Closing the Verification Loop: Self-Check Captioning for Long-Paragraph Detailed Audio Captioning

Fengji Ma[1,2‡], Yan Rong[1,2‡], Xu Li[2†], Chen Zhang[2], Pengfei Wan[2], Li Liu[1*]
[1]The Hong Kong University of Science and Technology (Guangzhou)
[2]Kling Team, Kuaishou Technology
[†]Project Leader
[‡]Work conducted during internships at Kling Team, Kuaishou Technology.
[*]Corresponding Author: avrillliu@hkust-gz.edu.cn

Abstract

Long-paragraph detailed audio captioning, which requires dense and transcript-faithful descriptions of fine-grained audio content, remains unsolved for current audio-visual multimodal language models. We attribute this failure to two structural problems. The first is *data poverty*, as no public corpus jointly provides long clips, paragraph captions, and verbatim-transcript fidelity. The second is *generation-mode failure*, evidenced by a 44.8 to 46.4 percentage-point gap between right-audio and shuffled-audio multiple-choice question (MCQ) accuracy. We address both within Self-Check Captioning (SCC), a unified framework that instantiates audio-grounded question answering as the verification primitive at every lifecycle stage. SCC yields three artifacts. Long-paragraph Audio Caption 50k (LACap-50k) is a 50,222-clip audio-visual corpus with 491.5-word captions and a post-hoc automatic speech recognition (ASR) audit. Layer-Curvature Supervised Fine-Tuning (LC-SFT) is the first on-policy supervised fine-tuning method to weight tokens by intermediate-layer evidence, motivated by our identification of Late-Layer Semantic-Entropy Collapse (SEC). SCC-Verifier arbitrates among caption rollouts via audio-grounded self-answering at inference. Across multiple benchmarks, our system attains state-of-the-art among open-source captioners and is competitive with proprietary baselines. We release LACap-50k to fill the resource gap for long-paragraph detailed audio captioning research.

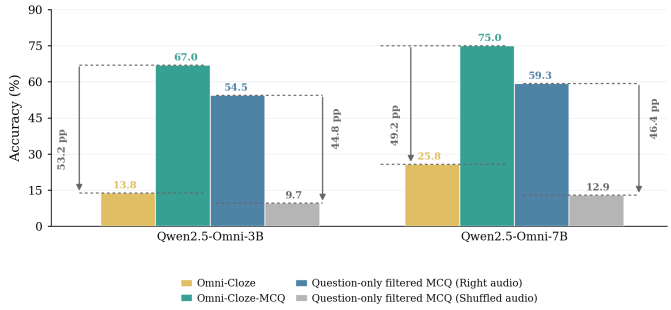
1 Introduction

Detailed long-paragraph captioning of audio–visual clips has become a demanding evaluation task for multimodal large language models (MLLMs). Such captions must track sound events, report acoustic attributes, transcribe speech verbatim, and remain faithful to the clip. Recent audio–visual and omni-modal MLLMs have advanced short-form audio understanding (Xu et al., 2025a,b; Tang et al., 2024, 2025; Ghosh et al., 2026), while agent systems synthesise diverse audio types from multimodal inputs (Rong et al., 2025). Yet fine-tuning on existing long-caption corpora consistently yields outputs that fabricate sound events, misidentify speaker attributes such as gender, and corrupt verbatim transcripts of spoken content. We argue that this persistent failure stems from two structural problems, neither of which has been jointly addressed by prior work.

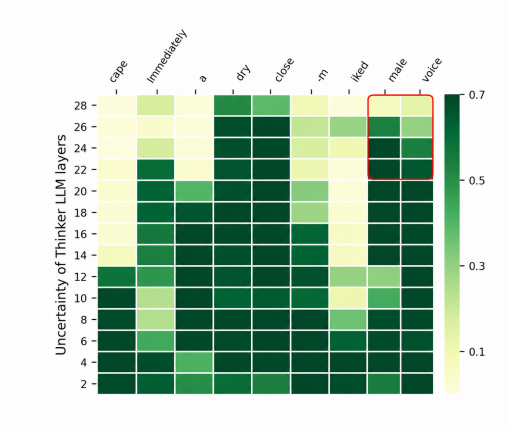
Data poverty. Training a paragraph-detailed long-form audio captioner requires data that is simultaneously long in clip duration, long in caption length, and faithful at the transcript level. No public corpus satisfies all three (Table 1). Large language model (LLM)-distilled corpora (Bai et al., 2025; Yuan et al., 2025; Chen et al., 2025) cap caption length below 50 word-like tokens, long-clip corpora (Mei et al., 2024; Munakata et al., 2026) reduce to single-sentence captions, and MECAT (Niu et al., 2025)'s 164-word per-clip aggregate concatenates six sentence-level sub-captions across disjoint perspectives rather than a coherent paragraph. None cross-check their verbatim speech transcripts against an independent automatic speech recognition (ASR) model after construction.

Generation-mode failure. The bottleneck of long-paragraph detailed audio captioning lies in free-form generation rather than in audio perception alone. Pre-trained Qwen2.5-Omni-3B/7B attain 67.0%/75.0% on Omni-Cloze (Ma et al., 2025) in multiple-choice question (MCQ) form but only 13.8%/25.8% in cloze form (Figure 1a). On a question-only filtered audio-dependent subset, MCQ accuracy collapses from 54.5%/59.3% under correct audio to 9.7%/12.9% under shuffled audio, ruling out the universal recognition-generation format effect and text-prior shortcuts. Recent on-policy supervised fine-tuning (SFT) vari-

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 60999–61017
October 24-29, 2026 ©2026 Association for Computational Linguistics



(a) Omni-Cloze accuracy under four settings: cloze generation, multiple-choice question (MCQ) recognition, and question-only filtered MCQ with right vs. shuffled audio. In all settings the model receives only the audio stream as input.



(b) Token-level semantic entropy across the 28 Thinker layers of Qwen2.5-Omni-7B; red box marks the hallucinated " male voice " tokens.

Figure 1: **Generation-mode failure of long-paragraph audio captioning.** *(a)* Qwen2.5-Omni-3B/7B accuracy on Omni-Cloze under four settings: cloze generation, MCQ recognition, and the same MCQ format restricted to a question-only filtered audio-dependent subset with right vs. shuffled audio. *(b)* Token-level semantic entropy across the 28 Thinker layers; the red box marks tokens of the hallucinated " male voice " on a clip whose actual speaker is female.

Table 1: **Comparison of Long-paragraph Audio Caption 50k (LACap-50k) with representative audio captioning corpora.** MECAT denotes the Multi-Experts Constructed Benchmark for Fine-Grained Audio Understanding Tasks.

Dataset	#Clips	Average Duration (s)	Average Text Length	Unique Tokens
AudioCaps (Kim et al., 2019)	46k	10.00	9.03	5.5k
Clotho (Drossos et al., 2020)	5k	22.50	11.00	5.5k
WavCaps (Mei et al., 2024)	403k	67.59	7.80	23.1k
Auto-ACD (Sun et al., 2024b)	1.5M	10.00	18.10	20.3k
FusionAudio-1.2M (Chen et al., 2025)	1.2M	10.00	47.18	28.1k
MECAT (Niu et al., 2025)	40k	10.05	164.51	32.8k
LACap-50k (Ours)	50k	**61.12**	**491.51**	**78.6k**

ants (Wu et al., 2025a; Zhu et al., 2025; Gan et al., 2026; Qin and Springenberg, 2025; Zhao et al., 2026) reweight policy-sampled tokens using only the final-layer probability π_θ. Layer-by-layer probing of the language model further reveals that hallucinated tokens hold high semantic entropy through the middle and late layers and commit abruptly only at the final one or two (Figure 1b), an in-backbone signature we call **Late-Layer Semantic-Entropy Collapse (SEC)** invisible to any reweighting from π_θ alone.

To address these two challenges jointly, we introduce **Self-Check Captioning (SCC)**, a unified framework in which audio-grounded question answering serves as the verification primitive at every lifecycle stage of a long-paragraph audio caption. At the *data-construction* stage (§3.1), a single Gemini model drafts a caption from video+audio and re-answers its self-generated questions on the audio stream alone (*self-model* and *modality-subtraction self-check*), yielding the **Long-paragraph Audio Caption 50k (LACap-50k)** corpus. At the *training* stage (§3.2), an external text LLM scores candidate captions against MCQs derived from this gold. Combined with a layer-curvature reliability signal that down-weights SEC tokens, this constitutes **Layer-Curvature Supervised Fine-Tuning (LC-SFT)**. At the *inference* stage (§3.3), the trained model arbitrates among caption rollouts through audio-grounded MCQs answered from the audio alone (**SCC-Verifier**).

This work makes three contributions.

- We propose **Self-Check Captioning (SCC)**, a unified framework that operationalises a single audio-grounded verification primitive at

the three lifecycle stages of a long-paragraph audio caption: modality-subtraction self-check combined with self-model self-check at the data stage, a caption-only MCQ reward reweighting at the training stage, and an audio-grounded self-verifier (**SCC-Verifier**) that arbitrates among multiple caption rollouts by exploiting the captioner's MCQ-recognition strength at the inference stage.

- We release **LACap-50k**, the data-stage product of SCC: a 50,222-clip audio–visual corpus of paragraph-detailed captions (491.5 words on average) over predominantly long clips, with sub-word vocabulary 2.4× richer than the strongest prior multi-expert corpus and post-hoc transcript-level ASR verification absent from prior datasets. LACap-50k is released to fill the resource gap for long-paragraph detailed audio captioning, enabling paragraph-level captioner training and the construction of fine-grained audio question answering (QA), long-form caption fidelity, and modality-interaction benchmarks beyond the scope of this paper.

- We are the first to formally diagnose long-paragraph detailed audio captioning failure as a **generation-mode phenomenon rather than a perception bottleneck**, evidenced by the 44.8 to 46.4 percentage-point gap between right-audio and shuffled-audio MCQ accuracy on a question-only filtered audio-dependent subset of Omni-Cloze (Figure 1a), and we identify **Late-Layer Semantic-Entropy Collapse (SEC)** as its in-backbone token-level signature (Figure 1b). Motivated by SEC, we propose **LC-SFT**, the first on-policy SFT variant to weight tokens by intermediate-layer evidence rather than by the final-layer probability used in prior variants.

