# ACE-CAP: ACTIVE EVIDENCE ACQUISITION VIA AGENTIC CO-EVOLUTION FOR LONG-PARAGRAPH FINE-GRAINED AUDIO CAPTIONING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Long-paragraph fine-grained audio captioning requires models to recover diverse acoustic facts while avoiding omissions and unsupported details. However, prevailing captioners remain passive one-shot generators: once a relevant detail is overlooked, they cannot identify the evidence gap, query the audio for targeted information, or adaptively determine when sufficient evidence has been collected. We formulate this task as active evidence acquisition and introduce **A**gentic **C**o-**E**volution for **Cap**tioning (ACE-Cap), a framework that couples a Composer and an Instruct model through multi-turn interaction. A Captioner first produces an initial description. Conditioned only on this description and the interaction history, a text-only Composer asks targeted questions about unresolved acoustic attributes, while an audio-conditioned Instruct model provides grounded answers. The Composer then decides when to terminate the interaction and synthesizes the accumulated evidence into a final caption. ACE-Cap trains these roles through a unified gold-to-prediction reward derived from fixed, gold-grounded multiple-choice questions and a frozen caption-only judge. To assign credit within variable-length interactions, we introduce **L**eave-**O**ne-**O**ut **P**er-turn **G**roup **R**elative **P**olicy **O**ptimization (LOOP-GRPO). It replaces the trajectory-wide scalar advantage with span-aligned signals for individual questions, stopping decisions, and final synthesis. These signals measure leave-one-out evidence contributions, evidence sufficiency, and evidence preservation, respectively. Role-wise warm-up followed by alternating Composer and Instruct optimization keeps each update a well-defined single-policy problem while allowing the two roles to co-evolve. Extensive experiments across diverse audio benchmarks demonstrate state-of-the-art performance among evaluated models with open-weight backbones under a common caption-as-evidence protocol.

## 1 INTRODUCTION

**A**utomated **A**udio **C**aptioning (AAC) translates an acoustic scene into a natural-language description. Advances in audio–language architectures and fine-grained datasets have improved caption fluency and accuracy (Mei et al., 2021; Ghosh et al., 2025; Chen et al., 2025). The emerging goal, however, is not merely to produce a longer description. A useful fine-grained caption must recover speech, music, sound events, ambience, recording characteristics, and temporal relations. Covering these diverse acoustic facts exposes a limitation of the prevailing captioning paradigm.

Most captioners remain *passive one-shot generators*: they encode the audio and commit to a complete description in a single decoding process. Once a detail is overlooked, they have no explicit mechanism to diagnose the gap or revisit the audio with a targeted query. Increasing model or data scale improves the underlying representation, but does not by itself provide an active information-acquisition procedure. Consequently, omitted or vague details can persist into the final caption without an opportunity for targeted correction.

Moving beyond this limitation calls for *active evidence acquisition*, rather than simply generating more text (Fig. 1). A captioner must identify what remains unresolved, obtain additional acoustic evidence, and determine when that evidence is sufficient for a final description. Learning such



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.

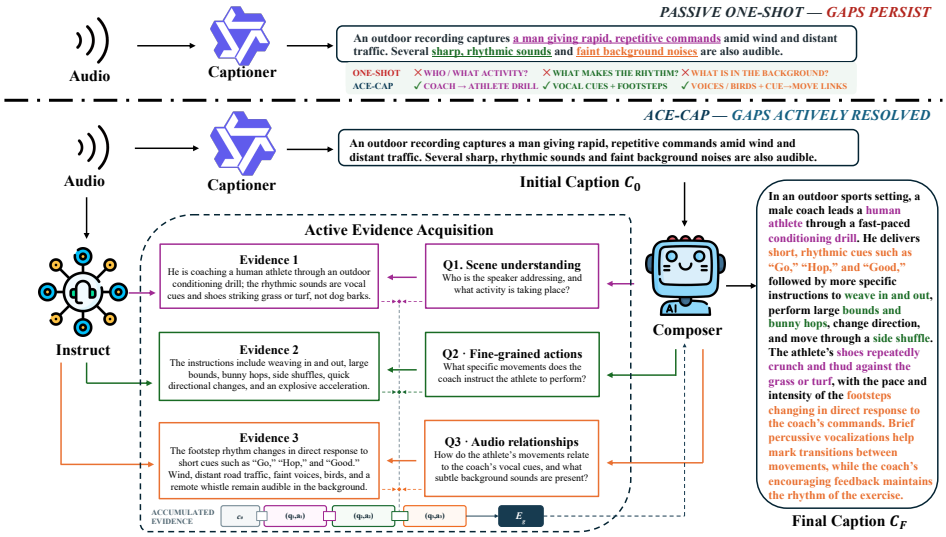



Figure 1: Passive captioning versus active evidence acquisition. A one-shot captioner produces a description directly (top); our framework acquires targeted answers before synthesis (bottom). For interaction $g$, $C_0 \equiv c_0$ is the initial caption, $C^F \equiv c_g^F$ the final caption, $(q_t, a_t)$ the question–answer pair at turn $t$, and $E_g$ the accumulated evidence. This example illustrates the mechanism, not exhaustive event recovery.

behavior requires addressing two challenges that completion-level caption supervision does not directly resolve.

The first is *credit assignment across heterogeneous actions*. Existing caption metrics primarily evaluate completed outputs (Zhou et al., 2022; Wang et al., 2024; Luong et al., 2025), while **G**roup **R**elative **P**olicy **O**ptimization (GRPO) applies a completion-level advantage across generated tokens (Shao et al., 2024). In an interaction, however, a question may uncover a unique event or repeat known information; stopping may be premature, and final synthesis may discard acquired evidence. A shared outcome score does not distinguish these contributions. Adjacent reward differences also depend on the order-specific history and can undervalue later questions after the score has partially saturated. Useful supervision must therefore distinguish evidence acquisition, termination, and evidence preservation.

The second is *non-stationarity between interacting roles*. As a questioner changes, the answerer encounters a different question distribution; as the answerer changes, the questioner receives different observations. Each role is thus optimized against a moving target (Papoudakis et al., 2019). Alternating updates can mitigate this coupling (Nekoei et al., 2023), but initially weak partners may still provide poor training signals. This motivates both role-specific preparation and updates that hold the interacting partner fixed.

To address these challenges, we propose **A**gentic **C**o-**E**volution for **Cap**tioning (ACE-Cap), a framework for fine-grained audio captioning. A Captioner produces a provisional description rather than an irrevocable output. A text-only Composer identifies unresolved attributes and asks targeted questions, while an audio-conditioned Instruct model supplies grounded answers. The Composer then decides when to stop and synthesizes the accumulated evidence. This separates the acquisition of missing information from its final expression as a caption.

For credit assignment, a unified **G**old-to-**P**rediction (G2P) reward evaluates textual states using fixed gold-derived **M**ultiple-**C**hoice **Q**uestions (MCQs) and a frozen caption-only judge. Our **L**eave-**O**ne-**O**ut **P**er-turn **G**roup **R**elative **P**olicy **O**ptimization (LOOP-GRPO) adapts counterfactual credit assignment (Foerster et al., 2018): question credit measures the reward lost when a turn is removed from the completed evidence state. Separate utilities train stopping for evidence sufficiency and final synthesis for evidence preservation. For coupled training, role-wise **S**upervised **F**ine-**T**uning (SFT)

and GRPO warm-up precede alternating Composer–Instruct updates, with the Captioner thereafter frozen.

Our contributions are threefold:

- We formulate fine-grained audio captioning as **active evidence acquisition** and introduce ACE-Cap. Starting from a provisional caption, a text-only Composer directs audio-grounded question–answer turns, adaptively terminates acquisition, and synthesizes accumulated evidence. This decomposition provides an explicit mechanism for revisiting unresolved acoustic attributes before committing to the final description.
- We introduce **LOOP-GRPO under a unified G2P reward**, assigning separate credits to what to ask, when to stop, and how to preserve evidence. Question credit measures each turn's leave-one-out contribution to the completed evidence state, rather than comparing successive prefixes. Type-wise normalization and token-span alignment connect these distinct utilities to their corresponding Composer actions.
- We develop a **warm-started co-evolution strategy** that prepares all three roles through interface-specific SFT and GRPO before alternating Composer–Instruct optimization. The Captioner then remains frozen, and each interacting policy is updated against a fixed partner. This schedule supports mutual adaptation while keeping each update a well-defined single-policy optimization problem.

## 2 Related Works

### 2.1 Audio Captioning and Agentic Interaction

Audio captioning developed through AudioCaps and Clotho (Kim et al., 2019; Drossos et al., 2020), transformer encoder–decoders (Mei et al., 2021), and weakly labeled pretraining in WavCaps (Mei et al., 2024). General audio–language models include Pengi (Deshmukh et al., 2023), Audio Flamingo (Kong et al., 2024), Audio Flamingo 2 (Ghosh et al., 2025), and FusionAudio (Chen et al., 2025). Agentic systems instead couple reasoning with actions (Yao et al., 2022), coordinate perceptual experts (Yang et al., 2023), support audio dialogue (Kong et al., 2024), or acquire omni-modal evidence (Xing et al., 2026). These systems broaden acoustic coverage and interaction, but captioners still produce descriptions in one pass and agents primarily optimize task answers. ACE-Cap instead uses interaction to identify and repair missing acoustic evidence before synthesizing a long-form caption.

### 2.2 Reinforcement Learning for Interactive Audio Captioning

Captioning has been optimized with self-critical training (Rennie et al., 2017) and audio-specific reinforcement learning (Xu et al., 2020), supported by semantic, **L**arge **L**anguage **M**odel (LLM)-based, and factorized caption metrics (Zhou et al., 2022; Wu et al., 2025b; Wang et al., 2024). For multi-turn policies, GRPO supplies group-relative optimization (Shao et al., 2024), turn-level methods assign intermediate credit (Zeng et al., 2025), and counterfactual learning estimates action contributions (Foerster et al., 2018). Adaptive agents additionally face non-stationarity (Papoudakis et al., 2019), which multi-timescale updates can mitigate (Nekoei et al., 2023). ACE-Cap combines these directions through a gold-derived G2P reward evaluated by a frozen caption-only judge, LOOP-GRPO credits for questioning, stopping, and synthesis, and alternating Composer–Instruct optimization. This design targets evidence-preserving caption generation rather than answer generation alone.

