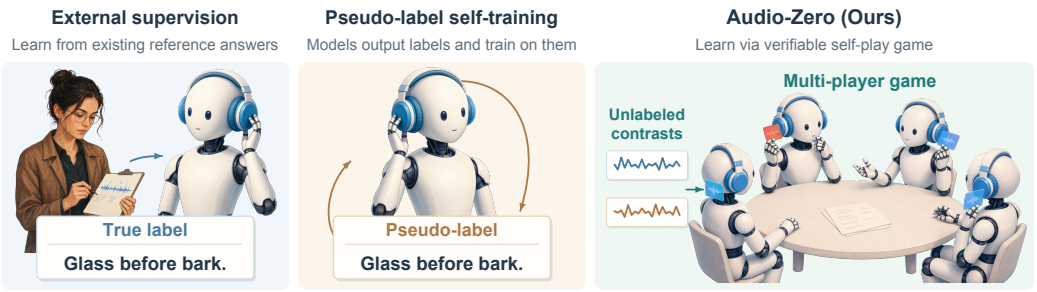


Figure 1: From external supervision to self-evolution. Existing audio-language learning paradigms depend on either external supervision or self-generated supervision. Audio-Zero instead converts unlabeled audio contrast pairs into a verifiable auditory self-play game, where supervision is derived from game outcomes, enabling self-evolution without external labels or reasoning traces.

matching or static inspection, meaning that an effective label-free audio self-improvement framework fundamentally demands an audio-specific verification mechanism capable of capturing fine-grained temporal dynamics and evaluating acoustic differences without external supervision.

Remaining within the audio domain, existing unsupervised or weakly supervised routes attempt to mitigate this annotation bottleneck, yet they face two fundamental challenges. First, self-supervised representation learning and self-improvement methods driven by pseudo-labels (Wang et al., 2025c; Zhang et al., 2025) inherently fall short in fine-grained perception. Because coarse event tags or global acoustic summaries are sufficient to optimize standard training objectives, models rarely possess the incentive to describe details. Second, this lack of precise perception impairs multi-step logical deduction, creating a structural perception-reasoning gap (Yang et al., 2025a; Shih et al., 2025). Crucially, without a verifiable grounding mechanism, training on these ungrounded descriptions triggers cascading hallucinations and error accumulation, where the policy recursively reinforces its own perceptual mistakes. While reinforcement learning and chain-of-thought distillation can alleviate this gap (Fan et al., 2025; Wang et al., 2026; Yang et al., 2026), they remain strictly bound to labeled QA pairs, teacher rationales, or manually crafted, task-specific rewards. These combined limitations heavily motivate the need for a unified, label-free framework that simultaneously incentivizes fine-grained auditory perception and multi-step reasoning through intrinsic verification.

To bridge this gap, we introduce **Audio-Zero**, the first label-free self-evolution framework tailored for LALMs. Audio-Zero converts unlabeled paired audios into an auditory self-play game. In each game, most players hear a reference audio, while one odd listener hears a subtle variant. During the listening stage, each player describes its audio in a concise clue. During the attribution stage, the model identifies the odd listener by comparing clues. The odd listener identity is generated automatically by the game construction, providing verifiable rewards without human annotation. Audio-Zero optimizes the model with Group Relative Policy Optimization (GRPO) (Shao et al., 2024), alternating between clue generation and odd-listener identification, so that better clues improve decisions and stronger decision pressure encourages more discriminative clues.

We evaluate Audio-Zero on TREA (Bhattacharya et al., 2025), MMAU-Test-mini (Sakshi et al., 2024) and MMAR (Ma et al., 2025b), covering temporal, general and deep audio reasoning. On Qwen2-Audio-7B-Instruct and Qwen2.5-Omni-7B, Audio-Zero improves temporal reasoning while also improving sound, music and speech categories, indicating that it does not trade off broad auditory competence for task-specific reasoning gains. Our contributions are summarized as follows:

- We propose Audio-Zero, a label-free self-evolution framework for improving fine-grained auditory perception and reasoning in LALMs without human captions, QA labels, or reasoning traces.
- We design an auditory self-play game that converts unlabeled audio contrast pairs into verifiable training signals by coupling clue generation with odd-listener identification.
- We comprehensive evaluate on TREA, MMAU Test-mini and MMAR using only 2k unlabeled pairs across Qwen2-Audio-7B-Instruct and Qwen2.5-Omni-7B, achieving substantial gains.

2 The Audio-Zero Framework

Audio-Zero is a self-evolution framework designed to improve fine-grained auditory perception and reasoning in LALMs without relying on external ground-truth labels. The core idea is to convert

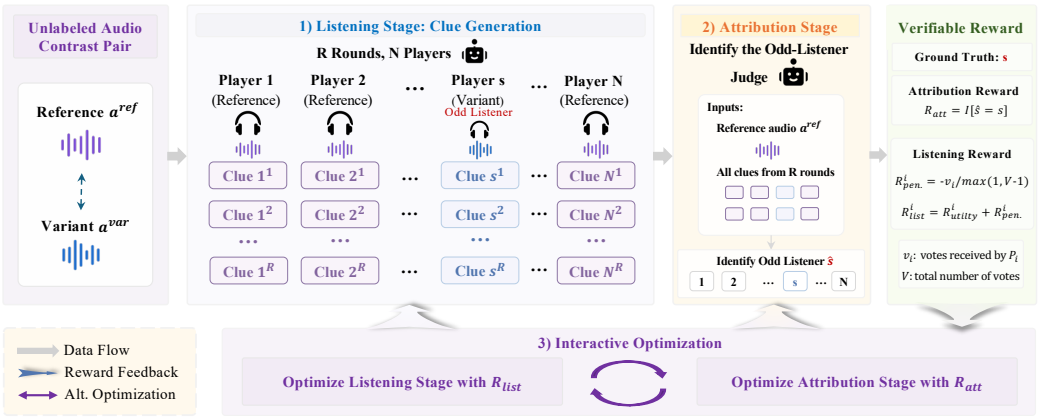


Figure 2: Overview of Audio-Zero. Audio-Zero formulates label-free auditory learning as a multi-round self-play game. Given an unlabeled audio contrast pair, players generate clues from different listening perspectives, while a judge identifies the odd listener. The known game outcome provides verifiable rewards that alternately optimize listening and attribution policies via GRPO, resulting in progressively improved auditory perception and reasoning.

unlabeled audio contrast pairs into a multi-player identification game. Specifically, crucial fine-grained acoustic discrepancies are inherently embedded within each audio pair, prompting the model to first verbalize what it hears and subsequently reason over these descriptions to pinpoint the odd listener. Because the odd listener's identity is procedurally determined, it provides an intrinsic ground truth that yields verifiable rewards within the game mechanics. Rather than relying on static labels, this design establishes a tight feedback loop where clue generation and logical attribution mutually reinforce each other, driving an autonomous evolutionary process as illustrated in Figure 2.

