TEMPO: Temporally-grounded Multi-task Post-training for Large Audio-Language Models

Anonymous ACL submission

Abstract

Large audio-language models (LALMs) describe audio at the clip level but cannot assign timestamps to the events, speakers, or sounds they identify. Despite being essential for downstream tasks like speech recognition and dense audio captioning, timestamping remains a key limitation of most LALMs. We present TEMPO (**Te**mporally-grounded **M**ulti-task **Po**st-training), the first unified model to handle audio, speech and music timestamping tasks. Our core contribution is a supervised fine-tuning (SFT) stage built on three innovations: atomic timestamp tokens, a time-aware projector that injects sinusoidal wall-clock encodings into audio frame embeddings, and a distance-aware Gaussian loss. Our training is based on a synthetic-to-real curriculum. We further present the first application of reinforcement learning to this problem, using Group Relative Policy Optimization (GRPO) with verifiable rewards to train directly against the metrics used at evaluation. To support this work, we build a training dataset containing 119K samples and an evaluation benchmark containing 10K samples, drawn from six established corpora across five tasks. On this benchmark, TEMPO outperforms Audio Flamingo Next and Qwen3-Omni, two state-of-the-art LALMs explicitly trained on timestamped data. Experiments confirm that SFT delivers most of these gains, with GRPO providing consistent but moderate refinements.

1 Introduction

Audio signals carry rich temporal information. Answering *what happens when* in an audio stream is fundamental to applications such as meeting transcription and music analysis, which require associating specific events with precise time intervals. Yet most existing Large Audio-Language Models (LALMs) treat audio understanding as a clip-level task, mapping a recording to a single caption or answer and discarding temporal structure (Ghosh et al., 2024; Goel et al., 2025; Kong et al., 2024; Ghosh et al., 2025b). Timestamping is therefore a foundational capability that current LALMs lack.

In this work, we focus on five timestamping tasks that span speech, music and audio: multi-speaker automatic speech recognition (ASR), speaker diarization, audio temporal grounding, dense audio captioning and timestamped music captioning. Despite differing in their outputs, all five share a common structure: each requires segmenting an audio into intervals and assigning a textual label to each one. Recent LALM-based approaches such as TAC (Kumar et al., 2026) and TimeAudio (Wang et al., 2025) address timestamping within a unified model but remain restricted to limited tasks.

We present TEMPO, the first unified LALM for timestamping across audio, speech, and music. TEMPO handles all five tasks with a single decoder, emitting text with timestamp tokens distinguished only by a task tag in the prompt. Our supervised fine-tuning (SFT) recipe builds on three components: atomic timestamp tokens, a time-aware multi-modal projector that injects sinusoidal wall-clock encodings into audio frame embeddings and a distance-aware Gaussian loss. Training follows a two-stage synthetic-to-real curriculum. Building on this, we present the first application of reinforcement learning to timestamped audio understanding, using Group Relative Policy Optimization (GRPO) (Guo et al., 2025) with verifiable rewards to train directly against the temporal metrics used at evaluation.

On a benchmark constructed from six established corpora, TEMPO substantially outperforms Audio Flamingo Next and Qwen3-Omni, two state-of-the-art LALMs explicitly trained on timestamped data. Gains are largest on speech: TEMPO cuts WER from 0.70 to 0.44 on multi-speaker ASR and improves diarization mIoU from 0.44 to 0.71 over Qwen3-Omni, with consistent improvements across audio and music tasks as well. Experiments

confirm that SFT delivers most of these gains, with GRPO providing consistent but moderate refinements, suggesting that careful post-training design is the primary lever for adding timestamping capabilities to LALMs. Our main contributions include:

- **First unified LALM for timestamping across speech, sound, and music.** TEMPO performs speaker diarization, multi-speaker ASR, audio temporal grounding, dense audio captioning, and timestamped music captioning in a single model, outperforming Qwen3-Omni, Audio Flamingo Next and other LALM-based timestamping baselines.

- **A post-training SFT recipe that teaches timestamping from scratch.** Three innovations, namely atomic timestamp tokens at 0.1 s resolution, a time-aware multi-modal projector with sinusoidal wall-clock encodings, and a distance-aware Gaussian loss, together teach a model with no prior temporal supervision to predict accurate timestamps.

- **First unified RL approach across speech, audio, and music timestamping.** We apply GRPO with temporal verifiable rewards, improving over the SFT baseline on all five tasks.

- **Multi-task benchmark and training corpus.** 119K training examples (51K synthetic, 68K real-world) and 10K evaluation examples drawn from six established corpora, covering all five timestamping tasks under a unified output format.

2 Related Work

2.1 Large Audio-Language Models and Timestamping

Recent LALMs including SALMONN (Tang et al., 2024), Qwen2-Audio (Chu et al., 2024), the Audio Flamingo series (Kong et al., 2024; Ghosh et al., 2025b; Goel et al., 2025; Ghosh et al., 2026), GAMA (Ghosh et al., 2024), and Kimi-Audio (KimiTeam et al., 2025) have demonstrated strong capabilities on captioning, question answering, and open-ended reasoning, but temporal grounding remains a persistent weakness. The closest prior work, TAC (Kumar et al., 2026) and TimeAudio (Wang et al., 2025), address timestamping within a unified model but are restricted to limited tasks and rely on supervised training. TEMPO extends timestamped generation to five tasks spanning speech, sound, and music.

2.2 Speech: Diarization and Multi-Speaker ASR

Diarization has evolved from clustering pipelines to end-to-end and target-speaker models; multi-speaker ASR extends this with joint transcription via serialized output training. Recent LLM-based methods include DiarizationLM (Wang et al., 2024), which refines outputs via post-processing and SpeakerLM (Yin et al., 2026) jointly models speaker and language. Unlike these speech-only systems, TEMPO jointly predicts *who*, *what*, and *when* within a general-purpose framework trained across non-speech tasks as well.