## 3 Methodology

### 3.1 Roles and Interaction

Figure 2(a) shows the three-role interaction. A Captioner produces initial caption $c_0$ from audio $x$. A text-only Composer acquires additional acoustic evidence through an audio-conditioned Instruct model.

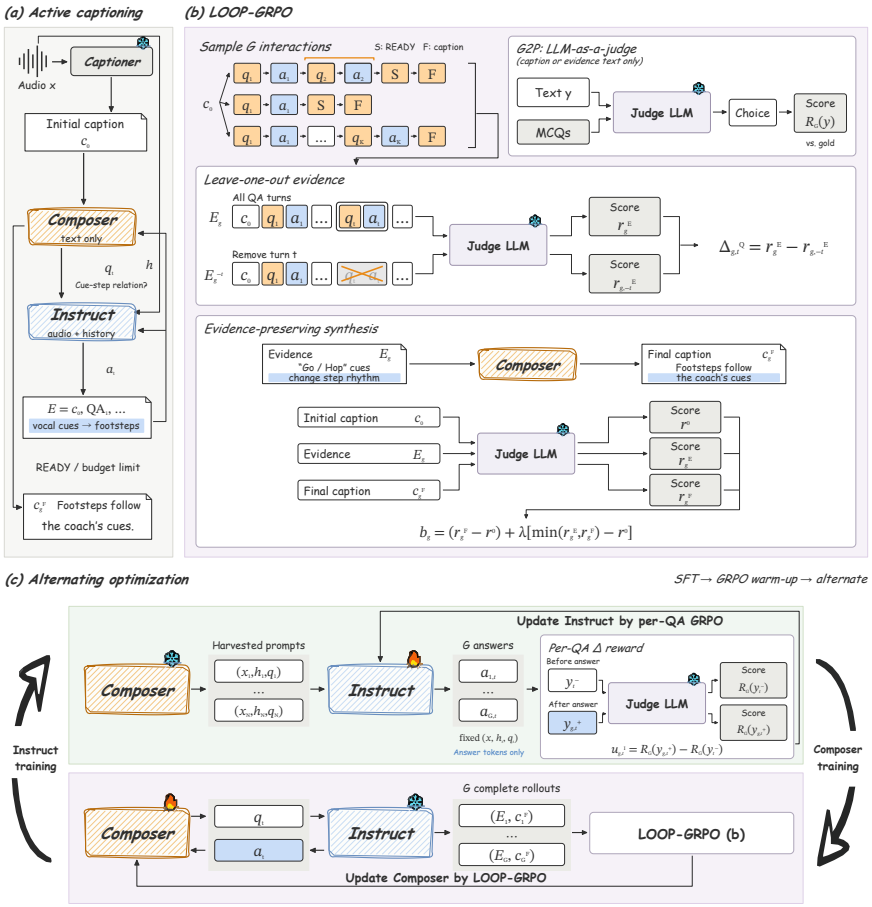



Figure 2: **Overview of ACE-Cap: active evidence acquisition and co-evolutionary training.** (a) A text-only Composer queries an audio-conditioned Instruct model to resolve details missing from the initial caption, then stops acquisition and synthesizes the accumulated evidence. (b) LOOP-GRPO assigns distinct credits to questioning, stopping, and final synthesis under a shared caption-only judge. Question credit measures the reward lost when one question–answer pair is removed from the completed evidence; the synthesis utility couples evidence and final-caption quality. (c) After role-wise warm-up, Composer and Instruct are optimized alternately against a frozen partner. Instruct learns from the reward increment of each answer; Composer learns from complete interactions using LOOP-GRPO. Snowflakes and flames indicate frozen and trainable models, respectively. All models are frozen at inference.

The Composer observes only text history $h$, asks question $q_t$ at turn $t$, and receives answer $a_t$ from the audio-conditioned Instruct. These **Q**uestion–**A**nswer (QA) pairs provide its acoustic evidence. We sample $G$ interactions per audio, indexed by $g$. Each interaction contains $T_g \leq K$ pairs, where $T_g$ is its query count and $K$ the query budget. Unindexed quantities in Figure 2(a) omit $g$. The Composer chooses ASK to query or READY to stop. *Accumulated Evidence* $E_g$ comprises $c_0$ and the completed pairs; $c_g^F$ is the final caption synthesized from it. Budget exhaustion forces finalization without a policy stop action. Full history and rollout definitions appear in Appendix A.2.

### 3.2 Unified Gold-to-Prediction Reward

For each audio $x$ and its gold training caption $c^\star$, we construct a fixed bank of $L$ gold-grounded MCQs offline. Question $j$ has gold option $o_j^\star$ among A–D; option E always means *Not Given*. For candidate text $y$, a frozen text-only judge predicts option $\hat{o}_j(y)$ without seeing audio or the gold

caption. The G2P reward averages item scores $s_j(y)$:

$$R_G(y) = \frac{1}{L}\sum_{j=1}^{L} s_j(y), \qquad s_j(y) = \begin{cases} 1, & \hat{o}_j(y) = o_j^\star, \\ -0.5, & \hat{o}_j(y) = \mathrm{E}, \\ -1, & \text{otherwise.} \end{cases} \tag{1}$$

The bank and judge are shared across training roles. This reward tests represented facts, not unsupported claims outside the bank. Appendix A.1 formalizes the bank and illustrates its construction.

We write $r^0 = R_G(c_0)$, $r_g^E = R_G(\mathrm{ser}(E_g))$, and $r_g^F = R_G(c_g^F)$ for initial-caption, accumulated-evidence, and final-caption rewards. Here ser deterministically concatenates the initial caption and completed pairs in interaction order, with role delimiters. Questions are marked non-assertive; only the initial caption and answers supply factual evidence.

### 3.3 Action-Aligned Credit with LOOP-GRPO

Figure 2(b) summarizes LOOP-GRPO, which distinguishes question ($Q$), stopping ($S$), and final-caption ($F$) actions instead of sharing one completion advantage (Shao et al., 2024).

**Question credit.** Let $E_g^{-t}$ be $E_g$ with its $t$-th question–answer pair removed, keeping all other pairs fixed, and let $r_{g,-t}^E = R_G(\mathrm{ser}(E_g^{-t}))$ be its reward. The **L**eave-**O**ne-**O**ut (LOO) utility assigned to question $t$ is

$$u_{g,t}^Q = r_g^E - r_{g,-t}^E. \tag{2}$$

This utility is denoted $\Delta_{g,t}^Q$ in Figure 2(b). Unlike adjacent reward differences, each turn is evaluated against the same completed evidence state without regenerating later observations (Foerster et al., 2018). Serialization and judge behavior can still introduce order sensitivity. Appendix A.3 gives the sequential alternative and derives the cancellation of a common question-cost offset.

**Stopping and final synthesis.** Both actions use an aggregate quality utility $b_g$:

$$b_g = (r_g^F - r^0) + \lambda\big[\min(r_g^E, r_g^F) - r^0\big], \tag{3}$$

where $\lambda \geq 0$ weights the bottleneck term ($\lambda = 1$ in our runs). The first term rewards final-caption improvement; the second rewards jointly high evidence and caption scores. It encourages preservation but does not guarantee overlap of individual correct facts. The raw stop and final-caption utilities are $u_g^S = b_g - c_S T_g$ and $u_g^F = b_g$, where $c_S \geq 0$ is an optional per-query stopping cost. We use $c_S = 0$: stopping is trained for evidence sufficiency, not by a length penalty. Only a sampled READY receives stop credit; forced finalization trains questions and synthesis but has no stop span.

**Type-normalized, span-aligned updates.** For each action type $\tau \in \{Q, S, F\}$, collect its valid actions from the same audio's $G$ rollouts. An action instance $\xi$ with raw utility $u(\xi)$ receives advantage $A(\xi) = (u(\xi) - \mu_\tau)/(\sigma_\tau + \varepsilon_{\mathrm{norm}})$, where $\mu_\tau$ and $\sigma_\tau$ are the within-type mean and standard deviation, and $\varepsilon_{\mathrm{norm}} > 0$ stabilizes normalization. The advantage is applied only to that action's Composer-token span in the clipped GRPO objective. Instruct answers, prompts, scheduler tokens, and padding are masked. Appendix A.4 supplies the token mapping, policy ratio, clipping, and **K**ullback–**L**eibler (KL) regularization.

### 3.4 Role-Wise Warm-Up and Alternating Training

Figure 2(c) outlines the schedule. Role-wise SFT and GRPO warm-up prepare each interface before alternating optimization. Captioner GRPO rewards the initial caption with $R_G(c_0)$. Instruct uses a *per-QA delta reward*. For a fixed audio/history/question tuple $(x, h_t, q_t)$, where $h_t$ is the pre-answer text history, we sample $G$ answers, indexed here by $g$. Let $y_t^-$ denote the serialized pre-answer evidence and $y_{g,t}^+$ the same text with answer $a_{g,t}$ appended. Its utility is

$$u_{g,t}^I = R_G(y_{g,t}^+) - R_G(y_t^-). \tag{4}$$

This compares alternative answers under the same fixed prompt, updating only Instruct's answer tokens; it is distinct from the Composer's completed-trajectory question credit. Composer warm-up

Table 1: Caption-as-evidence accuracy (%) on four audio-only benchmarks. Best and second-best reported scores within the open-weight-backbone group are bold and underlined. †: Original-paper results under a different evaluation protocol; 7B weights unreleased (Appendix C). A dash denotes an unreported result.

| Method | MMAU | MMAR | MMSU | Omni-Cloze |
|---|---|---|---|---|
| *Proprietary models* | | | | |
| GPT-4o Audio | 62.4 | 59.3 | 56.4 | 35.8 |
| Gemini2.0-Flash | 58.6 | 50.6 | 51.0 | 20.0 |
| Gemini2.5-Flash | 65.6 | 58.2 | 58.1 | 42.6 |
| Gemini2.5-Pro | 70.0 | 64.1 | 71.8 | 48.0 |
| Gemini3.1-Pro | 71.3 | 65.7 | 69.5 | 64.1 |
| Qwen3.5-Omni-Plus | 71.6 | 62.6 | 69.8 | 57.6 |
| *Models with open-weight backbones* | | | | |
| Audio-Flamingo 3-7B | 47.1 | 36.7 | 42.5 | 6.3 |
| Kimi-Audio-7B | 54.5 | 39.8 | 43.4 | 2.9 |
| Step-Audio-2-mini-8B | 62.5 | 44.6 | 53.6 | 5.7 |
| MiDashengLM-7B | 67.6 | 51.7 | 62.2 | 19.5 |
| Qwen2-Audio-7B | 63.3 | 44.2 | 53.3 | 22.2 |
| Qwen2.5-Omni-7B | 65.2 | 51.8 | 60.6 | 25.8 |
| Audio-Captioner-7B† | <u>70.0</u> | <u>59.8</u> | – | 53.2 |
| Qwen3-Omni-30B-A3B-Instruct | 67.9 | 54.5 | 67.4 | 35.3 |
| Qwen3-Omni-30B-A3B-Captioner | 66.4 | 55.3 | <u>68.2</u> | <u>57.5</u> |
| **ACE-Cap (ours)** | **73.0** | **64.1** | **70.4** | **64.4** |

uses LOOP-GRPO, with separate **Lo**w-**R**ank **A**daptation (LoRA) adapters for its Decision (questioning/stopping) and Final (synthesis) branches.