2 Related Work

2.1 Long-Paragraph Detailed Captioning

Paragraph-length detailed captioning spans images (Garg et al., 2024), video (Chai et al., 2025), audio–visual clips (Tang et al., 2025; Ma et al., 2025), time-aware dense audio (Rong et al., 2026b). The audio corpora that condition this task span short single-sentence captions (Kim et al., 2019; Drossos et al., 2020), large-scale LLM-distilled audio-only corpora (Mei et al., 2024; Sun et al., 2024b; Bai et al., 2025; Yuan et al., 2025; Chen et al., 2025), and recent multi-expert or long-clip corpora (Niu et al., 2025; Munakata et al., 2026; Niu et al., 2026); none pairs paragraph-length detail with predominantly long clips and post-construction transcript-level audit, the gap LACap-50k fills (Table 1). Agentic audiobook systems coordinate speech, sound effects, and music (Rong et al., 2026c). At the model side, the Qwen-Omni (Xu et al., 2025a,b), SALMONN (Tang et al., 2024), and Audio Flamingo (Kong et al., 2024; Ghosh et al., 2025, 2026) lines have established strong short-form audio-visual understanding. AudioGenie-Reasoner uses iterative evidence refinement for coarse-to-fine audio reasoning (Rong et al., 2026a). Our diagnostic (Fig. 1) shows that the same models attain substantially higher audio-only MCQ accuracy than cloze accuracy on questions derived from gold captions, indicating that the bottleneck lies in long-form generation rather than perception alone.

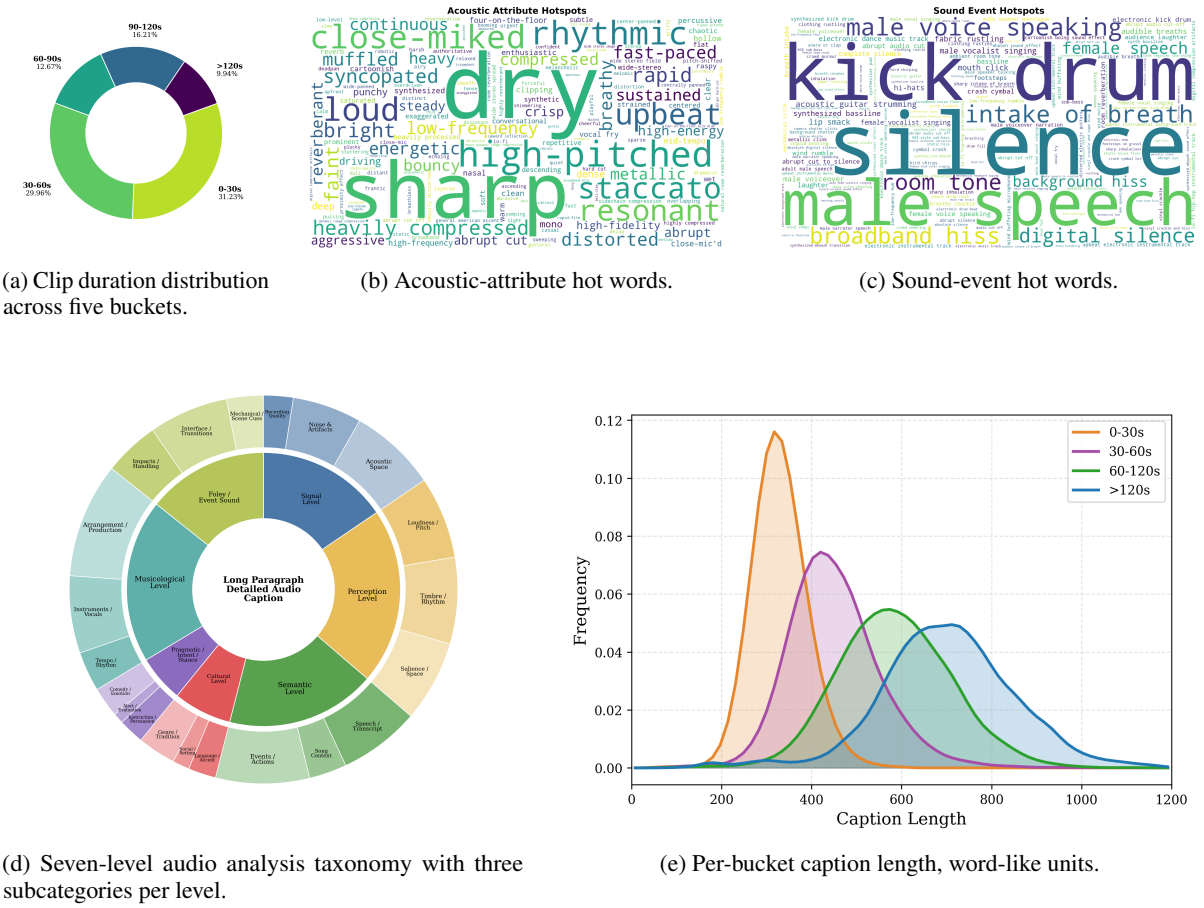
2.2 On-Policy SFT and Self-Verification

A growing line of work re-examines supervised fine-tuning through an on-policy lens (Wu et al., 2025a; Zhu et al., 2025; Gan et al., 2026; Qin and Springenberg, 2025; Zhao et al., 2026), with related recipes pursuing full reinforcement learning under verifiable rewards (Shao et al., 2024). All of these compute the token reweighting signal from the model's final-layer probability; LC-SFT extends the lineage by using late-layer semantic-entropy collapse, a structural in-backbone reliability signal invisible to any final-layer reweighting. Iterative verification spans self-critique (Madaan et al., 2023), sample aggregation (Wang et al., 2022), evaluator scoring (Zheng et al., 2023), and agentic tool calls for captioning (Ma et al., 2025). ACE-Cap acquires missing evidence through multi-turn audio-grounded questioning (Ma et al., 2026a). SCC applies one audio-grounded verification primitive across data construction, on-policy training, and inference-time arbitration.

3 Self-Check Captioning (SCC)

3.1 Data Stage: LACap-50k Construction

LACap-50k is the data-stage product of Self-Check Captioning: a 50,222-clip audio–visual corpus paired with one long-paragraph caption per clip (491.5 word-like tokens on average, 78.6 k unique sub-word tokens). We sample the 50,222 source



(a) Clip duration distribution across five buckets.

(b) Acoustic-attribute hot words.

(c) Sound-event hot words.

(d) Seven-level audio analysis taxonomy with three subcategories per level.

(e) Per-bucket caption length, word-like units.

Figure 2: **LACap-50k overview.** *Top*: clip-duration distribution and hot-word clouds for acoustic attributes and sound events. *Bottom*: seven-level audio analysis taxonomy and per-bucket caption-length kernel density.

videos from ASID-1M (Li et al., 2026) based on audio type, duration, and sound-event coverage; Figure 2 shows their distribution. The entire construction loop uses Gemini3.1-Pro for both caption generation and verification. We term this design *self-model self-check*: the same model that authors the initial caption from the audio–visual clip also generates the verification questions and supplies their answers, removing the inter-model disagreement noise that a multi-model pipeline would introduce. Crucially, the verification questions are open-form queries targeting the fine-grained acoustic details of specific audio events, and they are re-answered on the audio stream alone, with the video and the initial caption deliberately withheld. We call this audio-only re-listening pass *modality-subtraction self-check*: it strips away visual evidence so that the revised caption is anchored in audio content rather than in details inferred from the video, mitigating the modality bias that audio–visual inputs would otherwise leave in the draft. Captions span the seven-level audio analysis taxonomy of Figure 2d, covering signal, perception, semantic, cultural, pragmatic, musicological, and Foley event-sound facets. Subcategory definitions and lexical evidence are deferred to Appendices A.1 and A.2.

Distribution. LACap-50k deliberately spans short, medium, and long clips: 31.23% in 0–30 s, 29.96% in 30–60 s, and the remaining 38.81% beyond 60 s (Figure 2a). Per-bucket caption length grows monotonically with clip duration (Figure 2e), the mode shifting from $\sim$ 320 tokens for 0–30 s clips to $\sim$700 tokens beyond 120 s.

Transcript fidelity. We cross-check the verbatim speech segments inside each caption against the open-source Qwen3-ASR-1.7B model (Shi et al., 2026) after construction. Over 10,975 speech-heavy captions yielding 36,809 transcript spans, LACap-50k attains a micro-averaged character error rate (CER) of 6.23% and word error rate (WER) of 9.21% (Table 2); the English subset is markedly cleaner than the non-English subset, consistent with the relative coverage of the reference ASR's training data. Metric definitions and threshold-

Subset	Micro CER	Macro CER	Micro WER	Macro WER
English	5.58	7.41	8.60	10.86
Non-English	9.31	11.62	10.20	13.88
All	**6.23**	**8.32**	**9.21**	**11.51**

Table 2: **Transcript fidelity of LACap-50k** (%) on 36,809 transcript spans from 10,975 speech-heavy captions, against the open-source Qwen3-ASR-1.7B reference (Shi et al., 2026).

share statistics are deferred to Appendix A.4.

Post-hoc non-speech event grounding. To complement the transcript-fidelity audit, we evaluate non-speech event grounding on 1,000 final LACap-50k captions: 500 clips of 0–30 s and 500 of 30–60 s. Qwen3.6-27B (Qwen Team, 2026) extracts explicit atomic event claims from each caption. We canonicalise, filter, and deduplicate these claims using the eligibility rules before audio verification.

The open-source Frame-wise Language-Audio Modeling system OpenFLAM (Wu et al., 2025b) receives the source audio and each event query. Of 3,985 eligible claims, 3,770 (94.60%) are corroborated, 151 (3.79%) are not corroborated, and 64 (1.61%) are uncertain. Appendix A.5 provides the complete protocol and reporting definitions.

Positioning and release. A head-to-head comparison against representative audio captioning corpora is reported in Table 1 (§1): LACap-50k is roughly an order of magnitude longer in average caption length than the strongest prior LLM-distilled corpus, three times longer than the strongest prior multi-expert corpus, and the only listed corpus paired with a post-construction transcript-level ASR cross-check. LACap-50k is released under a permissive license to fill the resource gap for long-paragraph detailed audio captioning research; together with the paired multimodal initial draft and audio-only revised final caption, and the audio-only verification question–answer pairs from the SCC self-check loop, the release enables paragraph-level captioner training and the construction of fine-grained audio QA, long-form caption fidelity, and modality-interaction benchmarks beyond the scope of this paper.

3.2 Training Stage: Layer-Curvature SFT (LC-SFT)

We instantiate the on-policy SFT scheme as **LC-SFT**, an on-policy fine-tuning objective in which a verifiable caption-only reward selects which candidate captions drive the gradient, and a layer-curvature reliability signal computed from the policy's intermediate representations selects which tokens within each candidate carry that gradient.