2.3 Sound: Temporal Grounding and Dense Captioning

Audio grounding localizes a text query to a time interval, while dense captioning requires segmenting an audio stream and describing each event with timestamps. Standard benchmarks like AudioCaps (Kim et al., 2019) and Clotho (Drossos et al., 2019) provide only clip-level annotations, inducing *semantic collapse* where distinct events are compressed into global summaries (Kumar et al., 2026). Recent efforts address this via synthetic dense annotation pipelines and frame-level temporal indexing (Wang et al., 2025).

2.4 Music: Chord Estimation and Captioning

Automatic chord estimation traditionally combines acoustic classifiers with harmonic language models, evaluated on McGill Billboard (Burgoyne et al., 2011) using MIREX-standard metrics, but these systems are specialized classifiers that do not generalize beyond chord recognition. Music Flamingo (Ghosh et al., 2025a) introduces rotary time embeddings for temporal localization and produces theory-aware captions that may reference chord progressions in natural language, but does not generate structured, timestamped chord labels. To our knowledge, no existing audio-language model produces structured, timestamped chord progressions from audio; we integrate this as one of five tasks in TEMPO.

2.5 Reinforcement Learning for LALMs

The DeepSeek-R1 pipeline (Guo et al., 2025), which combines an SFT cold start with Group Relative Policy Optimization (GRPO) (Shao et al.,

Task	Question	Ground Truth Response
ASR	Provide a timestamped transcription of the speech in this audio.	[speech:asr] <|0.0|> The memory uh <|2.2|> [speech:asr] <|3.4|> Pops up the options <|4.5|> [speech:asr] <|4.9|> Yeah yeah that would be possible yeah sure <|7.1|> [speech:asr] <|7.5|> I th dont think thats uh that takes a lot of storage space . . . <|14.0|>
Diarization	How many speakers are in this audio and when does each one speak? Include timestamps.	[speech:diar] <|0.0|> Speaker 1 <|6.2|> [speech:diar] <|4.6|> Speaker 2 <|6.1|> [speech:diar] <|5.9|> Speaker 3 <|16.9|> [speech:diar] <|13.5|> Speaker 1 <|27.9|>
Audio Grounding	What is the time interval (start and end) for the query 'a bird tweets quietly and repeatedly outdoors' in the audio?	[audio:ground] <|0.9|> to <|1.4|> [audio:ground] <|2.3|> to <|2.8|> [audio:ground] <|3.7|> to <|4.3|> [audio:ground] <|5.2|> to <|5.7|>
Dense Audio Captioning	Which sound events occur, and what are their time intervals and descriptions?	[audio:caption] <|0.0|> A river flows loudly outdoors. <|23.9|> [audio:caption] <|0.9|> A bird tweets quietly and repeatedly outdoors. <|1.4|> [audio:caption] <|2.3|> A bird tweets quietly and repeatedly outdoors. <|2.8|> [audio:caption] <|3.7|> A bird tweets quietly . . . <|4.3|>
Music Captioning	Caption this music with timestamps. Describe the instrument, tempo, chord progression, dynamics, and note statistics.	[audio:music][instrument] Organ enters at <|0.0|> exits at <|30.0|> [tempo] 120.0 BPM at <|0.0|> [tempo] 125.0 BPM at <|9.5|> [chord] B:min from <|0.0|> to <|3.5|> [chord] D:sus2 from <|3.5|> to <|21.0|> [chord] C:sus2 from <|21.0|> to <|28.5|> [chord] D:min7 from <|28.5|> to <|29.0|> [chord] G:sus2 from <|29.0|> to <|29.5|> [chord] D:sus2 from <|29.5|> to <|30.0|> [stats] Note density:0.6 notes/sec, range: D5 to G6

Table 1: Example instances from each evaluation task. Ground truth responses are truncated for brevity.

2024) and verifiable rewards, has proven effective for text and vision-language models. Recent audio work has begun to apply RL post-training to LALMs: Audio-Thinker (Wu et al., 2025) uses GRPO to guide audio reasoning while R1-AQA (Li et al., 2025) explores RL post-training for audio QA. However, these efforts target reasoning or classification accuracy and optimize text-level metrics. We are the first to apply GRPO with temporal verifiable rewards across a unified multi-task LALM.

3 Tasks and Datasets

We train TEMPO on five timestamping tasks spanning speech, sound, and music using a combination of real and synthetic data. Table 1 shows example inputs and outputs for each task.

3.1 Real Data

Multi-Speaker ASR and Diarization. We use AMI (Carletta et al., 2005), ICSI (Shriberg et al., 2004), and Switchboard (Godfrey et al., 1992), three conversational corpora with speaker-labeled transcripts and precise timestamps, enabling evaluation of both joint transcript-timing prediction and speaker turn segmentation, yielding 16,045 ASR and 16,053 diarization samples.

Audio Temporal Grounding. We use AudioSet Strong (Hershey et al., 2021), a strongly-labeled subset of AudioSet (Gemmeke et al., 2017) that provides onset and offset times for over 500 sound classes. To convert the categorical event labels into natural-language queries, we prompt GPT-4o to generate paraphrased queries for each class, yielding 13,500 training samples.

Dense Audio Captioning. We directly use TACOS (Primus et al., 2025), which pairs audio recordings with captions and onset/offset timestamps, yielding 9,858 data points for this task.

Timestamped Music Captioning. We use Slakh2100 (Manilow et al., 2019), a collection of 2,100 multi-track songs synthesized from MIDI, from which we derive timestamped chord, tempo, and instrument annotations via the aligned MIDI files, yielding 13,000 samples.