During alternating training, the Captioner stays frozen. A fixed Composer harvests $N$ audio/history/question prompts for Instruct updates using Eq. (4), where $N$ is the harvested set size; a fixed Instruct supplies answers for Composer updates. Only one policy changes per phase. All roles are frozen at inference, without gold captions, reward MCQs, or the judge. Algorithm 1 and the role-specific objectives in Appendix A specify the complete training procedure.

# 4 Experiments

## 4.1 Experimental Setup

**Training Data and Implementation Details.** ACE-Cap is trained on 50,000 audio clips sampled from ASID-1M Li et al. (2026). We use Gemini 3.1 Pro-Preview[1] (hereafter Gemini 3.1 Pro) to generate gold captions for all clips and Gold-MCQs from a 3,000-pair subset. The Captioner and Instruct models are initialized from Qwen2.5-Omni-7B Xu et al. (2025a), while the Composer is initialized from Qwen3-8B Yang et al. (2025). Further data-construction and implementation details are provided in Appendix B, with the complete annotation prompts in Appendix E.

**Baselines and Evaluation Benchmarks.** Following the caption-as-evidence formulation of Omni-Captioner (Ma et al., 2026b), our common-protocol comparisons evaluate proprietary and open-weight caption generators on Omni-Cloze Ma et al. (2026b), MMAR Ma et al. (2026a), MMAU Sakshi et al. (2025), and MMSU Wang et al. (2025), using a frozen Qwen3.6-27B Qwen Team (2026) text-only judge. Table 1 lists the systems; Appendix C gives their references, benchmark descriptions, and the fixed prompting, decoding, and parsing protocol, including the cited-result exception marked by †.

Accuracy is the percentage of benchmark questions answered correctly from the supplied text. We use $\mathrm{Acc}(E_g)$ and $\mathrm{Acc}(c_g^F)$ as dataset-level shorthand for judging serialized accumulated evidence

[1]Official model card: `https://deepmind.google/models/model-cards/gemini-3-1-pro/`.

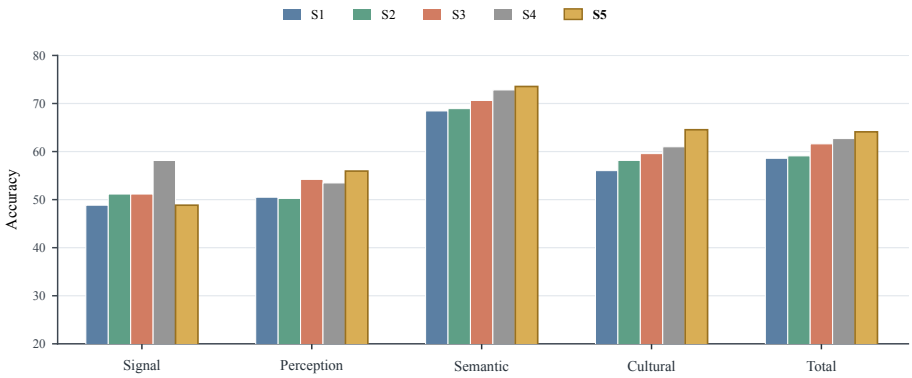



Figure 3: MMAR accuracy across training stages: S1, all-SFT pipeline; S2, Captioner GRPO; S3, Instruct GRPO; S4, Composer LOOP-GRPO warm-up; S5, alternating co-evolution. Each stage builds on the preceding one. The four capability layers and total accuracy are shown; the y-axis begins at 20%.

and final captions, respectively, not for a single rollout's score. Their aggregate retention gap is $\Delta_{\text{ret}} = \text{Acc}(E_g) - \text{Acc}(c_g^F)$, measured in percentage points. It is distinct from the training reward $R_G$ and does not directly count individual retained facts.

### 4.2 Main Comparison

**Caption utility on audio question answering.** ACE-Cap achieves 73.0%, 64.1%, and 70.4% on MMAU, MMAR, and MMSU (Table 1), exceeding the strongest open-weight baseline under the common evaluation protocol by 5.1, 8.8, and 2.2 percentage points, respectively. Across 18 proprietary-model comparisons, it records 15 higher scores, one tie, and two lower scores. It exceeds GPT-4o Audio and Qwen3.5-Omni-Plus on all three benchmarks; Gemini 2.5 Pro leads it on MMSU and Gemini 3.1 Pro on MMAR. These results support strong caption utility across varied audio reasoning tasks without implying uniform dominance.

**Fine-grained evidence coverage on Omni-Cloze.** Table 1 reports the corresponding audio-only Omni-Cloze comparison. ACE-Cap achieves 64.4%, the highest accuracy under the common protocol, exceeding Qwen3-Omni-30B-A3B-Captioner by 6.9 points. It also exceeds all proprietary baselines, including GPT-4o Audio, Gemini 2.5 Pro, Qwen3.5-Omni-Plus, and Gemini 3.1 Pro, by 28.6, 16.4, 6.8, and 0.3 points, respectively. The last margin is modest: this is the best point estimate, not a large separation. The results indicate utility for resolving masked acoustic details under the caption-only protocol.

### 4.3 Training-Stage Analysis

Figure 3 traces the ordered stages of Algorithm 1. Overall MMAR accuracy increases from 58.6% at S1 to 59.1%, 61.6%, 62.7%, and 64.1% at S2–S5. The respective gains are 0.5 points from Captioner GRPO, 2.5 from Instruct GRPO, 1.1 from Composer warm-up, and 1.4 from alternation. The largest stage gain follows Instruct training, consistent with the importance of answer quality to the evidence available for synthesis. Composer optimization then improves how that evidence is requested and used.

These are cumulative pipeline comparisons, not isolated component-removal effects: S1 already uses the all-SFT interactive pipeline, and S2 changes its Captioner rather than evaluating the Captioner alone. Perception, Semantic, and Cultural accuracy improve by 5.44, 5.09, and 8.51 points from S1 to S5. Signal peaks at S4 and returns to its S1 value at S5, so improvement is not uniform across capability layers; this subset contains only 43 examples.

The modality breakdown in Appendix D.1 further shows gains of 11.17 points for music and 6.06 for sound, versus 1.02 for speech. Sound–speech rises from 69.27% to 74.77% and music–speech

from 69.51% to 71.95%. This pattern indicates that the overall gain extends beyond speech-only understanding. The smallest mixed-modality subsets remain variable and are reported separately rather than generalized from.

### 4.4 Captioner and Credit-Assignment Ablations

Table 2: Captioner and credit-assignment comparisons on Omni-Cloze and MMAU (accuracy, %). (a) Initial captions $c_0$ versus final captions $c_g^F$. (b) Alternative credit-assignment schemes.

**(a) One-stage versus full system**

| Representation | Omni-Cloze | MMAU |
|---|---|---|
| SFT Captioner $c_0$ | 57.1 | 68.5 |
| GRPO Captioner $c_0$ | 62.3 | 70.8 |
| Final caption $c_g^F$ | **64.4** | **73.0** |

**(b) Credit-assignment schemes**

| Scheme | Omni-Cloze | MMAU |
|---|---|---|
| Trajectory-level GRPO | 63.0 | 71.3 |
| Adjacent difference | 63.4 | 70.9 |
| LOOP-GRPO (ours) | **64.4** | **73.0** |

Table 2(a) shows that GRPO improves initial-caption accuracy by 5.2 and 2.3 percentage points on Omni-Cloze and MMAU, respectively. ACE-Cap gains a further 2.1 and 2.2 points over GRPO captions, supporting the utility of interaction beyond a stronger one-stage captioner. In panel (b), LOOP-GRPO exceeds trajectory-level GRPO by 1.4 and 1.7 points, and adjacent difference by 1.0 and 2.1 points, respectively. Adjacent difference does not consistently outperform trajectory-level GRPO, whereas LOOP-GRPO achieves the highest accuracy on both benchmarks.

Table 3: Credit ablations and cross-Judge evaluation (accuracy, %). (a) Question ($Q$) and stopping ($S$) credit with the Final LoRA fixed; Omni denotes Omni-Cloze. (b) Final-caption ($F$) credit with fixed Decision trajectories; $\Delta_{\mathrm{ret}}$ is in percentage points. (c) The same frozen MMAR captions evaluated by three text-only Judges.