Setting and two-stage training. Let $\pi_\theta(y\,|\,x) = \prod_{t=1}^{T}\pi_\theta(y_t\,|\,x, y_{<t})$ denote the autoregressive policy of the audio-visual language model. We train within the on-policy SFT regime of Zhao et al. (2026); Qin and Springenberg (2025) in two stages. Stage 1 applies standard SFT on $\mathcal{D} = \{(x, y^\star)\}$ to grant paragraph-level capacity. Stage 2 (LC-SFT, defined below) refines the Stage-1 checkpoint by addressing residual generation-mode failures visible only in the SFT-tuned policy's own rollouts, from which all Stage-2 response groups (§3.2.1) are sampled. Full hyperparameters appear in Appendix C.3.

3.2.1 Response Group and Verifiable Reward

For each input x we construct a response group of K caption candidates,

$$\mathcal{G}_x \;=\; \{\, y_{\text{gold}},\; y_{r_1},\; \ldots,\; y_{r_{K-1}} \,\}, \qquad (1)$$

where y_{gold} is the LACap-50k reference for x (§3.1) and $\{y_{r_i}\}_{i=1}^{K-1}$ are $K-1$ multinomial samples freshly drawn from the current policy $\pi_\theta(\cdot\,|\,x)$ at every gradient step. The choice of K is reported in §4.

Each candidate $y_k \in \mathcal{G}_x$ is scored by a caption-only MCQ reward. Offline, a frozen external text language model converts y_{gold} into a pool $\mathcal{Q}_x$ of cloze-style questions about fine-grained audio facts in the clip, each paired with four close distractors and a "not given" option. At training time, the same evaluator answers $\mathcal{Q}_x$ conditioned only on the candidate text y_k (audio and video removed), and we record the accuracy

$$R(y_k) \;=\; \frac{1}{|\mathcal{Q}_x|}\sum_{q\in\mathcal{Q}_x} \mathbf{1}\big[\text{eval}(y_k, q) = \text{answer}(q)\big]. \qquad (2)$$

The caption-only protocol forces y_k to retain every fact that $\mathcal{Q}_x$ probes, so $R(y_k)$ measures retention of y_{gold}'s verified audio-fact content rather than stylistic similarity (a comparison with the inference-time verifier identity appears in Appendix F). We standardise R within each group to obtain the per-candidate advantage

$$A(y_k) \;=\; \frac{R(y_k) - \mu(\mathcal{G}_x)}{\sigma_R(\mathcal{G}_x) + \epsilon}, \qquad k = 1, \ldots, K, \qquad (3)$$

where $\mu(\mathcal{G}_x)$ and $\sigma_R(\mathcal{G}_x)$ are the empirical mean and standard deviation of R over $\mathcal{G}_x$ and $\epsilon > 0$ a small constant. Because $\mathcal{Q}_x$ is derived from y_{gold}, $R(y_{\text{gold}}) \approx 1$ by construction and $A(y_{\text{gold}})$ is the largest value in the group almost surely, serving as a supervised anchor. Below-mean candidates receive $A(y_k) < 0$ and contribute an unlikelihood gradient.

3.2.2 Layer-Curvature Reliability Signal

A scalar score $A(y_k)$ cannot distinguish reliably-learned tokens from individual hallucinated tokens inside an otherwise high-reward candidate. We obtain a per-token reliability signal directly from the policy's own intermediate representations through a parameter-free early-exit probe that reuses the model's final layer-norm $\mathrm{Norm}(\cdot)$ and language-model head W_{lm} without modification.

For each assistant token $y_{k,t}$ and each layer $\ell \in \mathcal{L} = (l_1, \ldots, l_m)$ from a fixed set of late Thinker layers, the early-exit predictive distribution and the per-layer negative log-likelihood of the realised token are

$$\begin{aligned} p_{k,t}^{(\ell)} &= \mathrm{softmax}\big(W_{\text{lm}}\,\mathrm{Norm}(h_{k,t}^{(\ell)})\big), \\ U_{k,t}^{(\ell)} &= -\log p_{k,t}^{(\ell)}(y_{k,t}), \end{aligned} \tag{4}$$

where $h_{k,t}^{(\ell)}$ is the hidden state at layer ℓ. The trajectory $(U_{k,t}^{(l_1)}, \ldots, U_{k,t}^{(l_m)})$ describes how the model resolves its prediction for $y_{k,t}$ across depth. For computational tractability, we use realised-token negative log-likelihood (NLL) U (surprisal) as a proxy for token-level semantic entropy rather than computing entropy over the full predictive distribution; the SEC name refers to the resulting late-layer collapse pattern.

SEC and its detection. On audio-faithful tokens this trajectory descends smoothly across the late layers. On hallucinated tokens it remains elevated through the middle and late layers and drops abruptly only at the final one or two layers (Late-Layer Semantic-Entropy Collapse, SEC; Fig. 1b). Because the final-layer probability $\pi_\theta(y_{k,t} \mid \cdot)$ is unchanged by whether the preceding trajectory was smooth or abrupt, it is structurally blind to SEC, and capturing SEC requires a feature that compares predictions across depth. We define the *layer-curvature* of token (k,t) as the smoothed second-order difference of $U_{k,t}^{(\ell)}$ summed over the intermediate layer indices,

$$\mathrm{Curv}_{k,t} = \sum_{j=2}^{m-1} \sqrt{\big(\Delta U_{k,t}^{(j+1)} - \Delta U_{k,t}^{(j)}\big)^2 + \epsilon^2}, \tag{5}$$

where $\Delta U_{k,t}^{(j)} = U_{k,t}^{(l_j)} - U_{k,t}^{(l_{j-1})}$ and the smoothed envelope $\sqrt{u^2 + \epsilon^2}$ approximates $|u|$ differentiably. The second-order form is essential. $\mathrm{Curv}_{k,t}$ vanishes on any constant-slope trajectory and is large only on trajectories with a true kink, isolating the SEC signature. A formal justification for this choice appears in Appendix C.1.

We convert the curvature to a multiplicative stop-gradient token weight,

$$w_{k,t} = \mathrm{sg}\Big(\exp\big(-\kappa\,\mathrm{Curv}_{k,t}\big)\Big) \in (0, 1], \tag{6}$$

where $\mathrm{sg}[\cdot]$ is the stop-gradient operator that prevents the model from masking the kink by reshaping intermediate representations, and $\kappa > 0$ is a validation-selected suppression strength (value in §4).

3.2.3 The LC-SFT Objective

Combining $A(y_k)$ from Eq. (3) and $w_{k,t}$ from Eq. (6), the per-candidate term is the w-weighted average of the per-token log-probability $\ell_{k,t} := \log \pi_\theta(y_{k,t} \,|\, x, y_{k,<t})$,

$$\widetilde{\mathcal{L}}(y_k;\, x) = -\frac{\sum_{t=1}^{T_k} w_{k,t}\,\ell_{k,t}}{\sum_{t=1}^{T_k} w_{k,t} + \epsilon}, \tag{7}$$

and the group-level loss is the A-weighted linear combination

$$\mathcal{L}(\mathcal{G}_x) = \sum_{k=1}^{K} A(y_k)\,\widetilde{\mathcal{L}}(y_k;\, x), \tag{8}$$

yielding the training objective $\mathcal{L}_{\text{LC}}(\theta) = \mathbb{E}_{x\sim\mathcal{D}}[\mathcal{L}(\mathcal{G}_x)]$, approximated by minibatch averaging over freshly sampled rollouts at every step. The advantage $A(y_k)$ thus selects which candidates contribute to the gradient (positive as maximum-likelihood, negative as unlikelihood) while $w_{k,t}$ selects which tokens within each candidate carry it. The full gradient and its two-axis interpretation appear in Appendix C.2.

3.3 Inference Stage: SCC-Verifier

A single inference pass remains exposed to sampling stochasticity even after LC-SFT (§3.2). We close the SCC loop at inference with **SCC-Verifier**,

which re-uses the LC-SFT-trained model itself as an audio-grounded multiple-choice answerer to arbitrate among candidate captions in three steps. *(i)* The policy π_θ produces M candidates for an audio-visual clip x, one greedy decoding $y^{(0)}$ and $M-1$ multinomial samples $y^{(1)}, \ldots, y^{(M-1)}$ at temperature T. *(ii)* For each $y^{(m)}$, a frozen external text language model extracts N atomic audio-grounded claims and converts each into a five-option MCQ whose correct option is the claim asserted by $y^{(m)}$ and whose distractors are minimally perturbed alternatives along the same audio facet plus a "not given" option, yielding a pool $\mathcal{Q}^{(m)}$ that encodes precisely what $y^{(m)}$ claims about the audio. *(iii)* For each $(y^{(m)}, q)$ with $q \in \mathcal{Q}^{(m)}$, the LC-SFT-trained model answers q in audio-only MCQ mode on x_{audio} with $y^{(m)}$ withheld, and the candidate score is the fraction of its own claims that survive,

$$S(y^{(m)}) = \frac{1}{|\mathcal{Q}^{(m)}|} \sum_{q \in \mathcal{Q}^{(m)}} \mathbf{1}\big[a_\theta(x_{\text{audio}}, q) = a_q^{(m)}\big], \tag{9}$$

where a_θ denotes the model's audio-only answer and $a_q^{(m)}$ the option asserted by $y^{(m)}$. SCC-Verifier outputs $y^\star = \arg\max_m S(y^{(m)})$, complementary by construction to the training-time MCQ reward of §3.2.1 (Appendix F).

4 Experiments

This section reports the empirical performance of LC-SFT (§3.2) and the SCC-Verifier inference protocol (§3.3) on four benchmarks that jointly probe audio understanding through downstream question answering and direct detailed-caption fidelity: **Massive Multi-Task Audio Understanding and Reasoning (MMAU)** (Sakshi et al., 2025), **MMAR (Deep Reasoning in Speech, Audio, Music, and Their Mix)** (Ma et al., 2026b), **Massive Multi-Task Spoken Language Understanding and Reasoning (MMSU)** (Wang et al., 2025), and **Omni-Cloze** (Ma et al., 2025). The first three follow a *caption-as-evidence* protocol in which the generated caption is the sole audio-derived evidence supplied to a frozen text-only judge language model (LLM-as-a-judge), which then answers the benchmark questions; the fourth scores the captions directly via cloze-style multiple-choice blanks. Higher scores indicate greater caption fidelity; Appendix G details the setup. All Gemini results use Gemini3.1-Pro-Preview (Gemini Team, 2023), denoted **Gemini3.1-Pro** throughout.