In total, the real-world corpus contains 68,456 training examples across all five tasks.

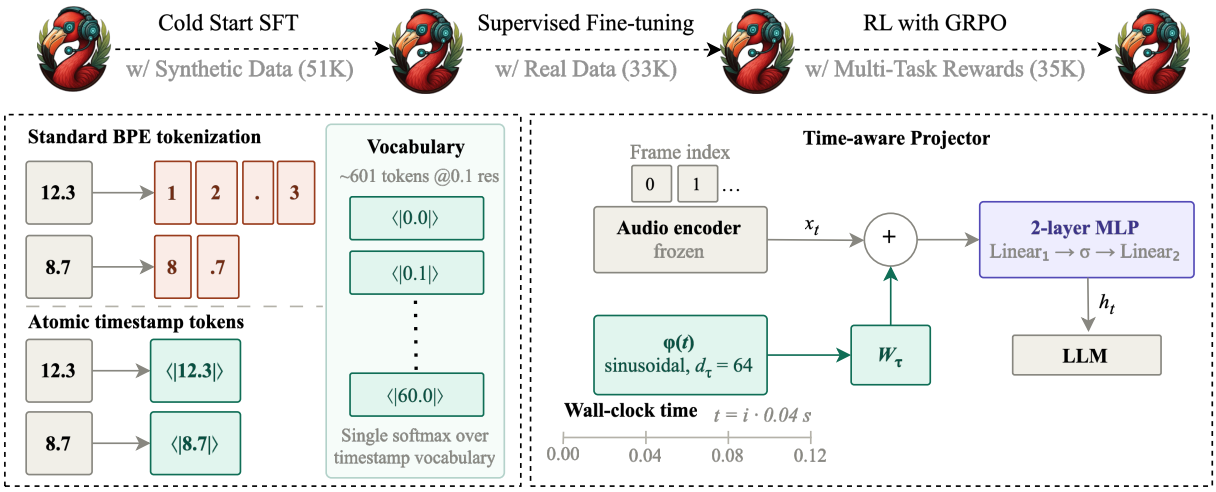

Figure 1: Overview of TEMPO, built on Audio Flamingo 3. **Top:** Three-stage curriculum from synthetic SFT (51K) through real-data SFT (33K) to GRPO with verifiable rewards (35K). **Bottom-left:** Atomic timestamp tokens replace BPE fragmentation with ∼601 dedicated tokens at 0.1 s resolution. **Bottom-right:** Time-aware projector injects sinusoidal wall-clock encodings $\varphi(t)$ into frozen audio encoder features before the MLP.

3.2 Synthetic Data

To improve temporal calibration before real-data training, we also construct a synthetic Stage-1 corpus from LibriSpeech (Panayotov et al., 2015) and ESC-50 (Piczak, 2015) covering ASR, diarization, DAC, and AG tasks, with timestamp tokens balanced uniformly across all possible values. We generate 8,783 ASR samples, 8,783 diarization samples, 11,962 AG samples and 11,984 DAC samples. We add 10,000 samples from Slakh2100 to this stage to ensure exposure to all tasks. This yields 51,512 training examples in total. Full details are in Appendix A.

4 Method

TEMPO is built on Audio Flamingo 3 (AF3) (Goel et al., 2025), which couples a Whisper-large audio encoder to a Qwen2-7B language model through a two-layer MLP multi-modal projector. The audio encoder is frozen throughout training, and the five tasks are distinguished by task-specific tags in the prompt: `[speech:asr]`, `[speech:diar]`, `[audio:caption]`, `[audio:ground]`, and `[audio:music]`. Figure 1 gives an overview of TEMPO's architecture and three-stage training pipeline.

4.1 Supervised Fine-Tuning

We begin with a supervised fine-tuning (SFT) stage that teaches the model the timestamped output format and provides the temporal calibration needed before reinforcement learning. A standard teacher-forced cross-entropy objective is insufficient for fine-grained timestamping for three reasons. First, numeric timestamps rendered as text are tokenized inconsistently by the BPE vocabulary, fragmenting supervision across subword pieces. Second, the AF3 audio encoder produces features indexed by frame rather than wall-clock time, requiring the model to learn the correspondence implicitly. Third, token-level cross-entropy penalizes all incorrect timestamps equally, which is misaligned with the downstream temporal IoU (tIoU) metric. Our SFT recipe addresses each of these in turn through a corresponding architectural or objective change: atomic timestamp tokens, a time-aware multi-modal projector, and a distance-aware Gaussian loss.

4.1.1 Atomic Timestamp Tokens

To prevent BPE from fragmenting timestamps across subword pieces, we extend the tokenizer with atomic timestamp tokens of the form $\langle|t|\rangle$, where $t \in \{0.0, 0.1, \ldots, 60.0\}$ at 0.1 s resolution. Each timestamp is represented by a single token, so a timestamp prediction becomes a single categorical decision over approximately 600 candidates. New token embeddings are initialized as the mean of the BPE decomposition of the corresponding numeric value.