2.1 Problem Definition

Let π_θ denote a LALM parameterized by θ. We consider an unlabeled training collection of audio contrast pairs:

$$\mathcal{D} = \left\{ \left(a_k^{\text{ref}}, a_k^{\text{var}} \right) \right\}_{k=1}^{K}, \tag{1}$$

where recordings within each pair share semantic contexts but diverge in fine-grained acoustic attributes, such as event order, repetition count, duration, missing sounds, or ambient noise. This contrastive structure derives learning signals strictly from the relative discrepancy between paired recordings, completely circumventing the need for absolute semantic ground truth, transcripts, or reasoning traces.

We cast label-free optimization over $\mathcal{D}$ into an asymmetric N-player auditory game. For each instance $(a^{\text{ref}}, a^{\text{var}})$, the environment uniformly samples a hidden odd-listener index $s \sim \text{Uniform}(\{1, \ldots, N\})$ and assigns the acoustic inputs according to:

$$a_i = \begin{cases} a^{\text{var}}, & i = s, \\ a^{\text{ref}}, & i \neq s. \end{cases} \tag{2}$$

The identity s remains hidden from the model throughout interaction and serves as an intrinsic reference retained solely by the environment for verifiable reward computation. Sampling s uniformly further prevents the policy from exploiting positional or role-specific shortcuts.

Through this setup, the objective naturally couples descriptive clue generation across player roles with odd-listener attribution under an adjudicator perspective, establishing an externally verifiable closed loop without human supervision.

2.2 Auditory Self-Play Game

Under this formulation, each game instance unfolds in two coupled stages that share the underlying policy parameters θ while differing in their information access and action spaces.

Listening stage. For role $i \in \{1, \ldots, N\}$ and round $r \in \{1, \ldots, R\}$, a natural-language clue is sampled according to:

$$z_i^r \sim \pi_\theta(\cdot \mid a_i, H_i^r)\,. \tag{3}$$

Each clue describes concrete auditory observations—such as sound-event identities, temporal order, repetition counts, durations, acoustic attributes, or scene context—while strictly refraining from disclosing whether the speaker receives a^{ref} or a^{var}. Conditioning on the shared public history H_i^r requires the model to review preceding clues before writing its own, naturally promoting descriptions that highlight discriminative acoustic details and remain mutually informative across rounds.

Once all $N \times R$ clues are generated, the complete clue set is assembled as $Z = \{z_i^r \mid i \in [N],\, r \in [R]\}$.

Attribution stage. Following the listening stage, the policy assumes an adjudicator perspective to identify the odd listener:

$$\hat{s} \sim \pi_\theta\big(\cdot \mid a^{\text{ref}}, Z\big)\,. \tag{4}$$

Here, the model contrasts clues across roles to formulate a consistent explanation for their discrepancies, distinguishing genuine acoustic differences (e.g., conflicting event orders) from superficial phrasing ambiguities. This stage effectively maps open-ended auditory language to a discrete decision verifiable by the environment.

Separating the two stages bridges auditory perception and reasoning. Rather than scoring descriptions in isolation, the framework evaluates clue quality by whether the adjudicator can successfully resolve the discrepancy. In turn, reliable attribution relies entirely on fine-grained clues, establishing a closed loop where perceptual precision and comparative inference jointly improve.

2.3 Label-Free Rewards

Both stages derive their learning signals directly from the game, without requiring external transcripts or annotations.

Attribution reward. The policy receives a binary reward when it correctly identifies the hidden odd listener:

$$R_{\text{att}} = \mathbb{I}[\hat{s} = s]. \tag{5}$$

Because the ground-truth index s is held by the environment and never provided in the model input, this metric provides an objective, verifiable check on the model's odd-player identification.

Listening reward. The listening reward evaluates whether generated clues are informative, acoustically grounded, and strategically sound within the game. We combine a rule-based content score with a vote-based penalty.

The content component, R^i_{utility}, uses deterministic checks to verify the presence of descriptive acoustic terms (such as event sequence, duration, and background context) while penalizing verbatim repetition across rounds and explicit self-identification statements.

To prevent deceptive or uninformative clues, we introduce a vote-based penalty based on the judge's decisions. Let v_i be the number of predictions in the current GRPO group identifying role i as the odd listener, out of V total predictions. The listening reward for role i is defined as:

$$R^i_{\text{list}} = R^i_{\text{utility}} - \lambda \cdot \frac{v_i}{\max(1, V-1)}, \tag{6}$$

where $\lambda > 0$ controls the penalty weight. This penalty balances both roles: the odd listener is penalized for generating clues that make its identity obvious, while regular listeners are penalized if vague descriptions cause them to be misidentified.

During training, each update applies only the reward corresponding to the active stage:

$$R = \begin{cases} R^i_{\text{list}}, & \text{listening update for role } i, \\ R_{\text{att}}, & \text{attribution update.} \end{cases} \tag{7}$$

This separation prevents gradients from the inactive stage from interfering with the current update.