**(a) Decision credit**

| **Variant** | $Q$ | $S$ | **Omni** | **MMAR** |
|---|---|---|---|---|
| Full Decision | ✓ | ✓ | 64.4 | 64.1 |
| w/o $Q$ credit | – | ✓ | 63.2 | 62.6 |
| w/o $S$ credit | ✓ | – | 64.0 | 63.6 |

**(b) Final credit (Omni-Cloze)**

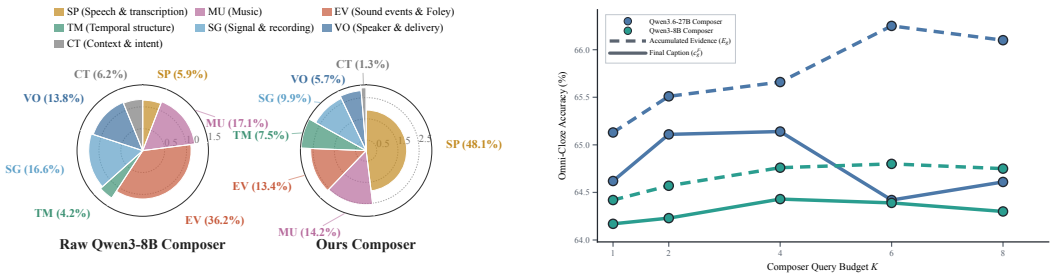
| **Variant** | Acc. $(E_g)$ | Acc. $(c_g^F)$ | $\Delta_{\mathrm{ret}}$ |
|---|---|---|---|
| Full Final | 64.8 | 64.4 | 0.4 |
| w/o $F$ credit | 64.8 | 63.4 | 1.4 |

**(c) Cross-Judge (MMAR)**

| **Judge** | **Signal** | **Perception** | **Semantic** | **Cultural** | **Total** |
|---|---|---|---|---|---|
| Qwen3.5-27B | 51.16 | 56.68 | 72.09 | 65.96 | 64.10 |
| Qwen3.6-27B | 48.84 | 55.94 | 73.54 | 64.54 | 64.10 |
| Gemini 3.1 Pro | 55.81 | 61.39 | 82.04 | 69.50 | 70.80 |

**Question and stopping credits.** Table 3(a) holds the Final LoRA fixed and removes one Decision credit at a time. Here $Q$ credit supervises the question's leave-one-out evidence contribution, and $S$ credit supervises the sampled stopping action; neither ablation removes the action itself. Without question credit, accuracy falls from 64.4% to 63.2% on Omni-Cloze and from 64.1% to 62.6% on MMAR. Without stopping credit, the corresponding scores are 64.0% and 63.6%. Direct question credit therefore has the larger effect in these comparisons, while stopping supervision also contributes.

**Final-caption credit with fixed evidence.** Table 3(b) keeps the Decision LoRA and its trajectories fixed, and compares the GRPO-trained Final LoRA with its SFT initialization. Both variants receive identical accumulated evidence with 64.8% accuracy on Omni-Cloze. Final-caption credit raises caption accuracy from 63.4% to 64.4%, reducing the aggregate gap from 1.4 to 0.4 points. This controlled comparison separates improved synthesis from the acquisition of different evidence. The approximately 71% reduction concerns the aggregate accuracy gap, not a measured fraction of individual facts retained.




**(a) Composer query behavior** **(b) Query-budget ablation**

Figure 4: Composer analysis on Omni-Cloze. (a) Query-category behavior: sector angle denotes the proportion of queries assigned to each category, while radius denotes the category-specific mean query turn. (b) Accuracy versus the maximum Composer query budget $K$. Dashed and solid lines evaluate the Accumulated Evidence $E_g$ and final caption $c_g^F$, respectively. Blue denotes the frozen base Qwen3.6-27B Composer, which was not trained with LOOP-GRPO or alternating co-evolution. Teal denotes our Qwen3-8B Composer, which was trained with both.

### 4.5 Query Behavior and Budget

**What the Composer asks.** Figure 4(a) summarizes query categories and their positions in the interaction. The Raw Composer favors sound-event queries (EV, 36.2%) and asks most categories near the first turn. Ours allocates 48.1% of queries to speech and transcription (SP), while placing temporal, speaker-related, and contextual questions later. These category summaries describe a shift toward more structured interactions; the credit ablations above separately test whether training signals affect downstream utility.

**More evidence does not always yield a better final caption.** Figure 4(b) varies maximum budget $K$. For the frozen Qwen3.6-27B Composer, increasing $K$ from 4 to 6 raises accumulated-evidence accuracy from 65.66% to 66.25%, yet lowers final-caption accuracy from 65.14% to 64.42%. Its gap consequently grows from 0.52 to 1.83 percentage points (pp), and remains 1.49 pp at $K = 8$. Thus, higher question-answer utility can be lost during synthesis of the longer history.

Our trained Qwen3-8B Composer maintains a gap of 0.25–0.45 pp across all tested budgets. Its final-caption accuracy peaks at 64.43% for $K = 4$ and is 64.39% and 64.30% at $K = 6$ and 8. Larger allowed budgets therefore do not improve final utility in this sweep. The policies differ in both backbone and training, so the comparison diagnoses budget-scaling behavior rather than isolating a LOOP-GRPO effect. Complete values are in Appendix D.2.

### 4.6 Evaluator Sensitivity

The same frozen MMAR captions achieve 64.10% under both Qwen3.5-27B and Qwen3.6-27B, and 70.80% under Gemini 3.1 Pro (Table 3(c)). All three evaluators rank the capability layers in the same order: Semantic, Cultural, Perception, then Signal. The stable profile supports the category-level interpretation, while the 6.70-point difference shows that absolute accuracy depends on the evaluator. This experiment evaluates one caption set across judges; it does not by itself establish cross-method ranking robustness. Further discussion is given in Appendix D.4.

## 5 Conclusion

ACE-Cap formulates fine-grained audio captioning as iterative evidence acquisition. A text-only Composer queries an audio-conditioned Instruct model and synthesizes the answers; LOOP-GRPO, role-wise warm-up, and alternating optimization train questioning, stopping, and synthesis. The benchmark results support active interaction for detailed audio understanding. The study is limited to audio-only inputs and caption-as-evidence evaluation; future work will consider audio-visual evidence, broader evaluation, and adaptive interaction budgets.

## STATEMENT OF AI USE

The research pipeline uses Gemini 3.1 Pro to construct synthetic caption and question supervision and frozen language models to evaluate textual evidence. These uses are detailed in the methodology and Appendix B. AI tools also assisted with LaTeX formatting, language editing, and checking mathematical explanations during manuscript preparation. This assistance did not generate experimental measurements.

## REPRODUCIBILITY STATEMENT

Appendix B specifies the training data, role-wise supervision, trainable modules, optimization settings, and alternating schedule. Appendix C describes the caption-as-evidence evaluation protocol, and Appendix E provides the core annotation prompts. Appendix A contains the complete training algorithm and token-masked objective.

## REFERENCES

Shunian Chen, Xinyuan Xie, Zheshu Chen, Liyan Zhao, Owen Lee, Zhan Su, Qilin Sun, and Benyou Wang. Fusionaudio-1.2 m: Towards fine-grained audio captioning with multimodal contextual fusion. *arXiv preprint arXiv:2506.01111*, 2025.

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.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

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.

Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, et al. Kimi-audio technical report. *arXiv preprint arXiv:2504.18425*, 2025.

Heinrich Dinkel, Gang Li, Jizhong Liu, Jian Luan, Yadong Niu, Xingwei Sun, Tianzi Wang, Qiyang Xiao, Junbo Zhang, and Jiahao Zhou. Midashenglm: Efficient audio understanding with general audio captions. *arXiv preprint arXiv:2508.03983*, 2025.

Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. Clotho: An audio captioning dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 736–740. IEEE, 2020.

Jakob Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. Counterfactual multi-agent policy gradients. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Sreyan Ghosh, Zhifeng Kong, Sonal Kumar, S Sakshi, Jaehyeon Kim, Wei Ping, Rafael Valle, Dinesh Manocha, and Bryan Catanzaro. Audio flamingo 2: An audio-language model with long-audio understanding and expert reasoning abilities. *arXiv preprint arXiv:2503.03983*, 2025.

Arushi Goel, Sreyan Ghosh, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang-gil Lee, Chao-Han Huck Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, et al. Audio flamingo 3: Advancing audio intelligence with fully open large audio language models. *arXiv preprint arXiv:2507.08128*, 2025.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. Audiocaps: Generating captions for audios in the wild. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 119–132, 2019.

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. *arXiv preprint arXiv:2402.01831*, 2024.

Yunheng Li, Hengrui Zhang, Meng-Hao Guo, Wenzhao Gao, Shaoyong Jia, Shaohui Jiao, Qibin Hou, and Ming-Ming Cheng. Towards universal video mllms with attribute-structured and quality-verified instructions. *arXiv preprint arXiv:2602.13013*, 2026.

Manh Luong, Gholamreza Haffari, Dinh Phung, and Lizhen Qu. Graph-score: A graph-grounded metric for audio captioning. In *Proceedings of the 23rd Annual Workshop of the Australasian Language Technology Association*, pp. 192–201, 2025.

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. *Advances in Neural Information Processing Systems*, 38, 2026a.

Ziyang Ma, Ruiyang Xu, Zhenghao Xing, Yunfei Chu, Yuxuan Wang, Jinzheng He, Jin Xu, Pheng-Ann Heng, Kai Yu, Junyang Lin, Eng Siong Chng, and Xie Chen. Omni-captioner: Data pipeline, models, and benchmark for omni detailed perception. In *The Fourteenth International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=Z091XLyVkJ.

Xinhao Mei, Xubo Liu, Qiushi Huang, Mark D Plumbley, and Wenwu Wang. Audio captioning transformer. *arXiv preprint arXiv:2107.09817*, 2021.

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.

Hadi Nekoei, Akilesh Badrinaaraayanan, Amit Sinha, Mohammad Amini, Janarthanan Rajendran, Aditya Mahajan, and Sarath Chandar. Dealing with non-stationarity in decentralized cooperative multi-agent deep reinforcement learning via multi-timescale learning. In *Conference on Lifelong Learning Agents*, pp. 376–398. PMLR, 2023.

Georgios Papoudakis, Filippos Christianos, Arrasy Rahman, and Stefano V Albrecht. Dealing with non-stationarity in multi-agent deep reinforcement learning. *arXiv preprint arXiv:1906.04737*, 2019.

Qwen Team. Qwen3.6-27B: Flagship-level coding in a 27B dense model, April 2026. URL https://qwen.ai/blog?id=qwen3.6-27b.

Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7008–7024, 2017.

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.

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.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

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

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. *arXiv preprint arXiv:2506.04779*, 2025.

Qian Wang, Jia-Chen Gu, and Zhen-Hua Ling. X-ace: Explainable and multi-factor audio captioning evaluation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 12273–12287, 2024.

Boyong Wu, Chao Yan, Chen Hu, Cheng Yi, Chengli Feng, Fei Tian, Feiyu Shen, Gang Yu, Haoyang Zhang, Jingbei Li, et al. Step-audio 2 technical report. *arXiv preprint arXiv:2507.16632*, 2025a.