4.1.2 Time-Aware Multi-Modal Projector

To supply the wall-clock information that frame-indexed features lack, we replace the AF3 projector, whose positional embeddings encode frame index rather than elapsed time, with one that injects sinusoidal wall-clock encodings into each frame em-

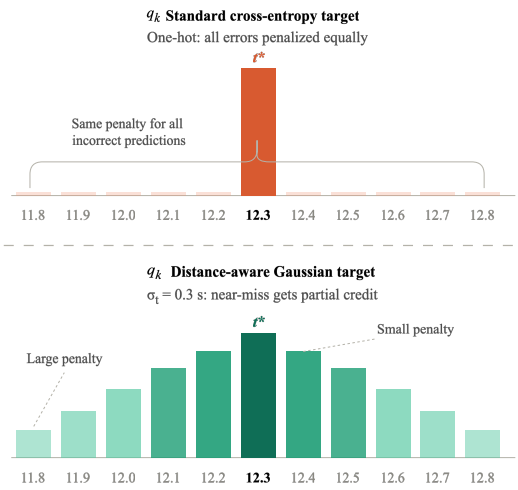

Figure 2: Standard cross-entropy (top) uses a one-hot target, penalizing all errors equally. Our distance-aware Gaussian target (bottom, $\sigma_t = 0.3$ s) assigns partial credit to near-miss predictions, encouraging the model to learn the ordinal structure of the timestamp vocabulary.

bedding:

$$\mathbf{h}_t = \mathrm{Linear}_2\big(\sigma(\mathrm{Linear}_1(\mathbf{x}_t + W_\tau \phi(t)))\big), \quad (1)$$

where $\mathbf{x}_t$ is the t-th audio frame, $\phi(t)$ is a fixed sinusoidal encoding of wall-clock time $t = i \cdot 0.04$ s, and W_τ is a learned projection initialized from $\mathcal{N}(0, 0.01^2)$. We use $d_\tau = 64$ with frequencies logarithmically spaced between periods of 0.08 s and 60 s, so that a single projector can span both sub-second acoustic events and long-range musical structure without task-specific parameterization.

4.1.3 Distance-Aware Timestamp Loss

Token-level cross-entropy treats every wrong timestamp as equally wrong, discarding the ordinal structure of the timestamp vocabulary and decoupling the training signal from the tIoU metric, under which a near miss is far better than a distant one. We restore this structure with an auxiliary loss that places a Gaussian soft label over the timestamp vocabulary, centered at the ground-truth time (Figure 2). If the target is $t^\star$ and the timestamp vocabulary corresponds to times $\{t_k\}_{k=1}^{K}$, the soft-label distribution is

$$q_k \propto \exp\left(-\frac{(t_k - t^\star)^2}{2\sigma_t^2}\right), \qquad \sum_{k=1}^{K} q_k = 1. \quad (2)$$

The auxiliary loss is

$$\mathcal{L}_{\text{time}} = -\sum_k q_k \log p_k, \quad (3)$$

where p_k is the model probability over the timestamp vocabulary obtained by restricting the output logits to timestamp tokens and applying a softmax. The overall training objective is

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda_{\text{time}} \mathcal{L}_{\text{time}} \quad (4)$$

with $\lambda_{\text{time}} = 0.5$ and $\sigma_t = 0.3$ s, where $\mathcal{L}_{\text{CE}}$ is the standard cross-entropy over all output positions including timestamps.

4.1.4 Two-Stage Synthetic-to-Real Curriculum

The three components above define *what* the model learns to predict; the curriculum determines the order in which it sees supervision. Stage 1 trains on the 51,512 synthetic examples described in Section 3 to establish temporal calibration before exposure to real data. Stage 2 then fine-tunes the merged Stage 1 checkpoint on 32,726 real-world examples taken from the real-world corpus.

Implementation details. During SFT, LoRA adapters (r=128, α=256, dropout 0.1) are attached to the language-model attention and MLP projection matrices, while the multi-modal projector and timestamp token embeddings are trained directly. Stage 1 runs for two epochs at learning rate 1×10^{-4} with 150 warmup steps; Stage 2 runs for two additional epochs at learning rate 5×10^{-5} with 50 warmup steps.

4.2 Reinforcement Learning with GRPO

Starting from the SFT checkpoint, we apply Group Relative Policy Optimization (GRPO) (Guo et al., 2025) with verifiable rewards designed to directly optimize the temporal metrics used at evaluation. For each task, we construct a reward that averages the eval metrics themselves, so that maximizing reward and improving eval performance are aligned. All rewards are bounded in $[0, 1]$ and gated by a binary format check that returns zero if the output does not parse into the expected task tag with valid timestamp tokens.

ASR and diarization. Both tasks use an equally-weighted average of three eval metrics. For ASR, the reward averages $(1 - \mathrm{WER})$, symmetric mIoU between predicted and ground-truth segments, and a boundary score $\exp(-\mathrm{MAE}/\tau)$ (with $\tau = 1.0$ s) computed over greedy-matched pairs with IoU above 0.3. For diarization, the same structure averages $(1 - \mathrm{DER})$, symmetric mIoU, and a soft speaker-labeled F1@IoU0.5 that requires exact speaker-label matches. Because both tasks frequently produce overlapping segments, we scale

this averaged reward by a soft factor $(0.8 + 0.2 \cdot p)$, where p is the fraction of non-overlapping output time.

Dense audio captioning. The reward averages three equally-weighted terms: soft event F1@IoU0.5, symmetric mIoU, and a per-pair METEOR computed over greedy-matched predicted and ground-truth segments. This rewards both correct localization of sound events and accurate textual descriptions of each, jointly aligning the model with the three eval metrics rather than treating them as independent objectives.

Audio grounding. The reward averages soft F1@IoU0.5, symmetric mIoU, and a boundary score with $\tau = 2.0$ s (larger than ASR's $\tau = 1.0$ s to accommodate the naturally noisier boundaries of environmental sound events). A light count-ratio term (weight 0.25 out of 3.25 total, or 7.7%) discourages large mismatches between predicted and ground-truth interval counts. The weight is intentionally small because precision and recall in the F1 term already penalize over- and under-prediction; the count ratio acts as a soft regularizer rather than a hard gate.