2.4 ALTERNATING GRPO OPTIMIZATION

Audio-Zero optimize the model using Group Relative Policy Optimization (GRPO) (Shao et al., 2024). Given an input prompt x, the current policy samples a group of G candidate outputs $\{y_j\}_{j=1}^{G}$. Each output receives a reward R_j, and the within-group advantage is calculated as:

$$A_j = \frac{R_j - \text{mean}_\ell(R_\ell)}{\text{std}_\ell(R_\ell) + \varepsilon}. \tag{8}$$

The policy parameters θ are updated using the clipped surrogate objective:

$$\mathcal{L}(\theta) = \mathbb{E}\left[\frac{1}{G}\sum_{j=1}^{G} \min\left(r_j(\theta)A_j,\ \text{clip}(r_j(\theta), 1-\epsilon, 1+\epsilon)A_j\right)\right] - \beta D_{\text{KL}}(\pi_\theta \parallel \pi_{\text{ref}}), \tag{9}$$

where $r_j(\theta) = \frac{\pi_\theta(y_j|x)}{\pi_{\theta_{\text{old}}}(y_j|x)}$ is the importance sampling ratio, and π_{ref} denotes the frozen initial policy.

The prompt and completion take different forms in the two stages.

- **Listening update.** $x = (a_i, H_i^r)$, $y_j = z_{i,j}^r$, and $R_j = R_{\text{list}}^i(y_j)$.
- **Attribution update.** $x = (a^{\text{ref}}, Z)$, $y_j = \hat{s}_j$, and $R_j = R_{\text{att}}(y_j)$.

The two updates are alternated throughout training. During an attribution update, the generated clues are treated as fixed context, and the policy is optimized to identify the odd role. During a listening update, the attribution decisions are held fixed, and their vote distribution supplies feedback for each clue. This schedule creates a directional learning loop. Improved attribution raises the standard for clue quality, while improved clues make the attribution context more informative. Alternation also prevents the binary attribution reward from dominating the denser clue-quality signal and provides a well-defined credit-assignment path for each action space.

3 EXPERIMENTS AND ANALYSIS

3.1 EXPERIMENTAL SETUP

Training data. Audio-Zero is instantiated with the paired audio preference dataset Audio-alpaca (Majumder et al., 2024). Each item contains a chosen recording and a rejected recording that share prompt-level semantics but may differ in event order, missing events, repetitions, or acoustic quality. After sampling and filtering, 2,000 pairs are used for post-training. The dataset provides a practical source of unlabeled auditory contrast, while its use represents an experimental instantiation rather than a methodological requirement.

Models. Two base models are considered. Qwen2-Audio-7B-Instruct (Chu et al., 2024) represents an audio-language model, whereas Qwen2.5-Omni-7B (Xu et al., 2025b) represents a unified omni-modal model.

Evaluation Benchmarks. Performance is measured by Multiple-Choice Question (MCQ) accuracy on three complementary benchmarks. MMAU Test-mini (Sakshi et al., 2024) contains 1,000 questions for general audio understanding and reasoning across Sound, Music, and Speech. MMAR (Ma et al., 2025b) evaluates deeper reasoning over speech, audio, music, and mixed-modality audio. TREA (Bhattacharya et al., 2025) focuses on temporal reasoning through the Order, Count, and Duration subsets, with 200 questions in each subset. The first two benchmarks assess broader audio reasoning ability, whereas TREA provides a more direct evaluation of the fine-grained temporal distinctions targeted during training.

Implementation details. Unless otherwise specified, each game contains $N = 4$ players and $R = 2$ listening rounds. GRPO uses $G = 6$ completions per prompt, a learning rate of 1×10^{-5}, bf16 precision, and a KL regularization coefficient of $\beta = 0.04$. The remaining hyperparameters are reported in Appendix A.1.

3.2 MAIN RESULTS ON GENERAL AND FINE-GRAINED AUDIO REASONING

To comprehensively evaluate Audio-Zero, we compare it against representative baselines spanning different learning paradigms. For the label-dependent group, we evaluate *R1-AQA* (Li et al., 2025)

Method	MMAU Test-mini				MMAR				TREA			
	Sound	Music	Speech	Avg.	Sound	Music	Speech	Avg.	Order	Count	Dur.	Avg.
Qwen2-Audio-7B-Instruct												
Base Policy	60.96	58.68	52.85	57.50	46.67	31.55	40.82	41.20	59.00	27.50	35.50	40.67
Label-Dependent												
R1-AQA (Li et al., 2025)	63.06	59.28	54.65	59.00	47.27	34.95	43.54	44.50	62.50	28.50	43.00	44.67
Audio-Thinker (Wu et al., 2026)	63.66	58.69	56.46	59.60	47.88	38.35	46.26	46.80	64.50	29.50	45.00	46.33
Label-Free												
AQA-TTRL (Zhang et al., 2025)	61.56	58.98	53.75	58.10	46.67	33.01	42.18	42.20	59.50	28.00	37.00	41.50
Audio-CoT (Ma et al., 2025a)	61.26	58.38	53.75	57.80	46.06	31.55	41.50	41.50	57.50	26.00	34.00	39.17
Audio-Zero (Ours)	**64.56**	**58.69**	**58.36**	**60.50**	**48.48**	**41.26**	**47.96**	**49.10**	**66.50**	**30.00**	**47.50**	**48.00**
Qwen2.5-Omni-7B												
Base Policy	70.57	68.26	63.06	68.30	56.36	47.09	59.86	56.20	90.50	58.50	61.00	70.00
Label-Dependent												
R1-AQA (Li et al., 2025)	76.58	70.66	70.27	72.50	63.64	50.49	60.54	60.80	92.00	61.00	63.50	72.17
Audio-Thinker (Wu et al., 2026)	79.28	71.56	72.97	74.60	66.67	53.40	61.90	63.70	92.50	63.00	64.00	73.17
Label-Free												
AQA-TTRL (Zhang et al., 2025)	72.67	69.46	66.67	69.60	59.39	48.54	60.20	57.50	91.00	59.50	62.00	70.83
Audio-CoT (Ma et al., 2025a)	71.47	68.86	66.37	68.90	57.58	47.57	59.86	56.80	89.00	57.00	59.00	68.33
Audio-Zero (Ours)	**81.68**	**72.75**	**74.47**	**76.30**	**69.70**	**55.83**	**60.54**	**66.20**	**93.00**	**64.50**	**64.50**	**74.00**

Table 1: Comparison with label-dependent and label-free baselines on general and fine-grained audio reasoning benchmarks. Bold indicates the best performance for each base model, and underline indicates the best result among the baseline methods.

and *Audio-Thinker* (Wu et al., 2026), which represent strong reinforcement learning and explicit reasoning baselines optimized via ground-truth answers. For the label-free group, we compare against (1) *AQA-TTRL* (Zhang et al., 2025), a test-time reinforcement learning paradigm that dynamically optimizes the model's policy via majority-consensus pseudo-label feedback and (2) *Audio-CoT* (Ma et al., 2025a), a training-free prompt-based method encouraging intermediate reasoning. We evaluate all methods across two foundation backbones: Qwen2-Audio-7B-Instruct and Qwen2.5-Omni-7B.