Model	MMAU	MMAR	MMSU
Proprietary Models			
GPT-4o Audio	62.4	59.3	56.4
Gemini 2.0 Flash	58.6	50.6	51.0
Gemini 2.5 Flash	65.6	58.2	58.1
Gemini 2.5 Pro	70.0	64.1	–
Open-Source Models			
SALMONN-13B	58.4	42.5	–
MiDashengLM-7B	59.4	50.7	–
Qwen2-Audio-7B	63.3	44.2	53.3
Qwen2.5-Omni-7B	65.2	51.8	60.6
Ours	**68.3**	**57.0**	**62.4**

Table 3: **Caption-as-evidence QA accuracy** (%) on MMAU, MMAR, and MMSU. A frozen text-only judge LLM (Qwen Team, 2026) answers benchmark questions using only the model-generated caption as evidence. The last row is our model.

4.1 Caption-as-Evidence QA on Audio Benchmarks

We adopt a *caption-as-evidence* protocol: the captioner generates a detailed audio caption for each clip, and a frozen text-only judge language model (LLM-as-a-judge) answers the benchmark question using only the caption as its source of audio-derived evidence. The judge accuracy on MMAU (Sakshi et al., 2025), MMAR (Ma et al., 2026b), and MMSU (Wang et al., 2025) therefore measures the extent to which the caption stands in for the audio itself: a higher score implies that the caption preserves the fine-grained acoustic information required for downstream reasoning, while a lower score indicates that audio facts are either omitted or hallucinated and cannot be recovered from the caption text. This protocol echoes the utility-based caption evaluation paradigm advocated by Yang et al. (2025), in which a frozen text-only LLM (Qwen Team, 2026) answers benchmark questions using only the caption as its evidence source; high judge accuracy is taken to mean that the caption successfully *stands in* for the underlying signal for downstream reasoning.

Training-evaluation consistency. This protocol is structurally identical to the training-time MCQ reward of §3.2.1, differing only in the source of the questions (gold-derived during training versus benchmark MCQs at evaluation). The LC-SFT objective therefore directly optimises the same caption-as-evidence quality that our evaluation measures.

Our LC-SFT-trained Qwen2.5-Omni-7B im-

Model	Accuracy (%) ↑
Proprietary Models	
Gemini 2.0 Flash	31.7
Gemini 2.5 Flash	36.2
Gemini 2.5 Pro	44.1
Gemini3.1-Pro	59.5
Gemini3.1-Pro (with SCC)	68.1
Open-Source Models	
video-SALMONN-13B	2.3
videoLLaMA 2-7B	4.3
Qwen2.5-Omni-7B	14.1
video-SALMONN 2-7B	40.3
Omni-Captioner-7B	54.5
Ours	**58.9**

Table 4: **Results on Omni-Cloze (audio-only split)**, audio-visual (video) input. The last row is our method based on Qwen2.5-Omni-7B.

proves over the vanilla backbone from 65.2 to 68.3 on MMAU (+3.1), 51.8 to 57.0 on MMAR (+5.2), and 60.6 to 62.4 on MMSU (+1.8), attaining the highest score among all open-source captioners on the three benchmarks. It further surpasses GPT-4o Audio on MMAU (68.3 vs. 62.4) and MMSU (62.4 vs. 56.4), and narrows the gap to Gemini 2.5 Pro to 1.7 points on MMAU. The larger gain on MMAR, which requires multi-step reasoning across mixed sound, music, and speech, suggests that LC-SFT's intermediate-layer reweighting most strongly benefits captions whose downstream reasoning depends on transcript-faithful audio facts, a pattern reinforced by MMSU's exclusive spoken-language focus.

4.2 Detailed Captioning Evaluation on Omni-Cloze

The caption-as-evidence judge accuracy of §4.1 measures the downstream utility of a caption but does not score the caption text itself. To complement it with a direct caption-quality measurement, we evaluate on **Omni-Cloze** (Ma et al., 2025), a cloze-style benchmark that masks fine-grained spans in a gold caption and forces the candidate captioner to fill them via multiple-choice selection, with a "not given" option that distinguishes omission from hallucination. Table 4 reports accuracy on the audio subset of Omni-Cloze under the audio-visual input setting, in which the model receives both the video and audio streams. The parallel audio-only input setting, where the model receives the audio stream alone, is reported in Appendix D.

On the audio subset of Omni-Cloze under audio-visual input, our LC-SFT-trained Qwen2.5-Omni-7B reaches 58.9%, a +44.8 absolute improvement over the vanilla backbone (14.1%) and +4.4 over the strongest open-source prior method Omni-Captioner-7B (Ma et al., 2025) (54.5%). We surpass Gemini 2.5 Pro by 14.8 points and trail Gemini3.1-Pro by only 0.6. Applying the data and inference stages of SCC to Gemini3.1-Pro yields a further +8.6 gain (59.5% to 68.1%) without any training-stage modification, demonstrating that SCC's self-check primitives function as model-agnostic plug-in modules.

Table 5: **Gemini3.1-Pro self-check outcomes.** Outcome rates (%) on the human-validated Omni-Cloze evaluation set. AV denotes audio-visual; Net is Correct minus Wrong.

Setting	Correct↑	Missing↓	Wrong↓	Net↑
Audio-visual input	59.54	37.48	**2.99**	56.55
Audio-only input	64.08	32.52	3.40	60.69
AV + audio-only self-check	**68.06**	**27.27**	4.66	**63.40**

4.3 Analysis

Self-check outcomes. Table 5 reports all three evaluated settings on the same human-validated Omni-Cloze evaluation set. Its Correct column provides the accuracy summarised in Table 4, while Missing and Wrong expose the outcome decomposition. Net is Correct minus Wrong, with Missing contributing zero. Adding the audio-only self-check to the audio-visual draft yields the highest Correct rate, lowest Missing rate, and best Net score, while Wrong rises from 2.99% to 4.66%.

Verifier-format alignment. We test whether the LC-SFT improvement transfers beyond the MCQ/verifier format and whether it depends on the text-judge family. First, SFT and LC-SFT receive audio only and freely generate one caption for each clip in the complete Clotho evaluation split (Drossos et al., 2020), using the same prompt, deterministic decoding, and a 512-token limit. This control contains no questions, answer options, caption-only judge, SCC-Verifier, or MCQ reranking. The five metrics are Best- and Mean-Chunk Contrastive Language-Audio Pretraining scores (B-CLAP and M-CLAP) (Wu et al., 2023), our RefAlign-F1 (R-F1), Sentence-BERT similarity (SBERT) (Reimers and Gurevych, 2019), and Fluency-Enhanced Sentence-BERT Evaluation (FENSE) (Zhou et al., 2022). LC-SFT improves all five metrics over SFT, including B-CLAP from

Table 6: **Open-ended Clotho captioning.** Audio-only free generation on the complete evaluation split. B-/M-CLAP denote best-/mean-chunk Contrastive Language-Audio Pretraining; R-F1 denotes our RefAlign-F1; SBERT denotes Sentence-BERT; and FENSE denotes Fluency-Enhanced Sentence-BERT Evaluation. All metrics are higher-is-better.

Model	B-CLAP	M-CLAP	R-F1	SBERT	FENSE
SFT	0.460	0.283	0.607	0.413	0.401
LC-SFT	**0.492**	**0.311**	**0.612**	**0.436**	**0.433**

Table 7: **Cross-model judge robustness.** Omni-Cloze accuracy (%) for fixed audio-visual captions; only the text judge changes. Protocol details are provided in Appendix H.

Captioner	Qwen3.6-27B	Gemini3.1-Pro
Qwen2.5-Omni-7B	14.13	16.34
SFT	57.10	59.51
LC-SFT	**58.92**	**61.70**

0.460 to 0.492 and FENSE from 0.401 to 0.433 (Table 6).

We next freeze the Omni-Cloze audio-visual captions and replace only the Qwen3.6-27B judge with Gemini3.1-Pro. LC-SFT remains above SFT, by +1.82 points under Qwen and +2.19 under Gemini (Table 7), so the gain transfers across judge families.

Recent baselines under a common protocol. Table 8 complements Tables 3 and 4 with three Qwen3-Omni variants (Xu et al., 2025b; Ma et al., 2025). Under the same audio-only caption-as-evidence protocol, LC-SFT outperforms all three variants on Omni-Cloze and MMAR, and is slightly higher on MMAU.

4.4 Ablation of the SCC Framework

We decompose SCC into its three lifecycle stages and ablate each contribution separately.

Data and inference stages. The last two proprietary rows of Table 4 (Gemini3.1-Pro with and without SCC) isolate the joint contribution of the data and inference stages, applied as a wrapper on a frozen backbone. SCC raises Omni-Cloze accuracy from 59.5% to 68.1% (+8.6 absolute) without any training-stage modification, confirming that these two stages function as model-agnostic plug-in modules.

LC-SFT decomposition. We isolate the two reweighting components of LC-SFT in a 2×2 design

Table 8: **Comparison with Qwen3-Omni variants under a common audio-only protocol.** Caption-as-evidence accuracy (%). Bold marks the best result.

Model	Omni-Cloze	MMAR	MMAU
Qwen3-Omni-Captioner	57.5	55.3	66.4
Qwen3-Omni-Instruct	35.3	54.5	67.9
Qwen3-Omni-Thinking	33.6	52.6	63.8
Ours	**59.3**	**57.0**	**68.3**

over the Stage-1 SFT baseline on the Omni-Cloze audio subset: response-level advantage reweight alone (LC-SFT$_R$, $A(y_k)$ from Eq. 3 with $w_{k,t} \equiv 1$), token-level layer-curvature reweight alone (LC-SFT$_T$, $w_{k,t}$ from Eq. 6 with $A(y_k) \equiv 1$), and the full objective combining both. Overall accuracy rises from 57.10 (SFT) to 57.28 with the advantage signal alone (+0.18), to 58.59 with the curvature signal alone (+1.49), and to 58.92 when both are combined (+1.82). The token-level curvature reweight is the dominant contributor, winning on 17 of 47 categories; the advantage signal is additive on top of curvature in 23 categories, especially speech- and narration-heavy ones. Full per-category results are reported in Appendix Table 10.

5 Conclusion

We introduced **Self-Check Captioning (SCC)**, an audio-grounded question-answering framework for verifying long-paragraph detailed audio captions during data construction, training, and inference. **LACap-50k**, **LC-SFT**, and **SCC-Verifier** attain state-of-the-art open-source results on MMAU, MMAR, MMSU, and Omni-Cloze. Applied at the data and inference stages, SCC improves Gemini3.1-Pro by 8.6 Omni-Cloze points, demonstrating model-agnostic utility.