Music captioning. The reward is a weighted sum of five terms gated by a format check (weight 0.05) that verifies all required section tags. The dominant component is chord accuracy (weight 0.45): each ground-truth chord is scored against its best-overlapping prediction as $0.3 \cdot \mathrm{tIoU} + 0.7 \cdot \mathrm{sim}$, where similarity weights root pitch (0.60), quality (0.25), and extension (0.15), with a false-positive penalty $\big(0.7 + 0.3 \cdot \min(|G|/|P|,\, 1)\big)$. Instrument accuracy (weight 0.20) averages name matching with entry/exit timing error under a 2.0 s tolerance. Tempo accuracy (weight 0.15) combines $0.7 \cdot \mathrm{BPM\ accuracy} + 0.3 \cdot \mathrm{timestamp\ accuracy}$ with the same tolerance. Statistics accuracy (weight 0.15) averages relative note-density error and exact pitch-range matching. The chord boundary tolerance is relaxed to 1.0 s (from 0.5 s) to reduce reward sparsity early in training.

Implementation details. GRPO training is initialized from the Stage 1+2 SFT checkpoint. We attach LoRA adapters (r=256, α=512) to the language model and train for $1{,}000$ steps at learning rate 1×10^{-5}. Each step generates 8 completions per prompt at temperature 1.2, with a maximum completion length of 768 tokens. The KL coefficient is set to $\beta = 0.01$.

5 Experiments and Results

We evaluate TEMPO on held-out test sets across all five tasks, using metrics that capture both the textual and temporal aspects of each output.

5.1 Evaluation Benchmarks

We evaluate TEMPO on held-out test sets across all five tasks. For dense audio captioning, we use the official TACOS test set: 2,000 question-answer pairs in which each query asks the model to caption all sound events in a clip along with their temporal boundaries. For audio temporal grounding, we reformat the TACOS test set into the grounding format used by TimeAudio (Wang et al., 2025), in which each strong caption becomes a natural-language query and its associated segment becomes the ground-truth interval, yielding 5,151 question-answer pairs. For speaker diarization and multi-speaker ASR, we evaluate on the AMI Meeting test set, segmented into chunks of up to 60 seconds for consistency with our training setup; this yields 1,181 question-answer pairs per task, with the same audio chunks used for both and the expected output differing only in whether transcription is required. For timestamped music captioning, we evaluate on the Slakh2100 test set (225 tracks), split into 30-second segments to obtain 1,000 question-answer pairs, where each segment requires the model to produce timestamped chord labels, instrument entry and exit points, and tempo.

5.2 Evaluation Metrics

We report standard metrics for each task; full computational definitions are given in Appendix B.

Speech metrics. For multi-speaker ASR we report *Word Error Rate (WER)*, computed over predicted and ground-truth segments matched by temporal IoU (Appendix B). For diarization we report *Diarization Error Rate (DER)* using the standard `pyannote.metrics` implementation (Bredin, 2017) with a 0.25 s collar; speaker labels are optimally permuted between hypothesis and reference before scoring, so a model that produces consistent but arbitrarily named speakers is not penalized for the naming alone.

Mean Intersection-over-Union (mIoU). For all tasks that produce time intervals, we report a *symmetric* mIoU computed as the harmonic mean of recall-side and precision-side mean-best-IoU. Unlike the more common recall-only mIoU, the sym-

metric formulation prevents over-prediction from inflating the score, since a model that emits many spurious intervals is penalized on the precision side.

F1 at IoU threshold and boundary MAE. For dense audio captioning and audio grounding we report event *F1@IoU0.5*, counting a predicted interval as a true positive if its IoU with a greedily-matched ground-truth interval is at least 0.5. For diarization we additionally report a *speaker-labeled F1@IoU0.5* that requires exact speaker-label agreement in addition to temporal overlap. For tasks with continuous time boundaries (ASR, grounding, dense captioning, music) we report boundary *MAE* in seconds, computed over greedy-matched pairs with $\mathrm{IoU} \geq 0.3$.

Caption and music metrics. For dense audio captioning we additionally report *METEOR* over matched segment pairs. For music captioning we report chord root and quality accuracy, chord F1@IoU0.5 and chord mIoU, instrument-name F1 and instrument boundary MAE, and tempo accuracy within a 4% tolerance. Full definitions of all caption and music metrics are deferred to Appendix B.

5.3 Compared Methods

We compare against three baseline families and ablations of our own pipeline. **Zero-shot LALMs** (Audio Flamingo 3, Audio Flamingo Next and Qwen3-Omni) are evaluated without task-specific training. **TimeAudio** (Wang et al., 2025), the closest prior open-source LALM-based approach, is run on the two tasks it supports (dense captioning and grounding). **Naive SFT** uses our two-stage curriculum without the three SFT innovations, isolating the recipe from the data. *Stage 1*, *Stage 2*, and *Stage 1+2* report the SFT model after each curriculum stage. Both **RL variants** start from the Stage 1+2 checkpoint: *single-task RL* runs a separate GRPO stage per task on that task's data and rewards, while *multi-task RL* runs one joint GRPO stage over all tasks. All results are shown in Table 2.

5.4 Main Results

Multi-speaker ASR. Existing LALMs perform poorly on multi-speaker ASR with timestamps under zero-shot evaluation. Qwen3-Omni performs best among baseline LALMs (WER of 69.7%) but still fails to align transcripts to accurate temporal boundaries (MAE of 1.12 s, mIoU of 31.6%). TEMPO with full SFT (Stage 1+2) achieves a WER of 44.7%, a 25-point absolute improvement over Qwen3-Omni, alongside the boundary MAE of 0.77 s and mIoU of 63.8%. Adding multi-task RL on top of the SFT checkpoint yields further improvements: WER drops to 43.5%, MAE to 0.76 s, and mIoU rises to 65.8%.