As reported in Table 1, Audio-Zero consistently achieves the best overall performance across all three benchmarks and both backbones, outperforming not only all label-free test-time approaches but also the strong label-dependent training baselines. Specifically, using merely 2k unlabeled contrast pairs, Audio-Zero reaches an average accuracy of 60.50% on MMAU Test-mini and 49.10% on MMAR under Qwen2-Audio-7B-Instruct, outperforming the base policy by 3.00% and 7.90% respectively. On the stronger Qwen2.5-Omni-7B backbone, Audio-Zero further yields substantial gains, boosting MMAU Avg. by 8.00% and MMAR Avg. by 10.00%. This comprehensive improvement confirms that converting unlabeled audio contrast pairs into an interactive self-play game can unlock latent perception capabilities, boosting both general multi-modal understanding and specialized audio reasoning without any external supervision.

Crucially, the performance gains are particularly pronounced on the TREA subsets, which explicitly target fine-grained, multi-step temporal reasoning (e.g., reaching 66.50% on Order and 47.50% on Duration under Qwen2-Audio-7B-Instruct). This dual improvement across both general and fine-grained benchmarks directly reflects the efficacy of our dual-reward framework. While standard label-free methods struggle with vague feedback, Audio-Zero establishes a dynamic game-theoretic constraint between the two optimization stages. Within this loop, the attribution stage acts as a judge that evaluates game outcomes, which inherently forces the listening stage to produce more detailed, grounded clues. Consequently, the interactive competition inherently drives the model to capture and process more fine-grained acoustic information, validating the potential of self-evolutionary training.

3.3 Ablation Study

The core design of Audio-Zero aims to achieve a joint self-evolution of both acoustic perception and multi-step reasoning. To verify whether this evolutionary loop requires the synergy of all components, we evaluate five architectural variants across both Qwen2-Audio-7B-Instruct and Qwen2.5-Omni-7B: (1) *Vanilla GRPO w/o game*, a standard single-stage RL baseline; (2) *Listening-stage only* and (3) *Attribution-stage only*, which restrict optimization to a single isolated stage; (4) *w/o Vote-aware feedback*, which removes the collective voting process; and (5) *Sequential training*, which optimizes the two stages in a static, one-way sequence without iterative interaction.

Variant	Components			Qwen2-Audio-7B-Instruct			Qwen2.5-Omni-7B		
	List.	Attr.	Inter.	TREA	MMAU Test-mini	MMAR	TREA	MMAU Test-mini	MMAR
Vanilla GRPO (w/o game)	✗	✗	✗	44.67	59.00	44.50	72.17	72.50	60.80
Listening-stage only	✓	✗	✗	45.33	59.40	45.80	72.50	73.30	61.90
Attribution-stage only	✗	✓	✗	45.83	59.10	47.20	72.83	72.70	63.10
w/o Vote-aware feedback	✓	✓	✓	45.50	59.20	46.80	72.67	72.90	62.70
Sequential training	✓	✓	✗	46.83	60.10	47.90	73.33	74.80	64.50
Audio-Zero (Full)	✓	✓	✓	**48.00**	**60.50**	**49.10**	**74.00**	**76.30**	**66.20**

Table 2: Ablation studies on the Qwen2-Audio-7B-Instruct and Qwen2.5-Omni-7B backbones. List., Attr., and Inter. denote listening-stage optimization, attribution-stage optimization, and vote-aware interaction feedback, respectively.

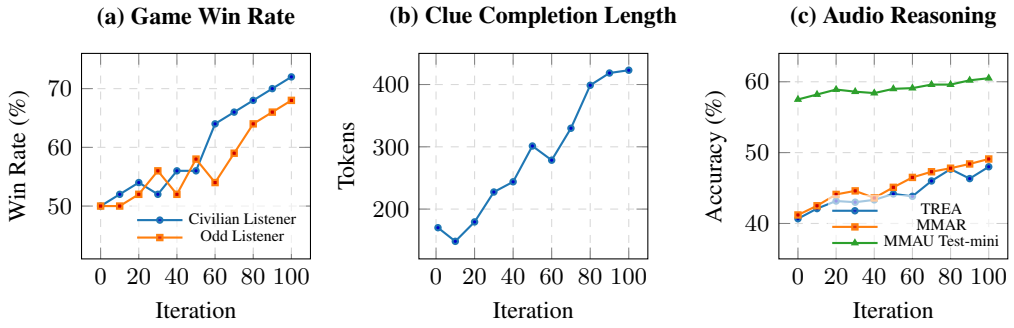


Figure 3: Analysis of self-evolution dynamics across training iterations. The subfigures show (a) game win rates under distinct role assignments, (b) average lengths of generated audio clues, and (c) performance on downstream audio reasoning benchmarks.

As shown in Table 2, removing any single component leads to a noticeable performance decline. Isolating either the listening or attribution stage yields marginal improvements, showing that single-role updates cannot sustain a continuous self-evolution loop. Crucially, disabling the vote-aware feedback causes a distinct performance drop on the TREA benchmark (from 48.00% to 45.50% on Qwen2; 74.00% to 72.67% on Qwen2.5). This empirical drop clearly verifies that collective voting is indispensable, as it suppresses sporadic hallucinations and guides the model with more reliable feedback. Furthermore, the performance gap between *Sequential training* and the full framework underscores that alternating optimization serves as the cornerstone of our framework. Instead of leading to training stagnation, this dynamic alternation drives the co-evolution of both roles: as the two stages update alternately, they continuously push each other forward, ultimately forcing the listening model to deliver increasingly precise and fine-grained acoustic facts in its reasoning steps.