Limitations

LACap-50k is designed as a high-quality *post-training* corpus for long-paragraph detailed audio captioning, providing 50,222 carefully verified clips that suffice for supervised fine-tuning and on-policy refinement on top of an already-capable audio-visual backbone. Constructing a comparably high-quality corpus at the scale required for *pre-training* (typically several orders of magnitude larger) remains prohibitively expensive under our SCC pipeline, and we leave this extension to future work.

Acknowledgments

This work was supported by the CCF-Kuaishou Large Model Explorer Fund (CCF-KuaiShou 2025002), the National Natural Science Foundation of China (No. 62471420), and the Guang-Dong Basic and Applied Basic Research Foundation (2025A1515012296).

References

Jisheng Bai, Haohe Liu, Mou Wang, Dongyuan Shi, Wenwu Wang, Mark D Plumbley, Woon-Seng Gan, and Jianfeng Chen. 2025. Audiosetcaps: An enriched audio-caption dataset using automated generation pipeline with large audio and language models. *IEEE Transactions on Audio, Speech and Language Processing*.

Wenhao Chai, Enxin Song, Yilun Du, Chenlin Meng, Vashisht Madhavan, Omer Bar-Tal, Jenq-Neng Hwang, Saining Xie, and Christopher Manning. 2025. Auroracap: Efficient, performant video detailed captioning and a new benchmark. In *International Conference on Learning Representations*, volume 2025, pages 16845–16889.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335. Association for Computational Linguistics.

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

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

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

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

Wangjie Gan, Miao Pan, Linbo Xi, Wenqi Zhang, Jintao Chen, Jianwei Yin, and Xuhong Zhang. 2026. Gft: From imitation to reward fine-tuning with unbiased group advantages and dynamic coefficient rectification. *arXiv preprint arXiv:2604.14258*.

Roopal Garg, Andrea Burns, Burcu Karagol-Ayan, Yonatan Bitton, Ceslee Montgomery, Yasumasa Onoe, Andrew Bunner, Ranjay Krishna, Jason Michael Baldridge, and Radu Soricut. 2024. Imageinwords: Unlocking hyper-detailed image descriptions. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 93–127.

Gemini Team. 2023. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Sreyan Ghosh, Arushi Goel, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang-gil Lee, Chao-Han Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, et al. 2026. Audio flamingo 3: Advancing audio intelligence with fully open large audio language models. *Advances in Neural Information Processing Systems*, 38:41819–41886.

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

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. 2019. 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)*, pages 119–132.

Zhifeng Kong, Arushi Goel, Rohan Badlani, Wei Ping, Rafael Valle, and Bryan Catanzaro. 2024. Audio flamingo: A novel audio language model with few-shot learning and dialogue abilities. *arXiv preprint arXiv:2402.01831*.

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

Fengji Ma, Yan Rong, Xu Li, Xuenan Xu, Chen Zhang, and Li Liu. 2026a. ACE-Cap: Active evidence acquisition via agentic co-evolution for long-paragraph fine-grained audio captioning. *arXiv preprint arXiv:2608.16162*.

Ziyang Ma, Yinghao Ma, Yanqiao Zhu, Chen Yang, Yi-Wen Chao, Ruiyang Xu, Wenxi Chen, Yuanzhe Chen, Zhuo Chen, Jian Cong, et al. 2026b. Mmar: A challenging benchmark for deep reasoning in speech, audio, music, and their mix. *Advances in Neural Information Processing Systems*, 38.

Ziyang Ma, Ruiyang Xu, Zhenghao Xing, Yunfei Chu, Yuxuan Wang, Jinzheng He, Jin Xu, Pheng-Ann Heng, Kai Yu, Junyang Lin, et al. 2025. Omni-captioner: Data pipeline, models, and benchmark for omni detailed perception. *arXiv preprint arXiv:2510.12720*.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2023. Self-refine: Iterative refinement with self-feedback. *Advances in neural information processing systems*, 36:46534–46594.

Xinhao Mei, Chutong Meng, Haohe Liu, Qiuqiang Kong, Tom Ko, Chengqi Zhao, Mark D Plumbley, Yuexian Zou, and Wenwu Wang. 2024. 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.

Hokuto Munakata, Takehiro Imamura, Taichi Nishimura, and Tatsuya Komatsu. 2026. Castella: Long audio dataset with captions and temporal boundaries. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 15352–15356. IEEE.

Yadong Niu, Tianzi Wang, Heinrich Dinkel, Xingwei Sun, Jiahao Zhou, Gang Li, Jizhong Liu, Xunying Liu, Junbo Zhang, and Jian Luan. 2025. Mecat: A multi-experts constructed benchmark for fine-grained audio understanding tasks. *arXiv preprint arXiv:2507.23511*.

Yadong Niu, Tianzi Wang, Heinrich Dinkel, Xingwei Sun, Jiahao Zhou, Gang Li, Jizhong Liu, Junbo Zhang, and Jian Luan. 2026. Acavcaps: Enabling large-scale training for fine-grained and diverse audio understanding. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 15347–15351. IEEE.

OpenAI. 2024. GPT-4o system card. *arXiv preprint arXiv:2410.21276*.

Chongli Qin and Jost Tobias Springenberg. 2025. Supervised fine tuning on curated data is reinforcement learning (and can be improved). *arXiv preprint arXiv:2507.12856*.

Qwen Team. 2026. Qwen3.6-27B: Flagship-level coding in a 27B dense model.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 3982–3992.

Yan Rong, Chenxing Li, Dong Yu, and Li Liu. 2026a. AudioGenie-Reasoner: A training-free multi-agent framework for coarse-to-fine audio deep reasoning. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 14497–14501. IEEE.

Yan Rong, Fengji Ma, Xu Li, Jinting Wang, Chen Zhang, and Li Liu. 2026b. Audiomap: Cloze-and-choice reinforcement learning for time-aware dense audio captioning. *arXiv preprint arXiv:2608.09559*.

Yan Rong, Jinting Wang, Guangzhi Lei, Shan Yang, and Li Liu. 2025. AudioGenie: A training-free multi-agent framework for diverse multimodality-to-multiaudio generation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 8872–8881.

Yan Rong, Shan Yang, Chenxing Li, Dong Yu, and Li Liu. 2026c. Dopamine audiobook: A training-free MLLM agent for emotional and immersive audiobook generation. *IEEE Transactions on Audio, Speech and Language Processing*, 34:2526–2538.

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

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

Xian Shi, Xiong Wang, Zhifang Guo, Yongqi Wang, Pei Zhang, Xinyu Zhang, Zishan Guo, Hongkun Hao, Yu Xi, Baosong Yang, et al. 2026. Qwen3-asr technical report. *arXiv preprint arXiv:2601.21337*.

Guangzhi Sun, Wenyi Yu, Changli Tang, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, Yuxuan Wang, and Chao Zhang. 2024a. video-salmonn: Speech-enhanced audio-visual large language models. *arXiv preprint arXiv:2406.15704*.

Luoyi Sun, Xuenan Xu, Mengyue Wu, and Weidi Xie. 2024b. Auto-acd: A large-scale dataset for audio-language representation learning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 5025–5034.

Changli Tang, Yixuan Li, Yudong Yang, Jimin Zhuang, Guangzhi Sun, Wei Li, Zejun Ma, and Chao Zhang. 2025. video-salmonn 2: Caption-enhanced audio-visual large language models. *arXiv preprint arXiv:2506.15220*.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. 2024. Salmonn: Towards generic hearing abilities for large language models. In *International Conference on Learning Representations*, volume 2024, pages 16607–16629.

Dingdong Wang, Junan Li, Jincenzi Wu, Dongchao Yang, Xueyuan Chen, Tianhua Zhang, and Helen Meng. 2025. Mmsu: A massive multi-task spoken language understanding and reasoning benchmark. *arXiv preprint arXiv:2506.04779*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Yongliang Wu, Yizhou Zhou, Zhou Ziheng, Yingzhe Peng, Xinyu Ye, Xinting Hu, Wenbo Zhu, Lu Qi, Ming-Hsuan Yang, and Xu Yang. 2025a. On the generalization of sft: A reinforcement learning perspective with reward rectification. *arXiv preprint arXiv:2508.05629*.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. 2023. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Yusong Wu, Christos Tsirigotis, Ke Chen, Cheng-Zhi Anna Huang, Aaron Courville, Oriol Nieto, Prem Seetharaman, and Justin Salamon. 2025b. Flam: Frame-wise language-audio modeling. *arXiv preprint arXiv:2505.05335*.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. 2025a. Qwen2.5-omni technical report. *Preprint*, arXiv:2503.20215.

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

Shijia Yang, Yunong Liu, Bohan Zhai, Ximeng Sun, Zicheng Liu, Emad Barsoum, Manling Li, and Chenfeng Xu. 2025. Captionqa: Is your caption as useful as the image itself? *arXiv preprint arXiv:2511.21025*.

Yi Yuan, Dongya Jia, Xiaobin Zhuang, Yuanzhe Chen, Zhuo Chen, Yuping Wang, Yuxuan Wang, Xubo Liu, Xiyuan Kang, Mark D Plumbley, et al. 2025. Soundvecaps: Improving audio generation with visually enhanced captions. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Hang Zhang, Xin Li, and Lidong Bing. 2023. Video-llama: An instruction-tuned audio-visual language model for video understanding. In *Proceedings of the 2023 conference on empirical methods in natural language processing: system demonstrations*, pages 543–553.

Anhao Zhao, Ziyang Chen, Junlong Tong, Yingqi Fan, Fanghua Ye, Shuhao Li, Yunpu Ma, Wenjie Li, and Xiaoyu Shen. 2026. On-policy supervised fine-tuning for efficient reasoning. *arXiv preprint arXiv:2602.13407*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

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

He Zhu, Junyou Su, Peng Lai, Ren Ma, Wenjia Zhang, Linyi Yang, and Guanhua Chen. 2025. Anchored supervised fine-tuning. *arXiv preprint arXiv:2509.23753*.

A LACap-50k Details

A.1 Audio Analysis Taxonomy

A long-paragraph audio caption must traverse multiple, qualitatively distinct facets of an audio signal from low-level acoustic properties to high-level pragmatic intent. The seven-level taxonomy summarised in Figure 2d of the main text is encoded into both the multimodal draft prompts and the audio-only relisten prompts of the construction pipeline, so that every level the draft attempts to describe is exactly a level the self-check pass probes. The levels are:

- **Signal level**: recording quality, noise and artifacts, acoustic space and reverberation, microphone distance, and other properties intrinsic to the captured waveform.
- **Perception level**: loudness and pitch, timbre and rhythm, salience and spatial impression; how the audio is perceived by a human listener.
- **Semantic level**: directly verbalisable content, including verbatim speech transcripts, song lyrics, identifiable sound events, and the relative ordering of actions in time.
- **Cultural level**: language and accent, genre tradition, and the social or situational context implied by the audio.