Speaker diarization. Qwen3-Omni achieves a reasonable DER of 44.2% but speaker-labeled F1 remains low at 31.5%. TEMPO reduces DER to 25.4% and raises speaker-labeled F1 to 54.9% and mIoU to 70.5%. The Stage 1+2 curriculum provides a clear gain over Stage 2 alone (54.9% vs 54.5% sF1, 70.5% vs 70.0% mIoU), validating the synthetic-to-real curriculum design. Multi-task RL further improves sF1 to 56.3% and mIoU to 71.1% at the same DER.

Dense audio captioning. TEMPO surpasses Qwen3-Omni on both metrics with 58.5% eF1 and 68.3% mIoU, while matching it closely on METEOR (22.3% vs 19.0%). TimeAudio, the closest prior LALM-based approach, achieves 43.6% eF1 and 64.6% mIoU on the same evaluation, which TEMPO exceeds by 14.9 and 3.7 absolute points respectively. Multi-task RL pushes performance further to 59.3% eF1 and 68.5% mIoU.

Audio temporal grounding. Audio grounding shows the largest gap between zero-shot baselines and trained models. Audio Flamingo 3 and Audio Flamingo Next achieve F1@IoU0.5 of only 3.3% and 3.6%, respectively, indicating that these models lack the ability to localize natural-language queries to specific time intervals. Qwen3-Omni performs better at 47.4% F1 but boundary MAE remains high at 1.10 s. TEMPO with full SFT reaches 46.2% F1 with boundary MAE of 1.19 s and mIoU of 48.5%. TimeAudio achieves 40.1% F1 and 46.3% mIoU on the same eval, which TEMPO exceeds by 6.1 and 2.2 absolute points. Multi-task RL improves F1 to 46.5% and mIoU to 49.4%.

Timestamped music captioning. On Slakh2100, the zero-shot baselines fail almost completely on chord-related metrics: all three models produce chord F1@IoU below 1% and chord root accuracy below 2%, indicating that none has learned to associate harmonic events with precise timestamps. Tempo accuracy is also poor for the LALM baselines (18.0% for Audio Flamingo 3, 40.3% for Qwen3-Omni). TEMPO substantially closes these

Table 2: Performance across all evaluation tasks. **Speech:** WER: word error rate (%); DER: diarization error rate (collar=0.25s) (%). **Audio:** eF1: event F1@IoU (%); MET: METEOR (%); F1: F1@IoU0.5 (%). **Music:** Root/Qual: chord root/quality accuracy; Ch.F1: chord F1@IoU0.5; Inst F1: instrument name F1; Tempo: tempo accuracy within 4% tolerance. **Shared:** MAE: boundary MAE (s, over matches with IoU≥0.3); mIoU: symmetric mean IoU (%); sF1: speaker-labeled F1@IoU (%). ↓: lower is better; ↑: higher is better.

Model	**ASR**			**Diarization**			**Dense Captioning**			**Audio Grounding**			**Chord**				**Instrument**		**Tempo**
	MAE↓	mIoU↑	WER↓	DER↓	mIoU↑	sF1↑	eF1↑	mIoU↑	MET↑	F1↑	MAE↓	mIoU↑	Root↑	Qual↑	F1↑	mIoU↑	F1↑	MAE↓	Acc↑
Audio Flamingo 3	0.99	19.4	209.7	103.2	2.9	1.0	3.8	5.7	2.7	3.3	1.25	5.6	0.004	0.009	0.001	0.021	0.337	2.84	0.180
Audio Flamingo Next	1.39	11.5	115.4	106.9	12.4	7.6	11.6	22.3	11.9	3.6	0.95	7.0	0.004	0.012	0.000	0.032	0.335	2.77	0.262
Qwen3-Omni	1.12	31.6	69.7	44.2	44.4	31.5	57.9	65.4	19.0	47.4	1.10	49.3	0.011	0.028	0.005	0.069	0.254	2.51	0.403
TimeAudio	–	–	–	–	–	–	43.6	64.6	19.9	40.1	1.32	46.3	–	–	–	–	–	–	–
SFT (Naive) - Stage 1 + 2	0.84	64.0	46.0	27.4	67.2	52.3	52.4	64.5	22.7	37.1	1.44	40.3	0.003	0.003	0.003	0.003	0.005	3.95	0.378
SFT (Ours) - Stage 1	2.42	41.7	94.0	74.3	43.3	25.1	47.6	50.9	10.5	34.3	3.10	36.2	0.194	0.324	0.058	0.755	0.964	2.71	0.475
SFT (Ours) - Stage 2	0.85	64.8	47.0	25.1	70.0	54.5	55.2	67.8	20.2	44.6	1.14	44.7	0.174	0.319	0.046	0.760	0.894	2.78	0.417
SFT (Ours) - Stage 1 + 2	0.77	63.8	44.7	25.4	70.5	54.9	58.5	68.3	22.3	46.2	1.19	48.5	0.196	0.327	0.062	0.762	0.963	2.63	0.473
SFT + Single-task RL	0.84	66.1	46.7	26.8	69.2	54.0	58.6	67.8	24.4	45.5	1.21	48.0	0.190	0.322	0.060	0.755	0.963	2.63	0.470
SFT + Multi-task RL	0.76	65.8	43.5	25.4	71.1	56.3	59.3	68.5	23.5	46.5	1.32	49.4	0.187	0.317	0.055	0.753	0.960	2.62	0.469

gaps: Stage 1+2 reaches 19.6% chord root accuracy, 32.7% chord quality accuracy, 6.2% chord F1@IoU0.5, and 76.2% chord mIoU, alongside 96.3% instrument F1 and 47.3% tempo accuracy.