3.4 Self-Evolution Dynamics

The training dynamics are examined to characterize how the policy changes over optimization. At fixed evaluation intervals from 0 to 100 iterations, the active policy plays self-play games against the frozen initial policy under both civilian-listener and odd-listener role assignments across identical game setups. Game win rates, average listening-stage clue length, and downstream performance on external audio reasoning benchmarks are tracked at the same intervals.

Figure 3 reveals three related trends. First, game win rates generally increase for both role assignments, despite local fluctuations. This pattern is consistent with improving identification of subtle acoustic differences and more reliable reasoning over generated evidence. Second, listening-stage clue length increases during the early phase and subsequently stabilizes. The stabilization suggests that the policy is not improving solely by generating longer descriptions, but is balancing descriptive detail with the requirements of the game without drifting into verbosity. Third, TREA, MMAR, and MMAU Test-mini exhibit overall upward trends, with larger gains on the more fine-grained benchmarks. The joint movement of game performance and external benchmark accuracy links the internal self-play signal to improvements in auditory detail perception and cross-description consistency. Together, these observations are consistent with progressive self-evolution during training and indicate that the learned behavior successfully transfers beyond the self-play environment.

3.5 Emergence of Fine-Grained Auditory Descriptions

The diagnostic analysis examines whether the game pressure in Audio-Zero is associated with more fine-grained auditory descriptions rather than with benchmark accuracy alone. Five metrics are tracked over training iterations. They are *Distinct Event Count*, *Temporal Relation Count*, *Quantity Expression Frequency*, *Acoustic Attribute Density*, and *Vagueness Rate*. Detailed definitions and matching rules are provided in Appendix C.

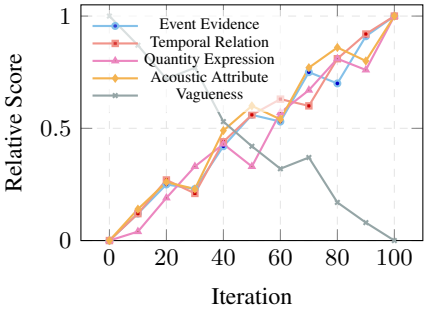


Figure 4: Diagnostic analysis of generated listening-stage descriptions across training iterations. All metrics are independently min-max scaled to $[0, 1]$ for visualization after length normalization where applicable. Higher values indicate stronger fine-grained auditory evidence, except for Vagueness, where lower values indicate better specificity.

As shown in Figure 4, the detail-related metrics generally increase during training, whereas Vagueness decreases. The Event Evidence and Temporal Relation scores indicate a shift from generic audio summaries toward descriptions that mention concrete sound events and organize them into coherent temporal structures. The increase in Quantity Expression is consistent with greater sensitivity to repetitions and counting-related details. The Acoustic Attribute trend suggests greater attention to properties such as loudness, distance, clarity, and acoustic texture. The steady decline in Vagueness further indicates reduced reliance on underspecified phrases such as "some sounds" or "various noises." Because the metrics are independently scaled, the analysis focuses on their relative trajectories rather than on absolute values across metrics. The results therefore consistently suggest that the self-play game changes the content of the descriptions, rather than merely increasing their length.

Attribution pressure is associated with descriptions that are more event-rich, temporally grounded, acoustically detailed, and discriminative across players. This pattern is consistent with the emergence of fine-grained auditory detail perception through the self-evolution process.

3.6 Case Study

Table 3 provides a qualitative illustration of the training dynamics on a representative temporal-reasoning example. At Stage 1, the base policy produces coarse and underspecified descriptions. The attribution stage consequently relies mainly on surface event overlap and fails to identify the odd listener. At Stage 2, the descriptions begin to represent the basic event sequence, allowing the attribution stage to detect a local ordering mismatch, although the reasoning remains limited. At Stage 3, the generated clues contain more fine-grained evidence, including event order, repetition count, relative acoustic strength, and background context. The attribution stage then compares the majority timeline with each player's description, localizes the temporal inconsistency, and correctly identifies the odd listener. This progression provides qualitative evidence that the framework improves both listening-stage detail perception and attribution-stage reasoning over training.

4 Related Work

Large Audio Language Models LALMs connect acoustic encoders with language models to support audio instruction following, reasoning, and spoken interaction. Qwen2-Audio trains a unified model across speech, sound, and music (Chu et al., 2024), while Qwen3-Omni and Qwen3.5-Omni (Xu et al., 2025c; Team, 2026) extend this toward multimodal interaction with strong audio-visual capabilities. GLM-4-Voice and MiniCPM-o 4.5 (Zeng et al., 2024; Cui et al., 2026) also explore audio foundation models for understanding, generation, and speech. While providing strong baselines, these models heavily rely on audio-text supervision, task annotations, or instruction data. Because systematic fine-grained annotation is costly, LALMs still struggle with temporal ordering, counting, duration, and deep audio reasoning (Kumar et al., 2025; Sakshi et al., 2024; Ma et al., 2025b; Bhattacharya et al., 2025). This motivates post-training methods that reduce reliance on content annotations while strengthening fine-grained auditory reasoning.

Post-Training and Self-Training for Audio Models A growing line of work aims to reduce annotation costs for audio models. SI-SDA uses self-improvement and pseudo-label filtering for unlabeled

Case Study: Behavioral Emergence on Complex Temporal Relation Reasoning	
Audio Inputs	**Reference Listeners (Majority):** [Glass Shattering → Dog Barking → Door Slamming] **Odd Listener (Anomaly):** [Dog Barking → Glass Shattering → Door Slamming]
Stage 1: Base Policy	**Reference Clue:** "There are several loud sounds in an indoor scene, including a dog and something breaking." **Odd-Listener Clue:** "I hear a dog barking and some sharp noise in the background." **Attribution (Judge):** "Both descriptions mention a dog and a sharp breaking sound, but neither gives a reliable event order. The inconsistency is unclear. **Guess: Player 1 (Wrong)**."
Stage 2: Mid-Train	**Reference Clue:** "A sharp glass-breaking sound occurs first, then a dog barks and a door slams near the end." **Odd-Listener Clue:** "A dog barks first, then glass shatters and a door slams." **Attribution (Judge):** "Most descriptions place glass before the dog bark, while one description places the dog bark before the glass. This suggests a temporal mismatch. **Guess: Player 2 (Correct)**."
Stage 3: Audio-Zero	**Reference Clue:** "'In a noisy indoor room with people talking in the background, I hear a sharp **glass shattering** sound first. Approximately **1 seconds later**, there are two quick **dog barks** and the clip ends with a heavy **door slam**." **Odd-Listener Clue (Masquerade):** "Inside a noisy indoor room, I can hear some people talking while a **sharp glass shattering**, a loud **dog bark** and a heavy **door slam** occur in close succession." **Attribution (Judge):** "Although Player 2 strategically incorporates the background context (noisy room, people talking) and matches the core event vocabulary to blend in with the majority, it completely omits the relative timeline. The most descriptions explicitly establish a strict [Glass → 1s → Dog] chronological sequence. Despite reporting identical acoustic facts, Player 2's ambiguous alignment exposes its anomaly. **Conclusion: Player 2 (Correct).**"