Tsung-Han Wu, Joseph E Gonzalez, Trevor Darrell, and David M Chan. Clair a: Leveraging large language models to judge audio captions. In *2025 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 1–7. IEEE, 2025b.

Zhenghao Xing, Ruiyang Xu, Yuxuan Wang, Jinzheng He, Ziyang Ma, Qize Yang, Yunfei Chu, Jin Xu, Junyang Lin, Chi-Wing Fu, et al. Native active perception as reasoning for omni-modal understanding. *arXiv preprint arXiv:2606.19341*, 2026.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, et al. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*, 2025a.

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*, 2025b.

Xuenan Xu, Heinrich Dinkel, Mengyue Wu, and Kai Yu. A crnn-gru based reinforcement learning approach to audio captioning. In *DCASE*, pp. 225–229, 2020.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. Mm-react: Prompting chatgpt for multimodal reasoning and action. *arXiv preprint arXiv:2303.11381*, 2023.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022.

Siliang Zeng, Quan Wei, William Brown, Oana Frunza, Yuriy Nevmyvaka, Yang Katie Zhao, and Mingyi Hong. Reinforcing multi-turn reasoning in llm agents via turn-level credit assignment. In *ICML 2025 Workshop on Computer Use Agents*, 2025.

Zelin Zhou, Zhiling Zhang, Xuenan Xu, Zeyu Xie, Mengyue Wu, and Kenny Q Zhu. Can audio captions be evaluated with image caption metrics? In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 981–985. IEEE, 2022.

# A Further Methodological Details

## A.1 Reward Bank and Policy Notation

Let $\mathcal{D} = \{(x_n, c_n^\star)\}_{n=1}^{|\mathcal{D}|}$ be the training set, where $n$ indexes an audio–gold-caption pair and $|\mathcal{D}|$ is its size. We omit $n$ for one example. The policies are Captioner $f_\phi$, Composer $\pi_\theta$, and Instruct $p_\psi$, with trainable parameters $\phi$, $\theta$, and $\psi$, respectively; the initial caption is $c_0 = f_\phi(x)$. For that example, the retained MCQ bank is

$$\mathcal{M}^\star(x, c^\star) = \{(m_j, o_j^\star)\}_{j=1}^{L}, \tag{5}$$

where $L$ is the bank size, $m_j$ is a question with five choices, and $o_j^\star$ is its correct option among A–D. The frozen judge $\mathcal{J}$ produces $\hat{o}_j(y) = \mathcal{J}(y, m_j)$ for text $y$; E is *Not Given*. Equation (1) maps these predictions to a common reward scale. The judge sees the candidate text and question choices, but not audio, the gold caption, or the gold option.

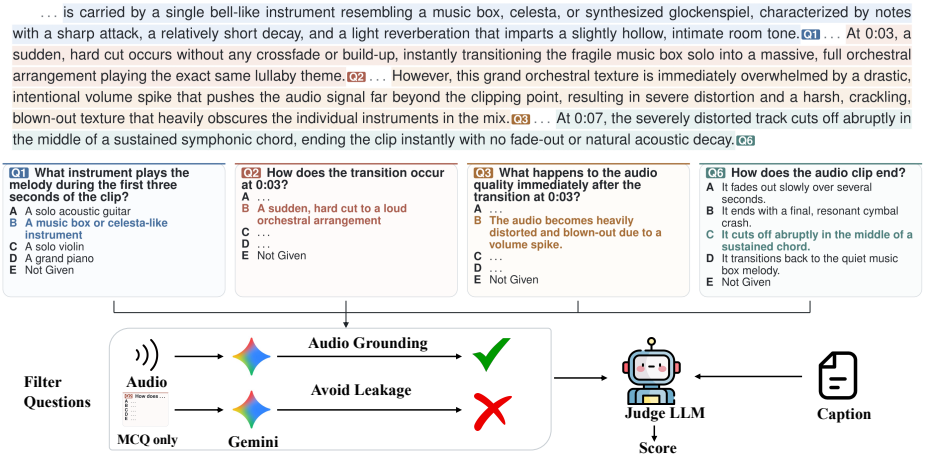



Figure 5: Gold-MCQ construction and G2P scoring. Colors link fixed MCQs to gold-caption evidence spans. Candidate questions are filtered offline; during GRPO, the frozen judge scores only candidate text against the retained questions, without audio or the gold caption.

## A.2 Complete Rollout and Evidence Definitions

The rollout index is $g \in \{1, \ldots, G\}$, where $G$ is the number of sampled interactions for one audio. At potential question turn $t$, let $s$ index earlier turns. The text history and sampled action $z_{g,t}$ are

$$h_{g,t} = \left(c_0, (q_{g,s}, a_{g,s})_{s=1}^{t-1}\right), \tag{6}$$

$$z_{g,t} \sim \pi_\theta(\cdot \mid h_{g,t}), \qquad z_{g,t} \in \{\mathrm{ASK}(q_{g,t}), \mathrm{READY}\}. \tag{7}$$

Here $q_{g,t}$ is a question and $a_{g,t}$ its answer. For an ASK action, Instruct samples

$$a_{g,t} \sim p_\psi(\cdot \mid x, h_{g,t}, q_{g,t}). \tag{8}$$

This answer is an environment observation, not a Composer action. If READY occurs before question $t$, the realized query count is $T_g = t-1$. Otherwise the scheduler finalizes at the maximum budget $T_g = K$, without introducing a policy stop token.

The completed evidence and the evidence with pair $t$ removed are

$$E_g = \left(c_0, (q_{g,t}, a_{g,t})_{t=1}^{T_g}\right), \tag{9}$$

$$E_g^{-t} = \left(c_0, (q_{g,s}, a_{g,s})_{1 \le s \le T_g,\, s \ne t}\right). \tag{10}$$

Removal does not regenerate later questions or answers. The final caption is sampled under the scheduler's FINALIZE instruction:

$$c_g^F \sim \pi_\theta(\cdot \mid E_g, \text{FINALIZE}). \tag{11}$$

With ser denoting deterministic, role-delimited evidence serialization, the three rewards are

$$r^0 = R_G(c_0), \qquad r_g^E = R_G(\text{ser}(E_g)), \qquad r_g^F = R_G(c_g^F). \tag{12}$$

Questions are marked non-assertive in the serialization; only $c_0$ and answers count as factual evidence. The leave-one-out score is

$$r_{g,-t}^E = R_G(\text{ser}(E_g^{-t})). \tag{13}$$

### A.3 Equivalent Utility Expression and Its Scope

For comparison with completed-evidence credit, let $E_{g,t}$ be the evidence prefix after $t$ pairs, with $E_{g,0} = (c_0)$. The adjacent reward difference $d_{g,t}$ is

$$d_{g,t} = R_G(\text{ser}(E_{g,t})) - R_G(\text{ser}(E_{g,t-1})). \tag{14}$$

It depends on which facts have already appeared and can favor early coarse questions under a saturating reward. In contrast, Eq. (2) removes each pair from the same completed state; it is not a counterfactual rollout under a different policy.

For the question actions in a same-audio group, write $v_{g,t} = r_g^E - r_{g,-t}^E$ and let $v$ denote their collection. Let $c_Q$ be a shared raw question-cost offset. With mean and std taken over those actions and $\varepsilon_{\text{norm}} > 0$ the stabilizer,

$$\frac{(v_{g,t} - c_Q) - \text{mean}(v - c_Q)}{\text{std}(v - c_Q) + \varepsilon_{\text{norm}}} = \frac{v_{g,t} - \text{mean}(v)}{\text{std}(v) + \varepsilon_{\text{norm}}}. \tag{15}$$

Thus, removing a common offset changes neither normalized advantages nor the policy objective. It is not an effective question-count penalty after centering. The recorded raw offset remains $c_Q = 0.02$ and the reported stopping coefficient remains $c_S = 0$; the algebraic rewrite does not change the experiment configuration.

The $\min(r_g^E, r_g^F)$ term in Eq. (3) is an aggregate bottleneck, not an intersection of item-level correct facts. Likewise, an aggregate evidence-to-caption accuracy gap does not directly measure same-item factual retention.

### A.4 Span Mapping and Clipped Token-Masked Objective

For action type $\tau \in \{Q, S, F\}$, let $\Xi_\tau$ collect its valid instances from the same audio's $G$ rollouts. The type-wise normalization in Section 3.3 uses the utilities in $\Xi_\tau$ to compute $\mu_\tau$ and $\sigma_\tau$. We abbreviate the resulting per-action advantages as $A_{g,t}^Q$ for a question, $A_g^S$ for a sampled stop, and $A_g^F$ for the final caption.

Let $k$ index tokens in rollout $g$. The sets $\mathcal{I}_{g,t}^Q$, $\mathcal{I}_g^S$, and $\mathcal{I}_g^F$ contain the Composer-token positions for those three actions. Their token-aligned advantage is

$$A_{g,k}^{\text{tok}} = \begin{cases} A_{g,t}^Q, & k \in \mathcal{I}_{g,t}^Q, \\ A_g^S, & k \in \mathcal{I}_g^S, \\ A_g^F, & k \in \mathcal{I}_g^F, \\ 0, & \text{otherwise.} \end{cases} \tag{16}$$

For forced finalization, $\mathcal{I}_g^S = \varnothing$. The binary mask $M_{g,k} \in \{0, 1\}$ selects tokens of the Composer branch being optimized; it excludes partner answers, prompts, user messages, scheduler tokens, and padding. In the sums below, $g$ ranges over all $G$ rollouts and $k$ over all token positions in each rollout.

For token $y_{g,k}$ and preceding text context $\mathcal{H}_{g,k}$, let $\pi_{\theta_{\text{old}}}$ be the fixed policy that sampled the rollouts. The importance ratio and clipped ratio are

$$\rho_{g,k}(\theta) = \frac{\pi_\theta(y_{g,k} \mid \mathcal{H}_{g,k})}{\pi_{\theta_{\text{old}}}(y_{g,k} \mid \mathcal{H}_{g,k})}, \tag{17}$$

$$\bar{\rho}_{g,k}(\theta) = \text{clip}\left(\rho_{g,k}(\theta), 1 - \varepsilon_{\text{clip}}, 1 + \varepsilon_{\text{clip}}\right), \tag{18}$$

where $\varepsilon_{\text{clip}} > 0$ is the policy-ratio clipping threshold, distinct from the normalization stabilizer. LOOP-GRPO minimizes the clipped loss $\mathcal{L}_{\text{LOOP}}$:

$$\mathcal{L}_{\text{LOOP}}(\theta) = -\frac{\sum_{g,k} M_{g,k} \min\left(\rho_{g,k} A^{\text{tok}}_{g,k}, \bar{\rho}_{g,k} A^{\text{tok}}_{g,k}\right)}{\sum_{g,k} M_{g,k}} + \beta \overline{D}_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}}). \tag{19}$$

Here $\overline{D}_{\text{KL}}$ is the masked token-average Kullback–Leibler (KL) penalty against a frozen reference policy $\pi_{\text{ref}}$, and $\beta = 0.07$ is its weight. This reference is distinct from the rollout policy in the importance ratio.