5.5 Ablation Study

Effect of the synthetic-to-real curriculum. Comparing Stage 1 alone and Stage 1+2 reveals the contribution of the two-stage curriculum. Stage 1 trained on synthetic data alone is insufficient for downstream tasks (94.0% WER for ASR, 74.3% DER for diarization, 34.3% F1 for grounding). Stage 2 fine-tuning on real annotated data brings dramatic improvements across the board, with the largest gains on audio grounding (F1 from 34.3% to 46.2%) and ASR (WER from 94.0% to 44.7%). Stage 1+2 also outperforms Stage 2 alone in most cases (0.77 s vs 0.85 s ASR MAE, 70.5% vs 70.0% diarization mIoU), indicating that the synthetic warmup provides priors that transfer beneficially to real data.

Effect of the full SFT recipe. On the full Stage 1+2 corpus, the Naive SFT baseline isolates the contribution of our three SFT innovations under the complete training budget. The full SFT recipe outperforms Naive SFT: 1.3 points lower WER on ASR, 2.0 points lower DER on diarization, +6.1 event F1 on dense captioning, and +9.1 F1 on audio grounding. A controlled study isolating each component individually is provided in Appendix C.

Effect of reinforcement learning. Adding GRPO on top of the Stage 1+2 SFT checkpoint produces consistent but moderate improvements across all four tasks (Table 2). Multi-task RL improves ASR WER by 1.2 points (44.7 to 43.5), diarization sF1 by 1.4 points (54.9 to 56.3), dense captioning eF1 by 0.8 points (58.5 to 59.3), and audio grounding mIoU by 0.9 points (48.5 to 49.4). Single-task RL, in which a separate RL stage is run for each task, yields gains of similar magnitude but does not consistently match multi-task RL across all metrics. However, RL did not help on music captioning, where nearly every metric declines slightly relative to the SFT checkpoint. The smaller magnitude of these RL gains relative to the SFT gains supports our overall finding: careful SFT design is the primary lever for installing timestamping in LALMs, with RL providing additional refinement of timestamp boundaries.

6 Conclusion

We presented TEMPO, the first unified LALM for timestamped generation across speech, sound, and music, covering multi-speaker ASR, speaker diarization, audio temporal grounding, dense audio captioning, and timestamped music captioning in a single decoder. Our SFT recipe combines atomic timestamp tokens, a time-aware multi-modal projector, and a distance-aware Gaussian loss, trained under a synthetic-to-real curriculum and refined with GRPO against verifiable temporal rewards. TEMPO substantially outperforms Audio Flamingo Next and Qwen3-Omni on a benchmark spanning six corpora and five tasks, with the largest gains on multi-speaker ASR (WER from 69.7% to 43.5%) and diarization (mIoU from 44.4% to 71.1%). Ablations show that careful SFT design is the primary lever for installing timestamping in LALMs, with RL providing additional refinement.

Limitations

TEMPO covers five timestamping tasks but does not address all forms of temporal audio understanding. In particular, we do not handle overlapping multi-domain events (e.g., simultaneous music, background sounds, and speech), which would require the model to jointly segment and label multiple concurrent streams.

Second, our temporal resolution is fixed at 0.1 s, which may be too coarse for tasks requiring sub-100 ms precision such as phoneme-level alignment or onset detection in percussive music, and unnecessarily fine for tasks with inherently ambiguous boundaries such as environmental sound grounding.

Third, TEMPO is built on Audio Flamingo 3 with a frozen Whisper-large encoder, inheriting any limitations of that encoder's spectral and temporal representations. Our results are thus conditioned on this specific encoder and may not generalize to other audio frontends.

Finally, the music captioning component is trained and evaluated exclusively on Slakh2100, a synthesized MIDI dataset. Performance on real-world recordings with natural acoustics, mixing artifacts, and production effects may differ substantially.

References

Hervé Bredin. 2017. pyannote.metrics: A Toolkit for Reproducible Evaluation, Diagnostic, and Error Analysis of Speaker Diarization Systems. In *Interspeech 2017*, pages 3587–3591.

John Ashley Burgoyne, Jonathan Wild, and Ichiro Fujinaga. 2011. An expert ground truth set for audio chord recognition and music analysis. In *ISMIR*.

Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Maël Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnès Lisowska, Iain McCowan, Wilfried Post, Dennis Reidsma, and Pierre Wellner. 2005. The AMI meeting corpus: A pre-announcement. In *Machine Learning for Multimodal Interaction (MLMI)*, volume 3869 of *Lecture Notes in Computer Science*, pages 28–39. Springer.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. 2024. Qwen2-audio technical report. *Preprint*, arXiv:2407.10759.

Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. 2019. Clotho: An audio captioning dataset. *Preprint*, arXiv:1910.09387.

Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780.

Sreyan Ghosh, Arushi Goel, Kaousheik Jayakumar, Lasha Koroshinadze, Nishit Anand, Zhifeng Kong, Siddharth Gururani, Sang gil Lee, Jaehyeon Kim, Aya Aljafari, Chao-Han Huck Yang, Sungwon Kim, Ramani Duraiswami, Dinesh Manocha, Mohammad Shoeybi, Bryan Catanzaro, Ming-Yu Liu, and Wei Ping. 2026. Audio flamingo next: Next-generation open audio-language models for speech, sound, and music. *Preprint*, arXiv:2604.10905.

Sreyan Ghosh, Arushi Goel, Lasha Koroshinadze, Sang gil Lee, Zhifeng Kong, Joao Felipe Santos, Ramani Duraiswami, Dinesh Manocha, Wei Ping, Mohammad Shoeybi, and Bryan Catanzaro. 2025a. Music flamingo: Scaling music understanding in audio language models. *Preprint*, arXiv:2511.10289.

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

Sreyan Ghosh, Sonal Kumar, Ashish Seth, Chandra Kiran Reddy Evuru, Utkarsh Tyagi, S Sakshi, Oriol Nieto, Ramani Duraiswami, and Dinesh Manocha. 2024. Gama: A large audio-language model with advanced audio understanding and complex reasoning abilities. *Preprint*, arXiv:2406.11768.