Table 3: Qualitative illustration of the self-evolution process on a temporal reasoning example. The examples show a progression from coarse event descriptions to temporally structured and discriminative auditory evidence.

speech adaptation (Wang et al., 2025c). AQA-TTRL performs test-time reinforcement learning for audio question answering using majority-vote pseudo labels (Zhang et al., 2025). Synthetic bootstrapping methods generate machine-authored audio-text pairs to mitigate hallucinations and augment training data without human intervention (Kuan & Lee, 2025). These methods reduce supervision needs, but they usually train models to imitate pseudo-labels, exploit synthetic supervision, or adapt to a specific test distribution. Audio-Zero differs by constructing a verifiable interaction environment: the model does not need to trust its own pseudo-answer as a label; success is determined by whether the hidden odd listener can be identified from generated clues.

Self-Play and Self-Evolution in LLMs and VLMs Self-play has emerged as a promising paradigm for bootstrapping language models without relying on static human annotations or external reward models. In the text domain, frameworks rooted in SPIN (Chen et al., 2024) enable models to self-improve through iterative generation and discrimination, a philosophy rapidly extended to zero-data self-training (Liu et al., 2025a; Xu et al., 2025a; Huang et al., 2025; Zhao et al., 2026) and specialized environments—such as strategic search, long-context reasoning, and multi-agent systems—where agents leverage self-guided exploration or mutual feedback to autonomously upgrade their capabilities (Liu et al., 2025b; Yang et al., 2025b; Wang et al., 2025b). Similarly, vision-language approaches like Visual Self-Play and Vision-Zero (He et al., 2025; Wang et al., 2025a; Tong et al., 2025) demonstrate that structured visual cooperative-competitive interactions can convert unlabeled multimodal differences into verifiable learning signals for robust decision-making.

5 Conclusion

In this work, we present Audio-Zero, a label-free self-evolution framework for improving fine-grained auditory perception and reasoning in large audio-language models. Instead of relying on human-annotated transcripts, QA labels, or reasoning traces, Audio-Zero converts unlabeled audio contrast pairs into an auditory self-play game, where the model alternates between generating listening-stage descriptions and performing attribution-stage reasoning to identify the odd listener. Experiments on Qwen2-Audio-7B-Instruct and Qwen2.5-Omni-7B demonstrate consistent gains on both general audio reasoning benchmarks and fine-grained temporal reasoning benchmarks. Specifically, Audio-Zero improves the two backbones by 7.33% and 4.00% on TREA, 7.90% and 10.00% on MMAR and 3.00% and 8.00% on MMAU Test-mini, respectively. Further analyses show that the game-derived feedback encourages the emergence of more event-rich, acoustically detailed and discriminative auditory descriptions. These results suggest that self-play over unlabeled audio contrasts is a promising direction for scalable post-training of LALMs toward stronger fine-grained auditory reasoning.

AI USE STATEMENT

We used AI tools to assist with manuscript writing and language polishing, as well as coding tasks during the experimental workflow. The authors take full responsibility for the content, experimental results, and conclusions presented in this paper.

REFERENCES

Debarpan Bhattacharya, Apoorva Kulkarni, and Sriram Ganapathy. Benchmarking and confidence evaluation of lalms for temporal reasoning. *arXiv preprint arXiv:2505.13115*, 2025.

Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, and Quanquan Gu. Self-play fine-tuning converts weak language models to strong language models. *arXiv preprint arXiv:2401.01335*, 2024.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, et al. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*, 2024.

Junbo Cui, Bokai Xu, Chongyi Wang, Tianyu Yu, Weiyue Sun, Yingjing Xu, Tianran Wang, Zhihui He, Wenshuo Ma, Tianchi Cai, et al. Minicpm-o 4.5: Towards real-time full-duplex omni-modal interaction. *arXiv preprint arXiv:2604.27393*, 2026.

Soham Deshmukh, Benjamin Elizalde, Rita Singh, and Huaming Wang. Pengi: An audio language model for audio tasks. *Advances in Neural Information Processing Systems*, 36:18090–18108, 2023.

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. *arXiv preprint arXiv:2510.20867*, 2025.

Yicheng He, Chengsong Huang, Zongxia Li, Jiaxin Huang, and Yonghui Yang. Visplay: Self-evolving vision-language models from images. *arXiv preprint arXiv:2511.15661*, 2025.

Chengsong Huang, Wenhao Yu, Xiaoyang Wang, Hongming Zhang, Zongxia Li, Ruosen Li, Jiaxin Huang, Haitao Mi, and Dong Yu. R-zero: Self-evolving reasoning llm from zero data. *arXiv preprint arXiv:2508.05004*, 2025.

Chun-Yi Kuan and Hung-yi Lee. From alignment to advancement: Bootstrapping audio-language alignment with synthetic data. *IEEE Transactions on Audio, Speech and Language Processing*, 2025.

Sonal Kumar, Šimon Sedláček, Vaibhavi Lokegaonkar, Fernando López, Wenyi Yu, Nishit Anand, Hyeonggon Ryu, Lichang Chen, Maxim Plička, Miroslav Hlaváček, et al. Mmau-pro: A challenging and comprehensive benchmark for holistic evaluation of audio general intelligence. *arXiv preprint arXiv:2508.13992*, 2025.