### A.5 Role-Specific Warm-Up and Training Algorithm

The Captioner utility $u^{\text{Cap}}$ is

$$u^{\text{Cap}} = R_G(c_0). \tag{20}$$

For Instruct warm-up, fix an audio/history/question tuple $(x, h_t, q_t)$ and sample $G$ answers, here indexed by $g$. Let $y_t^-$ be the pre-answer evidence and $y_{g,t}^+$ the same text with answer $a_{g,t}$ appended. Equation (4) defines the per-QA answer utility as the reward difference between these two texts. Only answer tokens are updated. Composer warm-up uses LOOP-GRPO with Instruct fixed. Subsequent alternation keeps the Captioner frozen and changes only one partner policy at a time.

**Algorithm 1** ACE-Cap Training

**Require:** Training set $\mathcal{D}$, per-example banks $\{\mathcal{M}_n^\star\}$, rollout size $G$, query budget $K$
1: SFT Captioner $f_\phi$, Instruct $p_\psi$, and Composer $\pi_\theta$
2: Warm up $f_\phi$ by GRPO using Eq. (20)
3: Warm up $p_\psi$ by per-QA GRPO using Eq. (4)
4: Warm up $\pi_\theta$ by LOOP-GRPO with fixed $p_\psi$
5: Freeze $f_\phi$ permanently
6: **for** each co-evolution cycle **do**
7: Freeze $\pi_\theta$; harvest prompts and update $p_\psi$
8: Freeze $p_\psi$; roll out $G$ trajectories and update $\pi_\theta$
9: **end for**
10: **return** $f_\phi$, $p_\psi$, and $\pi_\theta$

## B Training Data Construction and Implementation Details

ACE-Cap uses approximately 50,000 clips sampled from the ASID-1M Li et al. (2026) audio–visual corpus, with only the audio tracks retained. We use Gemini 3.1 Pro-Preview[2] (hereafter Gemini 3.1 Pro) to perform complementary analyses of signal condition, perceptual attributes, semantic events and speech, cultural and pragmatic context, music, and discrete sound effects. A separate synthesis prompt combines these observations into one long-form gold caption per clip. Each caption follows the audio chronologically, places timestamps only at major event boundaries, preserves intelligible speech or lyrics, and qualifies uncertain or inferred content.

We sample 3,000 audio–caption pairs for Gold-MCQ construction. Given only a gold caption, Gemini 3.1 Pro extracts structured evidence and generates five-option questions across the applicable acoustic taxonomy. Each question contains one gold answer among A–D, three closely matched hard negatives, and a fixed *Not Given* option E that is never correct. Questions are linked to supporting caption spans, taxonomy levels, and fine-grained subcategories. A deterministic validator rejects malformed outputs, invalid or duplicate choices, unsupported labels, missing evidence, and unverifiable questions. Failed annotations are regenerated, and only valid examples are retained.

[2]Official model card: `https://deepmind.google/models/model-cards/gemini-3-1-pro/`.

### B.1 ROLE-WISE SFT DATA

The confirmed role-wise SFT data are constructed from ASID-1M audio. Table 4 summarizes their policy-visible supervision and training scale.

### B.2 ROLE-WISE SUPERVISED FINE-TUNING

**Captioner.** We initialize the Captioner from Qwen2.5-Omni-7B Xu et al. (2025a) and train it for one epoch. All language-model parameters and the audio projector are updated, while the visual and audio encoders and the remaining aligner parameters are frozen. The language model and audio projector use learning rates of $2 \times 10^{-6}$ and $5 \times 10^{-6}$, respectively. Both rates follow a cosine schedule that decays to 10% of their initial values, with a 3% warm-up.

**Instruct.** The Instruct model uses the same Qwen2.5-Omni-7B initialization and trainable-module configuration, but is optimized on audio-conditioned evidence question answering. We train it for one epoch with the same language-model and audio-projector learning rates and cosine schedule as the Captioner, using a 5% warm-up.

**Composer.** The Qwen3-8B Yang et al. (2025) Composer is factorized into separate Decision and Final LoRA branches from the SFT stage onward. Decision SFT supervises only the ASK/READY outputs, whereas Final SFT supervises the final caption conditioned on the initial caption and completed question–answer pairs. Both use rank 32, scaling factor 64, dropout 0.05, and the same query, key, value, output, gate, up, and down projection target modules as their corresponding GRPO adapters. GRPO initializes each branch from its matching SFT weights without changing this adapter parameterization. Thus, SFT learns the two interfaces by token-level supervised likelihood without a reward, whereas GRPO retains the same interfaces and optimizes them with role-specific rewards.

Table 4: Role-wise SFT supervision and scale.

| Role | Supervision | Scale |
|---|---|---|
| Captioner | Audio → long-form gold caption | ∼50K pairs (∼853 h) |
| Instruct | Audio + query → detailed evidence answer | ∼170K QA pairs (∼50K audios) |
| Composer | Initial caption + QA history → decision or final-caption target | ∼1.5K shared audio seeds |

The Decision and Final Composer SFT branches use approximately 1.5K audio and initial-caption seeds, matching the Composer GRPO set. The SFT records provide only their respective supervised targets and contain no Gold-MCQ banks or reward baselines; those fields are introduced only for GRPO reward computation.

### B.3 ROLE-WISE GRPO DATA

Table 5 summarizes the static seed data used for role-wise GRPO. These counts describe the inputs available before online group sampling; sampled rollouts and runtime reward records are therefore not counted as dataset examples.

Table 5: Role-wise static GRPO seed data and scale. The Gold-MCQ banks and precomputed baselines are used only by the reward function and are hidden from the trainable policies.

| Role | GRPO Seed Unit | Scale |
|---|---|---|
| Captioner | Policy: audio prompt; reward: Gold-MCQ bank | ∼3.0K seeds; 24.8K MCQs |
| Instruct | Policy: $(x, h_t, q_t)$ QA prompt; reward: Gold-MCQ bank | 4.2K prompts (1.4K audios); 12.2K MCQs |
| Composer | Policy: initial caption; reward: baseline + Gold-MCQ bank | ∼1.5K seeds; 12.6K MCQs; $K = 4$ |

The Captioner set contains 3,000 audio-captioning seeds whose Gold-MCQ banks provide 24,811 reward questions. The Instruct set contains 4,226 unique $(x, h_t, q_t)$ states over 1,430 audios and spans four interaction rounds; its banks contain 12,169 unique questions. The Composer set contains 1,498 unique initial-caption environment seeds and 12,560 unique questions. Each Composer seed also stores a precomputed caption-only reward baseline, and all use the fixed interaction budget $K = 4$. Gold-MCQ coverage is complete for all three datasets. The Captioner sees only the audio prompt, the Instruct model sees the audio with the text history and current query, and the Composer initially sees only the caption. No trainable policy receives the reward questions, answer choices, or gold labels. The complete prompts used for gold-caption and Gold-MCQ construction are reported in Sec. E.

### B.4 Role-Wise GRPO Training

All role-wise GRPO rewards are computed from fixed Gold-MCQ banks using the frozen Qwen3.6-27B Qwen Team (2026) caption-only judge.

**Captioner.** We initialize both the Captioner policy and its frozen reference model from the preceding Captioner SFT checkpoint. As in SFT, all language-model parameters and the audio projector are updated, while the visual and audio encoders and the remaining aligner parameters are frozen. We train the policy for one epoch with learning rates of $3 \times 10^{-6}$ and $1.5 \times 10^{-5}$ for the language model and audio projector, respectively. Both rates follow a cosine schedule that decays to 10% of their initial values, with a 3% warm-up. For each input, we sample $G = 8$ captions using a temperature of 0.7, top-$p$ of 0.9, top-$k$ of 50, and a maximum completion length of 768 tokens. The token-level loss is normalized by the active-token count in the local batch (`loss_type=bnpo`), with batch-level reward scaling and a KL coefficient of $\beta = 0.07$.

**Instruct.** For each per-QA Instruct update, the policy and frozen reference model are initialized from the same incoming Instruct checkpoint, while the Composer and G2P judge remain fixed. We update all language-model parameters and the audio projector, while freezing the visual and audio encoders and the remaining aligner parameters. Each update runs for one epoch with learning rates of $1 \times 10^{-6}$ and $5 \times 10^{-6}$ for the language model and audio projector, respectively. Both rates follow a cosine schedule that decays to 10% of their initial values, with a 3% warm-up. For each replayed QA prompt, we sample $G = 8$ answers using a temperature of 0.7, top-$p$ of 0.9, top-$k$ of 50, and a maximum completion length of 512 tokens. Candidate answers are optimized using the per-QA G2P improvement defined in the main paper. We use the same token-normalized loss with batch-level reward scaling and a KL coefficient of $\beta = 0.07$.

**Composer.** The Decision and Final GRPO adapters are initialized from their corresponding SFT branches and retain the same LoRA configuration described above. Each is optimized for one epoch with a learning rate of $1 \times 10^{-6}$, a cosine schedule decaying to 10% of the initial rate, 3% warm-up, the same token-normalized loss, no additional reward rescaling, and $\beta = 0.07$. We sample $G = 8$ outputs with temperature 0.7, top-$p$ 0.9, top-$k$ 50, and a maximum completion length of 2,048 tokens.