- **Pragmatic / intent / stance level**: communicative purpose, emotional framing, and the speaker's attitude (explaining, persuading, joking, instructing, warning, or evaluating).
- **Musicological level**: tempo and rhythmic pattern, instrumentation and vocal style, arrangement and production texture, when music or musically structured audio is present.
- **Foley / event sound level**: discrete non-speech event sounds (impacts, handling sounds, interface tones, transitions, and mechanical or scene cues), treated as first-class caption content rather than collapsed into a vague "background noise" label.

Each level expands into three concrete subcategories (outer ring of Figure 2d); these subcategories are the units at which the audio-only verifier formulates open-form relisten questions during data construction.

A.2 Lexical Evidence of Taxonomy Coverage

To verify that the taxonomy is actually exercised by the released captions (rather than only by the prompts), we aggregate the vocabulary of all 50,222 final captions into two facet-conditional hotword clouds shown in Figures 2b and 2c of the main text. The *acoustic-attribute* cloud is dominated by descriptors that span the signal and perception levels of the taxonomy: dynamic and spectral terms (*loud*, *faint*, *bright*, *high-pitched*, *low-frequency*, *muffled*), articulation terms (*sharp*, *staccato*, *sustained*, *breathy*), recording-style terms (*close-miked*, *dry*, *reverberant*, *compressed*, *distorted*), and tempo-character terms (*rhythmic*, *syncopated*, *upbeat*, *fast-paced*, *rapid*). The *sound-event* cloud is dominated by terms that span the semantic and Foley levels: speech events (*male speech*, *female speech*, *male voice speaking*, *male voiceover*), human-body events (*intake of breath*, *audible breaths*, *lip smack*, *footsteps*), ambient and silence events (*silence*, *digital silence*, *room tone*, *background hiss*, *broadband hiss*), percussive musical events (*kick drum*, *hi-hats*, *crash cymbal*, *synthesized kick drum*), melodic events (*acoustic guitar strumming*, *sub-bass*, *bassline*), and editorial events (*abrupt cut to silence*, *abrupt audio cut*).

A.3 Construction Pipeline Details

The data-stage self-check of §3.1 is instantiated as a sequence of independent inference passes using Gemini3.1-Pro-Preview at temperature 0, with no contextual state propagated across passes. The first pass receives the complete original audio–video file, including clips longer than 120 s, without client-side sampling, segmentation, truncation, or additional compression, and produces an initial draft caption. Each subsequent verification pass is launched independently on the complete soundtrack, extracted as a 16-kHz mono WAV, and receives only that audio together with one verification question generated from the draft. The video frames, the draft caption, and earlier verification inputs and outputs are withheld. A final aggregation pass uses the verification answers to revise the draft into the released LACap-50k caption. The 0.2 fps and 2–32 frame setting in §G applies only to Qwen2.5-Omni training, not to Gemini data construction.

A.4 Transcript-Fidelity Metric Definitions

Qwen3.5-27B reads only the final caption—not the audio, ASR transcript, or error scores—and copies exact source spans containing explicit spoken, sung, or chanted text while preserving their language and script. We exclude translations, paraphrases, indirect speech, acoustic descriptions, and non-human sound effects, and retain spans containing at least six Unicode alphanumeric units (ten for short vocalisations). A caption is designated speech-heavy when it contains at least 20 retained units or at least two valid spans, yielding 10,975 captions and 36,809 spans. Qwen3-ASR-1.7B then transcribes each complete clip once. After Unicode normalisation, every caption span is aligned to the minimum-Levenshtein local subsequence of the full ASR transcript; residual paraphrased or indirect descriptions are excluded rather than counted as ASR errors. The resulting values are therefore an automatic consistency audit of explicit transcript spans, not a human-ground-truth ASR evaluation or a full-caption factuality measure.

For each transcript span $s \in \mathcal{S}$ extracted from a LACap-50k caption, let h_s denote the hypothesis transcript taken from the caption and r_s the Qwen3-ASR reference. Let $\mathrm{EditDist}_c(\cdot,\cdot)$ and $\mathrm{EditDist}_w(\cdot,\cdot)$ denote the Levenshtein edit distance at the character and word level, and let $|r_s|_c, |r_s|_w$ denote the reference length in characters and words. The per-span error rates are $\mathrm{CER}(s) = \mathrm{EditDist}_c(h_s, r_s)/|r_s|_c$ and $\mathrm{WER}(s) = \mathrm{EditDist}_w(h_s, r_s)/|r_s|_w$. We aggregate the character-level error rates across the span

set $\mathcal{S}$ in two complementary ways:

$$\text{Micro-CER} = \frac{\sum_{s\in\mathcal{S}} \text{EditDist}_c(h_s, r_s)}{\sum_{s\in\mathcal{S}} |r_s|_c}, \quad (10)$$

$$\text{Macro-CER} = \frac{1}{|\mathcal{S}|} \sum_{s\in\mathcal{S}} \text{CER}(s). \quad (11)$$

Micro aggregates edit counts before normalising, so each span contributes proportionally to its reference length and the dataset-wide rate is dominated by long spans; *macro* normalises within each span first and then averages, treating short and long spans equally. The word-level rates Micro-WER and Macro-WER are defined analogously by replacing character edit distance and character reference length with their word-level counterparts. We additionally report the share of spans whose per-span error rate stays below a conventional quality threshold of 0.2, namely $\text{CER}_{\leq 0.2} = |\{s : \text{CER}(s) \leq 0.2\}|/|\mathcal{S}|$ and the analogously defined $\text{WER}_{\leq 0.2}$; over LACap-50k, 87.69% of spans fall under the 0.2 CER threshold and 82.55% under the 0.2 WER threshold. The macro rates (8.32% CER and 11.51% WER overall) sit slightly above the micro rates (6.23% and 9.21%), indicating a mild concentration of errors on shorter spans rather than systematic long-span failure.

A.5 Post-hoc Non-speech Event Grounding Audit

This audit complements the transcript-fidelity analysis by testing whether explicit non-speech event claims in final captions are grounded in the source audio.

Cohort and caption-only claim extraction. The formal cohort contains 1,000 unique LACap-50k clips, stratified into 500 clips of 0–30 s and 500 clips of 30–60 s. Clips used to develop the extraction prompt, eligibility rules, or OpenFLAM thresholds are excluded. At temperature 0, Qwen3.6-27B (Qwen Team, 2026) receives only the final caption and extracts, for every explicit non-speech event, its source text, a short detector-oriented query, and any stated time hint. All 1,000 captions are parsed successfully.

Eligibility filtering. We exclude negated or absence claims, ambiguous alternatives, compound queries without a single detection outcome, silence and editing transitions, capture artifacts and vague background-noise descriptions, extremely low-salience details, speculative causal interpretations, and claims outside the detector's reliable semantic scope. The remaining queries are canonicalised into short, self-contained event phrases, and identical clip–query pairs are deduplicated. Retained claims span human non-speech vocalisations, animals, vehicles and machinery, alarms and electronics, impacts and household objects, natural sounds, movement, music and percussion, crowds, and environmental ambience.

OpenFLAM inference. We use OpenFLAM (Wu et al., 2025b). Audio is loaded as a 48-kHz mono waveform and processed in overlapping 10-s windows with a 5-s hop; short final windows are zero-padded. We run OpenFLAM in `unbiased` frame-level mode, apply a width-3 median filter to each similarity curve, and aggregate responses across all windows. Let $s_{\max}$ be the whole-clip peak response and d_{active} the accumulated duration above the positive threshold. We label a query `corroborated` when $s_{\max} \geq 0.50$ and $d_{\text{active}} \geq 0.20$ s, `not_corroborated` when $s_{\max} < 0.35$, and `uncertain` otherwise.

Results. After eligibility filtering, canonicalisation, and deduplication, the cohort contains 3,985 claims: 3,770 corroborated (94.60%), 151 not corroborated (3.79%), and 64 uncertain (1.61%). The expert-supported fraction is $3770/3985 = 94.60\%$; decision coverage is $(3770+151)/3985 = 98.39\%$; and support among decided claims is $3770/(3770+151) = 96.15\%$.

B Prompts

This appendix collects the prompts used at two stages of the work: (i) the caption-to-MCQ construction that produces the training-time reward of §3.2.1 (§B.1); and (ii) the cloze-to-MCQ conversion that translates Omni-Cloze blanks into self-contained five-option questions used in our evaluation pipeline (§B.2). All prompt blocks below are typeset verbatim from our released implementation; `{var}` markers denote placeholders filled at call time.

B.1 Caption-to-MCQ Construction (Training Reward)

At training time, the caption-only MCQ reward of §3.2.1 requires a pool $\mathcal{Q}_x$ of five-option multiple-choice questions derived from the gold caption y_{gold} of LACap-50k. The conversion is performed

offline by a frozen external text language model, once per gold caption, before training begins; the prompt below instructs the model to emit between 20 and 30 cloze-style MCQs spanning the seven audio analysis levels of §A.1, returning four model-generated options A–D per question (the "E: Not given" option is appended by our post-processor).

Caption-to-MCQ System Prompt

You are an expert audio comprehension MCQ generator. Given a detailed audio caption, generate 20–30 high-quality MCQs that together cover all seven audio analysis levels of §A.1. Each MCQ must test one specific, clearly answerable audio detail, with multiple questions per level wherever the caption provides evidence.

Seven audio analysis levels.

- `signal_level`: recording quality, mic technique, reverberation, background noise, dynamic range, mixing balance, stereo placement, artifacts.
- `perception_level`: loudness, pitch, timbre, rhythmic feel, speech clarity, foreground vs. background salience, perceived emotional intensity.
- `semantic_level`: exact words/phrases spoken (ASR-style), identifiable sound events, speaker turns, environmental sounds, relative ordering of audio events.
- `cultural_level`: speaker language and accent, musical genre tradition, social or situational context.
- `pragmatic_level`: communicative purpose, rhetorical style, speaker stance, whether the voice is human or synthetic.
- `musicological_level`: tempo, instrumentation, vocal style, music structure, production texture, genre.
- `sound_effect_level`: specific sound effect types, discrete event sounds, Foley-like action sounds.

Rules.