Gang Li, Jizhong Liu, Heinrich Dinkel, Yadong Niu, Junbo Zhang, and Jian Luan. Reinforcement learning outperforms supervised fine-tuning: A case study on audio question answering. *arXiv preprint arXiv:2503.11197*, 2025.

Bo Liu, Leon Guertler, Simon Yu, Zichen Liu, Penghui Qi, Daniel Balcells, Mickel Liu, Cheston Tan, Weiyan Shi, Min Lin, et al. Spiral: Self-play on zero-sum games incentivizes reasoning via multi-agent multi-turn reinforcement learning. *arXiv preprint arXiv:2506.24119*, 2025a.

Bo Liu, Chuanyang Jin, Seungone Kim, Weizhe Yuan, Wenting Zhao, Ilia Kulikov, Xian Li, Sainbayar Sukhbaatar, Jack Lanchantin, and Jason Weston. Spice: Self-play in corpus environments improves reasoning. *arXiv preprint arXiv:2510.24684*, 2025b.

Ziyang Ma, Zhuo Chen, Yuping Wang, Eng-Siong Chng, and Xie Chen. Audio-cot: Exploring chain-of-thought reasoning in large audio language model. In *2025 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 1–6. IEEE, 2025a.

Ziyang Ma, Yinghao Ma, Yanqiao Zhu, Chen Yang, Yi-Wen Chao, Ruiyang Xu, Wenxi Chen, Yuanzhe Chen, Zhuo Chen, Jian Cong, et al. Mmar: A challenging benchmark for deep reasoning in speech, audio, music, and their mix. *arXiv preprint arXiv:2505.13032*, 2025b.

Navonil Majumder, Chia-Yu Hung, Deepanway Ghosal, Wei-Ning Hsu, Rada Mihalcea, and Soujanya Poria. Tango 2: Aligning diffusion-based text-to-audio generations through direct preference optimization. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 564–572, 2024.

Xinhao Mei, Chutong Meng, Haohe Liu, Qiuqiang Kong, Tom Ko, Chengqi Zhao, Mark D Plumbley, Yuexian Zou, and Wenwu Wang. Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:3339–3354, 2024.

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.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Yi-Jen Shih, Desh Raj, Chunyang Wu, Wei Zhou, SK Bong, Yashesh Gaur, Jay Mahadeokar, Ozlem Kalinli, and Mike Seltzer. Can speech llms think while listening? *arXiv preprint arXiv:2510.07497*, 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*, volume 2024, pp. 16607–16629, 2024.

Qwen Team. Qwen3. 5-omni technical report. *arXiv preprint arXiv:2604.15804*, 2026.

Jingqi Tong, Jixin Tang, Hangcheng Li, Yurong Mou, Ming Zhang, Jun Zhao, Yanbo Wen, Fan Song, Jiahao Zhan, Yuyang Lu, et al. Game-rl: Synthesizing multimodal verifiable game data to boost vlms' general reasoning. *arXiv preprint arXiv:2505.13886*, 2025.

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*, 2026.

Qinsi Wang, Bo Liu, Tianyi Zhou, Jing Shi, Yueqian Lin, Yiran Chen, Hai Helen Li, Kun Wan, and Wentian Zhao. Vision-zero: Scalable vlm self-improvement via strategic gamified self-play. *arXiv preprint arXiv:2509.25541*, 2025a.

Shaobo Wang, Zhengbo Jiao, Zifan Zhang, Yilang Peng, Xu Ze, Boyu Yang, Wei Wang, Hu Wei, and Linfeng Zhang. Socratic-zero: Bootstrapping reasoning via data-free agent co-evolution. *arXiv preprint arXiv:2509.24726*, 2025b.

Shaowen Wang, Xinyuan Chen, and Yao Xu. Self-improvement for audio large language model using unlabeled speech. *arXiv preprint arXiv:2507.20169*, 2025c.

Shu Wu, Chenxing Li, Wenfu Wang, Hao Zhang, Hualei Wang, Meng Yu, and Dong Yu. Audio-thinker: Guiding large audio language model when and how to think via reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 33962–33970, 2026.

Fangzhi Xu, Hang Yan, Chang Ma, Haiteng Zhao, Qiushi Sun, Kanzhi Cheng, Junxian He, Jun Liu, and Zhiyong Wu. Genius: A generalizable and purely unsupervised self-training framework for advanced reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13153–13167, 2025a.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-omni technical report, 2025b.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025c.

Chih-Kai Yang, Neo Ho, Yen-Ting Piao, and Hung-yi Lee. Sakura: On the multi-hop reasoning of large audio-language models based on speech and audio information. *arXiv preprint arXiv:2505.13237*, 2025a.

Runyan Yang, Yuke Si, Yingying Gao, Junlan Feng, Chao Deng, and Shilei Zhang. Teaching audio models to reason: A unified framework for source-and layer-wise distillation. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 15357–15361. IEEE, 2026.

Ziyi Yang, Weizhou Shen, Chenliang Li, Ruijun Chen, Fanqi Wan, Ming Yan, Xiaojun Quan, and Fei Huang. Spell: Self-play reinforcement learning for evolving long-context language models. *arXiv preprint arXiv:2509.23863*, 2025b.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. Star: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems*, 35:15476–15488, 2022.

Aohan Zeng, Zhengxiao Du, Mingdao Liu, Kedong Wang, Shengmin Jiang, Lei Zhao, Yuxiao Dong, and Jie Tang. Glm-4-voice: Towards intelligent and human-like end-to-end spoken chatbot. *arXiv preprint arXiv:2412.02612*, 2024.

Haoyu Zhang, Jiaxian Guo, Yusuke Iwasawa, and Yutaka Matsuo. Aqa-ttrl: Self-adaptation in audio question answering with test-time reinforcement learning. *arXiv preprint arXiv:2510.05478*, 2025.

Andrew Zhao, Yiran Wu, Tong Wu, Quentin Xu, Yang Yue, Matthieu Lin, Shenzhi Wang, Qingyun Wu, Zilong Zheng, and Gao Huang. Absolute zero: Reinforced self-play reasoning with zero data. *Advances in Neural Information Processing Systems*, 38:105816–105879, 2026.