*Decision LoRA.* The Decision GRPO adapter is initialized from the Decision SFT LoRA and samples only Ask and Ready actions. The Instruct model is frozen and its answers are treated as loss-masked environment observations. The interaction budget is $K = 4$; reaching the budget forces finalization without creating a policy Ready span. The raw question utility includes an offset $c_Q = 0.02$, which cancels under action-type centering and is omitted from the equivalent main-text expression. There is no additional stopping cost ($c_S = 0$). Malformed actions receive a penalty of 1.0. A custom multi-turn trainer assigns advantages only to the Composer-generated decision spans, excluding Instruct answers, scheduler messages, and padding.

*Final LoRA.* The Final GRPO adapter is initialized from the corresponding Final SFT weights and optimizes only final-caption tokens using the evidence-preservation utility defined in the main paper, with $\lambda = 1$. Invalid final outputs receive an additional penalty of 1.0.

**Alternating co-evolution.** After the role-wise warm-ups, co-evolution reuses the 1,498 Composer GRPO seeds and alternates the trainable role every 50 optimization steps. During an Instruct phase,

the Composer is frozen and Instruct is updated from harvested audio-conditioned QA prompts. During the following Composer phase, Instruct is frozen and supplies environment answers while the Composer is updated. No phase updates both policies jointly, matching Algorithm 1.

Table 6: Omni-Cloze query-budget results for (a) the Base Composer and (b) Ours. Accuracy is in %, and the aggregate gap is $\Delta_{\text{ret}} = \text{Acc}(E_g) - \text{Acc}(c_g^F)$ in percentage points (pp). Bold indicates the highest accuracy within each representation and Composer.

**(a) Base Composer**

| $K$ | $\text{Acc}(E_g)$ | $\text{Acc}(c_g^F)$ | $\Delta_{\text{ret}}$ (pp) |
|---|---|---|---|
| 1 | 65.13 | 64.62 | 0.51 |
| 2 | 65.51 | 65.11 | 0.40 |
| 4 | 65.66 | **65.14** | 0.52 |
| 6 | **66.25** | 64.42 | 1.83 |
| 8 | 66.10 | 64.61 | 1.49 |

**(b) Ours Composer**

| $K$ | $\text{Acc}(E_g)$ | $\text{Acc}(c_g^F)$ | $\Delta_{\text{ret}}$ (pp) |
|---|---|---|---|
| 1 | 64.42 | 64.17 | 0.25 |
| 2 | 64.57 | 64.23 | 0.34 |
| 4 | 64.76 | **64.43** | 0.33 |
| 6 | **64.80** | 64.39 | 0.41 |
| 8 | 64.75 | 64.30 | 0.45 |

## C Baselines and Evaluation Benchmarks

**Baselines.** Following prior detailed-captioning evaluations Ma et al. (2026b), we compare with proprietary audio-capable models, including GPT-4o Audio Hurst et al. (2024), Gemini 2.0 Flash Team et al. (2023), Gemini 2.5 Flash and Gemini 2.5 Pro Comanici et al. (2025), Gemini 3.1 Pro, and Qwen3.5-Omni-Plus Team (2026). Representative open-weight audio-language baselines include Audio-Flamingo 3-7B Goel et al. (2025), Kimi-Audio-7B Ding et al. (2025), Step-Audio-2-mini-8B Wu et al. (2025a), MiDashengLM-7B Dinkel et al. (2025), Qwen2-Audio-7B Chu et al. (2024), Qwen2.5-Omni-7B Xu et al. (2025a), and the Qwen3-Omni-30B-A3B-Instruct and Qwen3-Omni-30B-A3B-Captioner variants Xu et al. (2025b). These comparison systems are used as caption generators; none receives a benchmark question, answer option, or ground-truth label while producing its caption.

**Results cited from Omni-Captioner.** Audio-Captioner-7B is the audio-only model in Ma et al. (2026b); Omni-Captioner-7B is their audio–visual model. The row marked † in Table 1 cites Audio-Captioner's MMAU 70.0% and MMAR 59.8% from their Table 3(a), and audio-only Omni-Cloze 53.2% from their Table 4(a). MMSU is not reported and is marked with a dash. Their MMAU/MMAR cascade uses GPT-4o as the QA backbone, rather than our Qwen3.6-27B Judge; their Omni-Cloze score is likewise quoted under the original protocol. These values are not matched re-evaluations. They are included in Table 1's reported-score highlights but excluded from common-protocol improvement margins. Both 7B models use a Qwen2.5-Omni-7B backbone, but their trained weights had not been released as of September 26, 2026. The official maintainer states that there are no plans to release the 7B model and instead points to the 30B-A3B version.[3] The released Qwen3-Omni-30B-A3B-Captioner is therefore a distinct checkpoint, not the paper's Audio-Captioner-7B.

**Evaluation Benchmarks.** We use four audio benchmarks. **Omni-Cloze** Ma et al. (2026b) contains cloze-style multiple-choice questions constructed by masking fine-grained spans in reference descriptions. Its *Not Given* option makes missing evidence distinguishable from an incorrect alternative. We evaluate its 2,320-sample audio subset using audio input only. **MMAR** Ma et al. (2026a) comprises 1,000 examples requiring deep reasoning over speech, general audio, music, and their mixtures. **MMAU** Sakshi et al. (2025) is a 1,000-example evaluation split spanning diverse audio-understanding and reasoning tasks. **MMSU** Wang et al. (2025) contains 5,000 evaluation items focused on spoken-language perception, understanding, and reasoning. Together, these benchmarks test whether a caption preserves fine-grained acoustic attributes, temporal and event structure, linguistic content, and the evidence needed for downstream reasoning.

For methods evaluated under our common protocol, we adopt a unified caption-as-evidence protocol across benchmarks: the generated caption is evaluated by the same frozen

[3] Official release-status response: `https://github.com/ddlBoJack/Omni-Captioner/issues/5#issuecomment-4885569721`.

Qwen3.6-27B Qwen Team (2026) text-only judge. On MMAR, MMAU, and MMSU, the judge receives the caption together with the benchmark question and answer choices, but has no access to the raw audio or gold answer. On Omni-Cloze, it receives the caption and cloze question with its answer choices under the same caption-only evidence constraint. We use the benchmark-specific official evaluator to align predictions with ground truth and report accuracy (%). Missing, invalid, or unparsed predictions are retained and scored as incorrect according to the corresponding official protocol. Judge prompting, decoding, and output parsing are fixed across methods.

# D Additional Ablation Results

## D.1 Modality-Wise Training-Stage Ablation

**Modality-Wise Results.** Fig. 6 shows that the training stages benefit the single-modality subsets differently. From S1 to S5, music improves by 11.17 points and sound by 6.06 points, whereas speech remains comparatively stable and gains 1.02 points. Among the larger mixed-modality subsets, sound–speech improves from 69.27% to 74.77%, and music–speech improves from 69.51% to 71.95%. The all-three subset rises from 41.67% to 70.83%, while sound–music decreases from 45.45% to 27.27%. These two large changes should be interpreted cautiously because they are based on only 24 and 11 examples, respectively. Together with the monotonic increase in total accuracy, the results show positive end-to-end changes on the well-represented sound, music, sound–speech, and music–speech subsets, while performance on very small modality combinations remains variable across training stages.

## D.2 Query-Budget Ablation

Table 6 reports the complete query-budget results. Here, $K$ is the maximum number of queries permitted by the interaction policy, rather than the realized number of queries in a trajectory. The Base Composer is a frozen Qwen3.6-27B model that was not trained with LOOP-GRPO or alternating co-evolution. Our Composer is a Qwen3-8B model trained with both.

For both Composers, accumulated-evidence accuracy increases through $K = 6$ and then saturates, whereas final-caption accuracy peaks at $K = 4$. This divergence is especially pronounced for the strong raw Qwen3.6-27B Base Composer. Increasing its budget from $K = 4$ to $K = 6$ raises $\mathrm{Acc}(E_g)$ from 65.66% to 66.25%, but lowers $\mathrm{Acc}(c_g^F)$ from 65.14% to 64.42%. Its retention gap therefore expands from 0.52 to 1.83 points and remains 1.49 points at $K = 8$. Thus, even a stronger raw Composer can accumulate additional QA evidence yet lose more of that information when synthesizing a longer interaction history. Our trained Qwen3-8B Composer instead keeps the retention gap between 0.33 and 0.45 points over $K = 4$–8, while its final-caption accuracy remains within 64.30–64.43%. Because the two systems use different backbones and no raw Qwen3-8B control is included, this table does not isolate the effect of LOOP-GRPO on either evidence acquisition or preservation. It instead documents their budget-scaling behavior: the strong raw Qwen3.6-27B Base accumulates additional QA evidence but loses more during final synthesis, whereas the deployed ACE-Cap Qwen3-8B Composer maintains a smaller and more stable retention gap as the history grows. Across both systems, $K = 4$ provides the best observed final-caption accuracy.

## D.3 LOOP-GRPO Credit Ablation

Table 3 separates the credit ablations for the two Composer adapters. For the Decision LoRA, we remove either question or stopping credit while keeping the Final LoRA fixed, and report downstream accuracy. For the Final LoRA, we keep the full Decision LoRA and its generated trajectories fixed, and replace the GRPO-trained Final LoRA with its SFT initialization when removing final-caption credit. This holds the acquired evidence fixed when testing final synthesis.

For the Decision LoRA, the full credit assignment obtains 64.4% on Omni-Cloze and 64.1% on MMAR. Removing question credit lowers accuracy by 1.2 and 1.5 points on the two benchmarks, respectively. Removing stopping credit reduces accuracy by 0.4 and 0.5 points. Thus, direct question credit provides the stronger quality signal, but stopping credit remains useful: it is not an additional stopping penalty, since $c_S = 0$, but supervision for deciding when the acquired evidence is sufficient. Each ablation removes one type of direct credit while retaining the policy's question and stop actions.