- Every question must be answerable solely by listening to the audio; the correct answer must be directly supported by the caption text.
- Generate four choices A–D only; the post-processor appends "E: Not given" automatically. Distractors must be realistic within the same category.
- Order-of-event questions must use relative terms ("first", "after", "last") and must not mention specific timestamp numbers.
- Each question is concise (max 25 words); `answer_text` must exactly match the chosen option.
- Output ONLY a valid JSON array; no markdown, no code fences, no reasoning.

Output format (per MCQ object).

```
{"level": "<level>",
 "subcategory": "<aspect>",
 "question": "<text>",
 "A": "<text>", "B": "<text>",
 "C": "<text>", "D": "<text>",
 "answer": "<A|B|C|D>",
 "answer_text": "<matches option>"}
```

B.2 Omni-Cloze to MCQ Conversion (Evaluation Preparation in Fig. 1a)

The Omni-Cloze benchmark (Ma et al., 2025) presents captions as cloze passages with placeholders such as `[BLANK_1]`, `[BLANK_2]`, . . . , each paired with five candidate options. To evaluate our LC-SFT-trained captioner against Omni-Cloze under the caption-as-evidence protocol of §4.1, we convert each blank into a self-contained natural-language question that can be answered by reading the caption text alone, while preventing the question from leaking answers to neighbouring blanks. The conversion is implemented as two complementary prompts: a per-blank prompt (used as the default and as the fallback for any sample-level failure), and a sample-level batched prompt (used for efficiency when many blanks share a passage). Visual-only blanks have their answers pre-substituted into the passage so that the LLM sees a fully grammatical context; only MCQ-relevant blanks (audio and audio-visual) remain as unknown `[BLANK_M]` markers.

Cloze-to-MCQ System Prompt (Per-Blank, Anti-Leakage)

You are an expert at creating video comprehension test questions. The passage contains blanks written as `[BLANK_1]`, `[BLANK_2]`, . . . , `[BLANK_30]`. Your task: write ONE concise question whose answer fills the TARGET blank specified in the prompt; every other `[BLANK_M]` is a completely unknown value.

Rules.

- Output ONLY the question text (no numbering, no explanation, no answer); one sentence, max 25 words.
- Do NOT state or imply the answer to the target blank in the question.
- Do NOT guess, infer, or mention specific values for any other `[BLANK_M]`: every other blank is unknown, never write a word you think fills it.
- Frame the question as something a viewer answers from the video using the modality hint provided.

Cloze-to-MCQ System Prompt (Sample-Level Batched)

You are an expert at creating video comprehension test questions. The passage contains blanks written as [BLANK_M]. Your task: write concise questions for the specified TARGET blanks only; every [BLANK_M] marker is an unknown value unless it appears as plain text.

Rules.

- Output ONLY valid JSON whose keys are blank numbers as strings; no markdown, no explanation.
- Each value is one natural question, one sentence, max 25 words.
- Do NOT state or imply the answer to a target blank in its own question.
- Do NOT guess, infer, or mention specific values for any other [BLANK_M].
- Frame each question as something a viewer answers from the video using the modality hint.

Output format.

```
{"3": "<question for blank 3>",
 "7": "<question for blank 7>"}
```

Two-layer anti-leakage protocol. The two prompts above are combined with a post-generation validation pass: for each generated question, we check whether any other MCQ blank's answer string appears verbatim (whole-word, case-insensitive); if so, the question is regenerated with an explicit forbidden-word constraint appended to the prompt, up to a fixed retry budget. Unresolved items are flagged with `_leakage_warning` for downstream filtering. Non-MCQ (visual-only) blanks have their answers pre-substituted into the passage rather than masked, since the LLM never sees a [BLANK_M] marker that it is not authorized to discuss; this also gives the LLM a fully grammatical reading context instead of a sentence riddled with placeholders.

C Method Details

C.1 Why Second-Order Curvature

We define $\mathrm{Curv}_{k,t}$ in Eq. 5 as the smoothed second-order difference of $U^{(\ell)}_{k,t}$ across late layers, rather than the first-order alternative $\sum_j |\Delta U^{(j)}_{k,t}|$. The choice is essential to isolating SEC. The second-order formulation evaluates to zero on any constant-slope trajectory (stable, linearly ascending, or linearly descending), and is large only on trajectories that exhibit a true kink between consecutive intermediate layers. The first-order alternative cannot make this distinction. It fires equally on a smoothly resolving common token, where each $\Delta U^{(j)}_{k,t}$ is small but nonzero, and on an abrupt late-layer commitment, where $\Delta U^{(j)}_{k,t}$ spikes at one position. Only the second-order statistic produces near-zero weight on smooth descents and concentrates suppression on the kinked trajectories that diagnose SEC.

C.2 Full Gradient Expression

Differentiating the group-level objective in Eq. 8 with respect to θ yields

$$\nabla_\theta \mathcal{L}_{\mathrm{LC}} \;=\; -\,\mathbb{E}_{x,\mathcal{G}_x}\left[\sum_{k=1}^{K} \frac{A(y_k)}{Z_k} \sum_{t=1}^{T_k} w_{k,t}\, \nabla_\theta \ell_{k,t}\right], \tag{12}$$

where $Z_k = \sum_t w_{k,t} + \epsilon$ is the per-candidate normalizer. The expression makes the two-axis reweighting explicit. The candidate axis is reweighted by $A(y_k)/Z_k$, which is positive for above-mean candidates (maximum-likelihood gradient pulling π_θ towards them) and negative for below-mean candidates (unlikelihood gradient pushing π_θ away). Within each candidate the token axis is reweighted by $w_{k,t}$, close to one for smoothly-resolved tokens and close to zero for SEC-prone tokens. The combined effect concentrates gradient mass simultaneously on high-reward candidates and on intermediate-layer-reliable tokens.

C.3 Two-Stage Training Details

Stage 1 applies standard cross-entropy SFT on $\mathcal{D} = \{(x, y^\star)\}$ where $y^\star$ is the LACap-50k gold caption for x. The model from this stage gains the basic capacity to produce paragraph-length captions for audio–visual clips. Stage 2 applies LC-SFT as defined in §3.2, initialized from the Stage 1 checkpoint. All rollouts $y_{r_1}, \ldots, y_{r_{K-1}}$ in the response group $\mathcal{G}_x$ are sampled from the current LC-SFT policy π_θ at every gradient step. The Stage 1 checkpoint is used only for initialization and is not consulted during Stage 2. The optimizer, learning rate, batch size, group size K, curvature suppression coefficient κ, GPU configuration, and total training steps for both stages are reported in §4.

C.4 Relation to the Training Reward

SCC-Verifier and the caption-only MCQ reward of §3.2.1 share the same verification primitive (audio-grounded multiple-choice question answering) but

differ in three respects, each dictated by the constraints of the lifecycle stage they operate at:

- **Source of MCQs.** At training, MCQs are derived from the data-stage SCC-verified gold caption y_{gold}; at inference, no gold caption is available, so MCQs are derived from each candidate $y^{(m)}$ itself.
- **Verifier identity.** At training, the verifier is a frozen external text language model that scores caption-text retention without listening to the audio, forcing the candidate caption to preserve gold-verified facts in its text. At inference, the verifier is the LC-SFT-trained model itself listening to the audio without seeing the candidate caption, forcing the candidate's own claims to be audio-grounded.
- **Role in the SCC framework.** Together with the data-stage modality-subtraction self-check (§3.1), the two complete SCC's audio-grounded verification primitive across all three lifecycle stages of a long-paragraph audio caption.

D Omni-Cloze Audio-Only Input Results

Each Omni-Cloze clip provides both an audio stream and a video stream. The main-text Table 4 reports accuracy on the audio subset of Omni-Cloze under the audio-visual input setting, in which the model receives both modalities. For completeness, we additionally evaluate the same audio subset under the audio-only input setting, in which the model receives the audio stream alone with no access to the video. This complementary setting isolates each model's audio-understanding capability from any visual disambiguation that the video stream might provide. Table 9 reports the results.

E LC-SFT Component Ablation

We isolate the two reweighting components of the LC-SFT objective (§3.2) in a 2×2 design on top of the Stage-1 SFT baseline. **SFT** is the Stage-1 standard SFT checkpoint on LACap-50k. **+Adv** applies only the response-level advantage reweight $A(y_k)$ from Eq. 3 while keeping $w_{k,t} \equiv 1$ (denoted LC-SFT$_R$). **+Tok** applies only the token-level layer-curvature reweight $w_{k,t}$ from Eq. 6 while keeping $A(y_k) \equiv 1$ (denoted LC-SFT$_T$). **LC-SFT** is the full objective combining both. All four variants are evaluated on the audio subset of Omni-Cloze under audio-visual input. Table 10 reports per-category correct accuracy (%) on all 47 Omni-Cloze audio-analysis categories, together with the overall accuracy aggregated across the full split.

Model	Acc (%) ↑
Proprietary Models	
GPT-4o Audio (OpenAI, 2024)	35.8
Gemini 2.0 Flash (Gemini Team, 2023)	20.0
Gemini 2.5 Flash (Gemini Team, 2023)	42.6
Gemini 2.5 Pro (Gemini Team, 2023)	48.0
Gemini3.1-Pro (Gemini Team, 2023)	64.1
Open-Source Models	
SALMONN-13B (Tang et al., 2024)	10.6
MiDashengLM-7B (Dinkel et al., 2025)	19.5
Qwen2-Audio-7B (Chu et al., 2024)	22.2
Qwen2.5-Omni-7B (Xu et al., 2025a)	25.8
Audio-Captioner-7B (Ma et al., 2025)	53.2
Qwen2.5-Omni-7B (Ours)	**59.3**

Table 9: **Results on Omni-Cloze (audio subset) under the audio-only input setting.** Each model receives only the audio stream, with no access to the video. The last row is our LC-SFT-trained model based on Qwen2.5-Omni-7B.

F Relation between SCC-Verifier and the Training Reward

SCC-Verifier (§3.3) and the caption-only MCQ reward of §3.2.1 share the same verification primitive, namely audio-grounded multiple-choice question answering, but operate at different lifecycle stages with complementary rather than conflicting roles.

At training time, MCQs are derived from the data-stage SCC-verified gold caption y_{gold}, and the verifier is a frozen external text language model that scores the candidate caption y_k on its text alone with no audio access. The reward thereby forces y_k to retain every audio fact already verified in y_{gold}, turning the training signal into a measure of verified audio-fact retention from the gold reference. The text-only verifier identity is deliberate. Allowing the training verifier to access the audio would let it bypass y_k and answer from perception alone, leaving the candidate text unused and the gradient uninformed about which facts the candidate actually retains.