A Appendix

A.1 Additional Implementation Details

A.1.1 Training Data Construction

We instantiate Audio-Zero using Audio-Alpaca (Majumder et al., 2024), a paired audio preference dataset where each example contains a prompt, a chosen audio, and a rejected audio. We treat the chosen and rejected audios as an audio contrast pair $(a^{\mathrm{ref}}, a^{\mathrm{var}})$. The two clips share prompt-level semantics but may differ in event order, missing or extra events, repetitions, acoustic quality, or background scene.

We sample and filter 2,000 pairs from the full collection for post-training. We remove examples whose audio files cannot be loaded, whose duration is too short or too long for stable batching, or whose paired audios are near-duplicates after manual or automatic inspection. All selected clips are resampled to 16 kHz and truncated or padded to a fixed duration during training. The dataset is only used as a source of unlabeled auditory contrast; no caption, QA label, or reasoning trace is used as supervision.

A.1.2 Game Configuration

Unless otherwise specified, each self-play game uses $N = 4$ players and $R = 2$ listening rounds. For each game, one player is uniformly sampled as the odd listener. The odd listener receives a^{var}, while the remaining players receive a^{ref}. The model sequentially plays all player roles and generates one listening-stage clue for each player in each round. After all clues are generated, the model acts as the judge and predicts the odd listener based on the full clue history.

The odd-listener identity is never provided to the model as input. It is only used by the environment to compute the attribution reward. This ensures that the training signal is verifiable by construction while remaining independent of human-annotated content labels.

A.1.3 Reward Implementation

The attribution reward is implemented as a binary correctness signal:

$$R_{\mathrm{att}} = \mathbb{I}[\hat{s} = s]. \tag{10}$$

For listening-stage descriptions, we combine a content utility reward with a vote-aware penalty. Mathematically, the total reward allocated to player i is formalized as:

$$R^i_{\mathrm{list}} = R^i_{\mathrm{utility}} \; - \; \lambda \cdot \frac{v_i}{\max(1,\, V-1)} \tag{11}$$

Here, v_i is the number of attribution votes received by player i, $V = G \times N$ is the total number of voters, and λ is the vote penalty coefficient. In our implementation, we set $\lambda = 0.6$.

The content utility reward R^i_{utility} is computed with an automated rule-based scorer. It rewards clues that mention concrete auditory events, temporal relations, quantity expressions, acoustic attributes, and scene context. It penalizes overly vague descriptions, identity leakage, and repeated phrasing across rounds. The vote-aware penalty discourages clues that make a player appear suspicious, encouraging descriptions that are informative and discriminative without directly revealing the hidden role.

A.1.4 GRPO Hyperparameters

Table 4 summarizes the main hyperparameters used in our post-training experiments.

B Evaluation Protocol

We evaluate all models using multiple-choice question accuracy. For MMAU Test-mini, we report results on Sound, Music, and Speech subsets as well as the overall average. For TREA, we report

Hyperparameter	Value
Number of players N	4
Listening rounds R	2
GRPO completions G	6
Training pairs	2,000
Learning rate	1×10^{-5}
KL coefficient β_{KL}	0.04
Vote penalty λ	0.6
Precision	`bfloat16`
Optimizer / Trainer	GRPO

Table 4: Main hyperparameters for Audio-Zero post-training.

Metric	Definition	Target Examples
Distinct Event Count	Number of unique auditory event words mentioned in a clue. It measures whether the model describes concrete sound events rather than generic audio content.	speech, footsteps, knock, music, glass shattering
Temporal Relation Count	Number of explicit ordering relations between events. It measures whether the model organizes events into temporal structure.	after, before, followed by, then, first
Quantity Expression Frequency	Frequency of quantity or repetition expressions. It measures whether the model captures counting-related auditory details.	twice, repeated, several, multiple, two barks
Acoustic Attribute Density	Density of acoustic or auditory attribute words. It measures whether the model describes sound quality and perceptual properties.	loud, faint, muffled, distant, clear, sharp, heavy
Vagueness Rate	Proportion of clues containing vague or underspecified phrases. Lower values indicate more specific descriptions.	some sounds, various noises, something happens

Table 5: Diagnostic metrics for analyzing listening-stage descriptions.

Order, Count, and Duration subsets. For MMAR, we report Sound, Music, and Speech subsets, and also compute the overall score over the full benchmark, including mixed-modality samples.

For each question, the model is prompted to output only the option letter or the exact choice text depending on the benchmark format. Generation is performed greedily without sampling. We use the same evaluation script and answer extraction rules for all methods to ensure fair comparison.

C Diagnostic Metrics

To analyze whether fine-grained auditory descriptions emerge during training, we compute five diagnostic metrics over listening-stage clues. All feature-based metrics are length-normalized where applicable and min-max scaled to $[0, 1]$ for visualization in Figure 4. Table 5 summarizes their rigorous matching definitions and conceptual descriptions.

D Prompt Templates

We provide the exact input templates wrapper used during training in the following structured layout.

D.1 Listening Stage Instruction Template

The model is instructed to describe what the current player hears without revealing its hidden role. The prompt emphasizes concrete auditory evidence, including events, order, repetitions, acoustic attributes, and background scene.

Prompt Template: Clue Generation (Listening Round r)

`[System Instruction]`
You are Player i in an audio self-play game. Listen to your private audio and provide a concise clue describing what you hear. Mention concrete sound events, temporal order, repetitions, acoustic properties, and scene context when relevant. Do not state whether you are the odd listener.

Input Audio:
Public Clue History:

D.2 Attribution Stage Instruction Template

The model is instructed to act as the judge and identify the odd listener from the full clue history.

Prompt Template: Odd-Listener Identification (Judge Perspective)

`[System Instruction]`
You are the judge in an audio self-play game. Most players heard the same reference audio, while one odd listener heard a subtle variant. Given the full set of player clues, compare their descriptions and identify which player most likely heard the variant audio. Return the player number.

Reference Audio Input:
Compiled Player Clues:

E Additional Notes on Reproducibility

All experiments are run with the same random seed for data sampling and role assignment unless otherwise specified. For checkpoint evaluation, we use the same benchmark splits and answer extraction rules across all methods. For diagnostic analysis, the same set of generated listening-stage clues is used to compute all metrics at each training iteration.