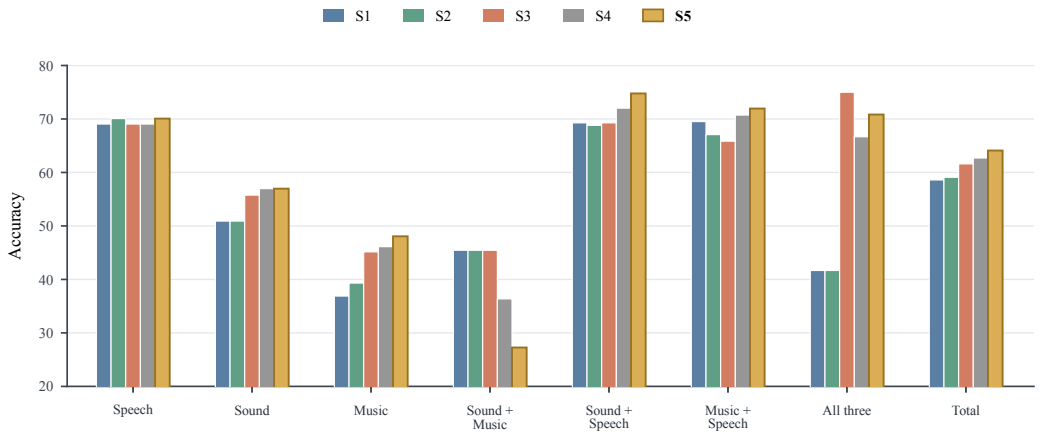



Figure 6: Modality-wise MMAR accuracy (%) across the ordered ACE-Cap training pipeline. S1 is the all-SFT initialization; S2 applies Captioner GRPO after S1; S3 further applies Instruct GRPO; S4 further applies Composer LOOP-GRPO warm-up; and S5 is the final model after alternating Composer–Instruct co-evolution. The eight groups comprise speech ($n = 294$), sound ($n = 165$), music ($n = 206$), sound–music ($n = 11$), sound–speech ($n = 218$), music–speech ($n = 82$), all three modalities ($n = 24$), and the full test set ($n = 1000$). The y-axis begins at 20%. Results for the small sound–music and all-three subsets are descriptive because their sample sizes are limited.

The Final-LoRA comparison more directly measures evidence preservation because the Decision trajectories and their Accumulated Evidence are identical. Both variants therefore start from $\mathrm{Acc}(E_g) = 64.8\%$ on Omni-Cloze. Final-caption credit raises $\mathrm{Acc}(c_g^F)$ from 63.4% to 64.4%, a gain of 1.0 point, and reduces the retention gap from 1.4 to 0.4 points. This is a reduction of approximately 71% in the aggregate accuracy gap during final synthesis, not a direct measurement of individual facts lost. The remaining 0.4-point gap also shows that preservation is not perfect, but the controlled comparison supports the intended role of final-caption credit: improving the transfer of already acquired evidence into the final long-form caption rather than changing which evidence is collected. Taken together, the ablation shows that question selection, adaptive stopping, and final evidence preservation provide complementary credit signals in LOOP-GRPO.

### D.4 Cross-Judge Robustness

We assess whether the MMAR conclusions are robust to the choice of text-only evaluator. The generated captions are frozen, and the same caption-as-evidence protocol and official MMAR evaluation are used throughout; only the Judge is replaced. Table 3(c) reports category-wise and total accuracy for Qwen3.5-27B, Qwen3.6-27B, and Gemini 3.1 Pro. The Signal, Perception, Semantic, and Cultural categories contain 43, 404, 412, and 141 examples, respectively. The two Qwen Judges yield the same total accuracy of 64.10%, despite small category-level redistributions. Gemini 3.1 Pro gives a higher total accuracy of 70.80%, an increase of 6.70 percentage points, and scores higher in all four categories; its largest gain is on the Semantic Layer (8.50–9.95 percentage points over the Qwen Judges). Importantly, all three evaluators preserve the same category ordering: Semantic is highest, followed by Cultural, Perception, and Signal. Taken together, these results support cross-Judge robustness: all three evaluators recover the same capability profile, so the category-wise conclusion is not tied to a single Judge. The 6.70-point offset under Gemini reflects a difference in absolute score calibration rather than a reversal of the observed evaluation pattern.

## E Dataset Annotation Prompts

All annotations use Gemini 3.1 Pro. We report only the core instructions that determine the gold captions and Gold-MCQs; implementation-specific orchestration and validation templates are omitted.

**Gold Caption Generation Prompt**

You are given an audio clip. Produce one detailed, fine-grained gold caption using only audible evidence. Describe the principal sound sources and events, speech or lyrics, music, acoustic properties, background sounds, and their temporal order. Do not introduce visual details or any fact unsupported by the audio; clearly qualify uncertain words and higher-level inferences. Write a single long paragraph in chronological order. Use timestamps only at major events or transitions, placing each timestamp at the beginning of the corresponding sentence or clause. Preserve intelligible speech and lyrics as faithfully as possible, including speaker changes, repetitions, and disfluencies. Output only the final caption.

**Gold-MCQ System Prompt**

You are given a gold audio caption. Generate 4–8 fine-grained MCQs that are answerable solely from facts explicitly stated in the caption. Each question must have exactly five choices, A–E, with exactly one correct answer among A–D. Choice E must be exactly "Not Given" and must never be correct. Construct the other choices as close hard negatives that differ from the answer in one subtle audio detail, such as the source, timing, count, acoustic property, speaker, instrument, or event order; avoid obvious distractors and unknown/not-mentioned wording in A–D. For each question, include the supporting evidence span and timestamp evidence when available. Assign one allowed focus level to `taxonomy_layer`, one allowed merged subcategory to `taxonomy_subcategory`, and one appropriate `aux_skill`. Return valid JSON only.

**Gold-MCQ User Prompt Template**

Use the following annotation space to construct diverse, fine-grained questions from the gold caption.
**Allowed focus levels for `taxonomy_layer`.**

- **Signal-Level:** recording physics and signal condition, including fidelity, clipping, hiss, hum, reverberation, dynamic range, mixing balance, compression artifacts, and transmission artifacts.
- **Perception-Level:** human-perceived qualities, including loudness, pitch, timbre, rhythm, speech clarity, perceived distance, foreground/background salience, and emotional intensity.
- **Semantic-Level:** direct content and events, including spoken words, lyrics, sound events, speaker turns, actions, environmental events, event order, counts, and explicit audio facts.
- **Cultural-Level:** language, accent, regionally marked speech, musical tradition, social setting, ecological context, and culturally recognizable sound cues.
- **Pragmatic / Intent / Stance-Level:** communicative goal, role, rhetorical function, interaction stance, attitude, emotion, turn-taking function, and evidence-grounded inference.
- **Musicological-Level:** tempo feel, rhythm, instrumentation, melody, harmony, lyrics, vocal style, layering, section change, texture, and genre.
- **Sound-Effect / Foley / Event-Sound Priority:** discrete non-speech event sounds, including designed effects, Foley actions, impacts, interface tones, transition whooshes, mechanical cues, crowd bursts, and vehicle cues.

**Allowed merged subcategories for `taxonomy_subcategory`.**

- Acoustic Quality and Recording Artifacts.
- Acoustic Source and Spatial Inference.
- Foreground-Background and Masking Relations.
- Acoustic Scene and Ambient Context.
- Sound Event Recognition and Reasoning.
- Temporal Event and Sequence Reasoning.
- Counting and Quantitative Statistics.
- Speech Content and Conversational Fact Retrieval.
- Phonological and Stress Pattern Decoding.
- Speaker Identity and Role Mapping.
- Key Highlight and Salient Detail Extraction.
- Emotion, Intent, and Stance Interpretation.
- Emotion Shift and Dissonance Detection.
- Socio-Cultural and Pragmatic Interpretation.
- Eco-Acoustic and Environmental Knowledge.
- Professional Domain Knowledge Reasoning.
- Instrumentation and Source Timbre.

- Rhythm, Tempo, and Meter Understanding.
- Melodic, Harmonic, and Lyric Reasoning.
- Musical Texture, Genre, and Aesthetic Reasoning.
- Foley and Designed Sound-Effect Analysis.
- Anomaly and Audio Difference Detection.

**Recommended subcategory guide by focus level.**

- **Signal-Level:** Acoustic Quality and Recording Artifacts; Acoustic Source and Spatial Inference; Foreground-Background and Masking Relations; Anomaly and Audio Difference Detection.
- **Perception-Level:** Acoustic Source and Spatial Inference; Foreground-Background and Masking Relations; Acoustic Scene and Ambient Context; Key Highlight and Salient Detail Extraction; Emotion, Intent, and Stance Interpretation.
- **Semantic-Level:** Sound Event Recognition and Reasoning; Temporal Event and Sequence Reasoning; Counting and Quantitative Statistics; Speech Content and Conversational Fact Retrieval; Phonological and Stress Pattern Decoding; Speaker Identity and Role Mapping; Key Highlight and Salient Detail Extraction.
- **Cultural-Level:** Socio-Cultural and Pragmatic Interpretation; Eco-Acoustic and Environmental Knowledge; Professional Domain Knowledge Reasoning; Musical Texture, Genre, and Aesthetic Reasoning.
- **Pragmatic / Intent / Stance-Level:** Speaker Identity and Role Mapping; Emotion, Intent, and Stance Interpretation; Emotion Shift and Dissonance Detection; Socio-Cultural and Pragmatic Interpretation; Speech Content and Conversational Fact Retrieval.
- **Musicological-Level:** Instrumentation and Source Timbre; Rhythm, Tempo, and Meter Understanding; Melodic, Harmonic, and Lyric Reasoning; Musical Texture, Genre, and Aesthetic Reasoning; Emotion, Intent, and Stance Interpretation.
- **Sound-Effect / Foley / Event-Sound Priority:** Foley and Designed Sound-Effect Analysis; Sound Event Recognition and Reasoning; Temporal Event and Sequence Reasoning; Counting and Quantitative Statistics; Acoustic Scene and Ambient Context; Anomaly and Audio Difference Detection.

**Auxiliary-skill labels.** `instrument_detail`, `source_attribution`, `sound_event_detail`, `event_knowledge`, `event_ordering`, `event_count`, `temporal_transition`, `acoustic_texture`, `acoustic_quality`, `spatial_position`, `foreground_background`, `masking`, `speaker_role`, `speaker_attribute`, `asr_keyword`, `lyric_detail`, `phonemic_stress`, `phonological_sequence`, `conversational_fact`, `key_highlight`, `emotion_state`, `emotion_flip`, `dissonant_emotion`, `pragmatic_intent`, `socio_cultural_context`, `eco_acoustic_context`, `tempo_rhythm`, `melody_harmony`, `music_texture`, `genre_reasoning`, `foley_effect`, `anomaly_event`, and `audio_difference`.
Create 4–8 high-quality MCQs and use fewer only when the caption is too sparse. Each question must remain answerable from the gold caption and use coherent focus-level, subcategory, and auxiliary-skill labels.