At inference time, no gold caption is available, so MCQs are derived from each candidate $y^{(m)}$ itself. The verifier is the LC-SFT-trained model, queried in audio-only MCQ mode with the candidate's text deliberately withheld. The score thereby forces each candidate's own assertions to be audio-

Category	SFT	+Adv	+Tok	LC-SFT	Category	SFT	+Adv	+Tok	LC-SFT
Artificial Intelligence	59.41	61.19	**61.98**	60.20	Lectures & Talks	57.23	58.42	**61.58**	60.40
Arts & Crafts	65.00	63.00	**67.20**	66.20	Listicles & Rankings	59.80	**61.98**	60.20	60.79
Athlete Profiles	50.92	51.74	50.51	**53.99**	Livestreams	**56.77**	52.62	55.90	56.55
Board Games & Puzzles	**58.20**	56.80	56.20	55.40	Miscellaneous	57.38	53.48	55.94	**58.81**
Business & Finance	58.45	59.67	**59.88**	57.43	Motorsports	52.68	55.27	57.46	**58.45**
Car Modifications	62.00	57.09	**62.38**	61.06	Movies & Trailers	51.59	**51.98**	50.60	50.99
Car Reviews	59.51	59.51	**63.97**	57.49	Music	40.84	40.08	40.27	**41.60**
Collecting	61.13	60.53	**63.56**	61.94	News Reports	57.74	59.52	59.52	**62.30**
Comedy	49.60	**51.00**	49.00	49.40	Off-Roading	52.26	51.32	**53.01**	52.44
Crafts & DIY	60.04	58.56	**62.16**	60.47	Opinion & Commentary	56.47	58.43	57.84	**59.41**
Debates	**56.50**	55.92	55.92	55.92	Outdoor Activities	57.51	57.14	58.44	**60.85**
Driving Tutorials	58.88	61.57	63.84	**64.67**	Parenting & Family	56.75	60.76	**61.60**	59.28
Engineering	67.57	64.64	**67.78**	**67.78**	Photography & Film	51.47	52.05	51.47	**52.84**
Environmental Science	57.79	57.99	**61.07**	59.84	Science & Technology	**67.36**	66.32	67.15	66.74
Fashion & Beauty	60.51	59.72	61.89	**63.26**	Science Explainers	55.76	57.41	59.88	**60.29**
Fitness & Training	63.95	64.77	63.95	**69.45**	Space Exploration	62.98	64.89	**65.46**	63.93
Food & Cooking	60.86	60.45	61.27	**62.09**	Sports Commentary	57.36	**62.05**	61.19	61.62
Gaming	53.41	56.63	**56.83**	56.02	Tech Reviews	59.27	58.25	59.06	**60.49**
Health & Fitness	55.98	52.39	54.78	**57.97**	Travel	57.61	61.32	**63.37**	60.08
Highlights & Replays	49.41	50.00	51.19	**51.78**	Tutorials	59.88	59.46	**62.58**	62.16
History & Culture	57.03	53.66	55.45	**58.02**	Visual Arts	56.14	57.95	59.56	**59.76**
Home & Garden	60.29	61.51	62.73	**63.34**	Vlogs	**53.58**	50.10	50.10	50.87
Interviews	53.60	53.81	57.63	**59.53**	Writing & Literature	54.09	55.58	54.99	**57.73**
Language Learning	53.35	55.91	58.27	**60.24**					
Overall accuracy:	SFT 57.10	+Adv 57.28	+Tok 58.59	**LC-SFT** 58.92					

Table 10: **LC-SFT component ablation on Omni-Cloze** (audio subset, audio-visual input). Per-category correct accuracy (%) on all 47 audio-analysis categories for the four variants on top of Qwen2.5-Omni-7B. **SFT**: Stage-1 standard SFT on LACap-50k. **+Adv**: Stage-2 with response-level advantage reweight only ($A(y_k)$ from Eq. 3, $w_{k,t} \equiv 1$; denoted LC-SFT$_R$). **+Tok**: Stage-2 with token-level layer-curvature reweight only ($w_{k,t}$ from Eq. 6, $A(y_k) \equiv 1$; denoted LC-SFT$_T$). **LC-SFT**: full objective combining both reweighting signals. Per-row best is in bold; ties are bolded jointly.

grounded, with the trained model acting as a self-consistency oracle on the audio side. The audio-grounded verifier identity is again deliberate. Allowing the inference verifier to read the candidate caption would collapse the check into textual self-agreement and would not test whether the assertions correspond to the actual audio.

The two mechanisms are therefore complementary by construction. The training reward exploits the data-stage SCC verification (the gold caption is already audio-checked) to teach the policy what audio-grounded captioning looks like, using a text-only judge to keep the supervision focused on caption-side fact retention. The inference verifier exploits the resulting capability of the trained model itself, using audio-grounded self-answering to arbitrate among samples without any gold reference. Together with the data-stage modality-subtraction self-check (§3.1), the two complete SCC's audio-grounded verification primitive across all three lifecycle stages of a long-paragraph audio caption, with the verifier's modality tuned to what each stage requires.

G Experiment Setup

Backbone and training framework. We adopt Qwen2.5-Omni-7B (Xu et al., 2025a) as the captioner backbone for both stages of training. Training is implemented with the Swift framework using DeepSpeed ZeRO-3 on 8 GPUs with `bf16` mixed precision and gradient checkpointing. The random seed is fixed at 42.

Two-stage training and data split. The 50,222-clip LACap-50k corpus is partitioned into two disjoint training subsets. Stage 1 (standard SFT) uses 44k clips for cross-entropy supervised fine-tuning on the gold long-paragraph captions, while Stage 2 (LC-SFT) uses a separate 1k clips for on-policy refinement under the caption-only MCQ reward and the layer-curvature reweighting signal (§3.2). The two stages share the same optimisation configuration. Both use learning rate 1×10^{-6} with cosine schedule and warmup ratio 0.03, per-device batch size 1, gradient accumulation 2, 1 training epoch, and a maximum sequence length of 32,768 tokens with the delete truncation strategy. The only

differences between the two stages are the training objective (cross-entropy in Stage 1 versus the LC-SFT loss of Eq. 8 in Stage 2) and the respective data subset.

Multimodal input processing. Each clip is processed jointly as a synchronised video and audio stream, with the audio output head disabled so that the model produces text only. Video frames are sampled at 0.2 frames per second within a range of 2 to 32 frames per clip. The video token budget per clip is 64 to 128 tokens, and the image token budget is 4 to 512 tokens.

LC-SFT and SCC-Verifier hyperparameters. For Stage 2 LC-SFT, the response group size is $K = 4$ (one LACap-50k gold caption plus three on-policy rollouts), the curvature suppression coefficient is $\kappa = 2$, and the late-layer set $\mathcal{L}$ comprises the last $m = 3$ Thinker layers. For inference-time SCC-Verifier, we use $M = 4$ candidate captions (one greedy plus three multinomial samples at temperature $T = 0.8$, symmetric with the training-time group size K) and N verification MCQs per candidate. We do not fix N but instead extract every atomic audio-grounded claim that the external LLM can identify in the candidate caption, so longer captions naturally yield more verification MCQs and a more fine-grained per-candidate score.

Evaluation setup. For the caption-as-evidence protocol of §4.1, the frozen text-only judge LLM is Qwen3.6-27B (Qwen Team, 2026). Prior-method numbers on MMAU and MMAR are reproduced from Ma et al. (2025).

Baselines. We compare against representative proprietary and open-source audio-visual captioners across MMAU (Table 3), MMAR (Table 3), MMSU (Table 3), and Omni-Cloze (Table 4 for audio-visual input and Table 9 for audio-only input). The proprietary baselines are GPT-4o Audio (OpenAI, 2024) and the Gemini family of multimodal models (Gemini Team, 2023), evaluated at the 2.0 Flash, 2.5 Flash, 2.5 Pro and 3.1 Pro-Preview release tiers. The open-source baselines on audio-only inputs are SALMONN-13B (Tang et al., 2024), MiDashengLM-7B (Dinkel et al., 2025), Qwen2-Audio-7B (Chu et al., 2024), and Qwen2.5-Omni-7B (Xu et al., 2025a), the last of which also serves as our backbone. The open-source baselines on audio-visual inputs additionally include the video-grounded captioners video-SALMONN-13B (Sun et al., 2024a), videoLLaMA 2-7B (Zhang et al., 2023), video-SALMONN 2-7B (Tang et al., 2025), and the long-paragraph detailed captioner Omni-Captioner-7B (Ma et al., 2025) (also denoted Audio-Captioner-7B in the audio-only setting).

H Analysis Protocols Beyond Verifier-Format Alignment

Open-ended Clotho evaluation. SFT and LC-SFT receive audio only and generate one caption for every clip in the complete 1,045-clip Clotho evaluation split (Drossos et al., 2020), using the same captioning prompt, deterministic decoding, and a 512-token limit. Generation and evaluation contain no questions, answer options, caption-only judge, SCC-Verifier, or MCQ reranking. Because the generated paragraphs are much longer than Clotho's five short references, Best- and Mean-Chunk CLAP (Wu et al., 2023) score consecutive semantic chunks of at most 30 words.

RefAlign-F1 (ours). We split each generated caption into ordered sentence or clause units $C = \{c_i\}_{i=1}^{m}$ and denote its five references by $R = \{r_j\}_{j=1}^{n}$, where $n = 5$. Splitting follows punctuation and line breaks; overlong units are further divided into consecutive word chunks. We encode all units and references with BGE-M3 (Chen et al., 2024), L_2-normalise the embeddings, and define $S_{ij} = \max\{0, \cos(e(c_i), e(r_j))\}$. With word-count weights $w_i = \max\{1, \mathrm{words}(c_i)\} / \sum_k \max\{1, \mathrm{words}(c_k)\}$, alignment precision and recall are $P = \sum_i w_i \max_j S_{ij}$ and $R = n^{-1} \sum_j \max_i S_{ij}$. RefAlign-F1 is $2PR/(P + R)$ when $P + R > 0$ and zero otherwise; we report the arithmetic mean of per-clip scores. This score measures soft alignment to the five references; audio grounding is evaluated separately. For SBERT and FENSE, BGE-M3 selects the complete generated sentence with the highest mean reference similarity before scoring.

Cross-family judge control. All captioners first generate captions from synchronised audio–visual input. We freeze these captions, after which Qwen3.6-27B and Gemini3.1-Pro each receive the identical caption, question, and answer options without audio/video access. Judge prompts, temperature (0), and answer parsing are fixed, so the only changed variable is judge identity.