John J. Godfrey, Edward C. Holliman, and Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pages 517–520. IEEE.

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

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, and 175 others. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638.

Shawn Hershey, Daniel P W Ellis, Eduardo Fonseca, Aren Jansen, Caroline Liu, R Channing Moore, and Manoj Plakal. 2021. The benefit of temporally-strong labels in audio event classification. *Preprint*, arXiv:2105.07031.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. 2019. Audiocaps: Generating captions for audios in the wild. In *NAACL-HLT*.

KimiTeam, Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, Zhengtao Wang, Chu Wei, Yifei Xin, Xinran Xu, Jianwei Yu, Yutao Zhang, Xinyu Zhou, Y. Charles, and 21 others. 2025. Kimi-audio technical report. *Preprint*, arXiv:2504.18425.

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. *Preprint*, arXiv:2402.01831.

Sonal Kumar, Prem Seetharaman, Ke Chen, Oriol Nieto, Jiaqi Su, Zhepei Wang, Rithesh Kumar, Dinesh Manocha, Nicholas J. Bryan, Zeyu Jin, and Justin Salamon. 2026. Tac: Timestamped audio captioning. *Preprint*, arXiv:2602.15766.

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

Ethan Manilow, Gordon Wichern, Prem Seetharaman, and Jonathan Le Roux. 2019. Cutting music source separation some slakh: A dataset to study the impact of training data quality and quantity. *Preprint*, arXiv:1909.08494.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210.

Karol J. Piczak. 2015. ESC: Dataset for Environmental Sound Classification. In *Proceedings of the 23rd Annual ACM Conference on Multimedia*, pages 1015–1018. ACM Press.

Paul Primus, Florian Schmid, and Gerhard Widmer. 2025. Tacos: Temporally-aligned audio captions for language-audio pretraining. *Preprint*, arXiv:2505.07609.

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

Elizabeth Shriberg, Raj Dhillon, Sonali Bhagat, Jeremy Ang, and Hannah Carvey. 2004. The ICSI meeting recorder dialog act (MRDA) corpus. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL 2004*, pages 97–100, Cambridge, Massachusetts, USA. Association for Computational Linguistics.

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. *Preprint*, arXiv:2310.13289.

Hualei Wang, Yiming Li, Shuo Ma, Hong Liu, and Xiangdong Wang. 2025. Listening between the frames: Bridging temporal gaps in large audio-language models. *Preprint*, arXiv:2511.11039.

Quan Wang, Yiling Huang, Guanlong Zhao, Evan Clark, Wei Xia, and Hank Liao. 2024. Diarizationlm: Speaker diarization post-processing with large language models. In *Interspeech 2024*, interspeech_2024, page 3754–3758. ISCA.

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

Han Yin, Yafeng Chen, Chong Deng, Luyao Cheng, Hui Wang, Chao-Hong Tan, Qian Chen, Wen Wang, and Xiangang Li. 2026. Speakerlm: End-to-end versatile speaker diarization and recognition with multimodal large language models. *Preprint*, arXiv:2508.06372.

A Dataset and Preprocessing Details

Multi-Speaker ASR and Diarization. AMI, ICSI, and Switchboard feature spontaneous, overlapping, naturalistic speech. Recordings are segmented into chunks of up to 60 seconds while preserving multi-speaker structure within each chunk.

Audio Temporal Grounding. AudioSet Strong provides categorical event labels rather than natural-language queries. We use GPT-4o (OpenAI et al., 2024) to generate diverse paraphrased queries for each event class (e.g., for `dog_bark`: "When does a dog start barking?"), with the strong temporal labels serving as ground-truth intervals.

Dense Audio Captioning. TACOS contains 12,358 FreeSound recordings (15–30 seconds each) annotated with 47,748 timestamped captions.

Timestamped Music Captioning. Slakh2100 contains 145 hours of audio split into 30-second segments. Chord labels are extracted by quantizing simultaneously sounding notes into standard chord symbols (e.g., C major, G7); instrument entries/exits, tempo, dynamics, and note statistics are read directly from the MIDI metadata.

Synthetic speech compositions. We synthesize multi-speaker audio by randomly selecting 1-4 speakers from LibriSpeech and concatenating their utterances onto a shared 60-second canvas with small gaps between segments. Speaker slots are balanced across 1-4 speakers in equal proportions. Segment start and end times are chosen to avoid over-representing any particular timestamp value, and the final audio is peak-normalized.

Synthetic audio event compositions. We synthesize audio event recordings by randomly placing ESC-50 sound clips onto a 60-second canvas, with each composition containing 1-6 overlapping or non-overlapping events drawn from different sound categories. Separate pools are generated for DAC and AG to decouple their category distributions. For AG, we ensure at least 50% of compositions contain repeated occurrences of the same sound category, so that the model must learn to localize multiple spans of the same event. A random 10% of AG compositions ground the full clip rather than a specific event. As with speech, timestamp values are balanced uniformly across all compositions.

B Evaluation Metric Details

This appendix gives the full computational definitions of the evaluation metrics. All interval matching uses the greedy IoU-based matching described below unless stated otherwise.

Word Error Rate (WER). For multi-speaker ASR, we first greedily match predicted segments to ground-truth segments by temporal IoU. For each matched pair we compute the word-level Levenshtein distance between the normalized predicted and reference transcripts, and accumulate it into a total edit count. Unmatched ground-truth segments contribute all of their reference words as deletions, and unmatched predicted segments contribute all of their words as insertions. The final WER is the total edit count divided by the total reference word count across all segments.

Diarization Error Rate (DER). We use the standard DER as implemented in `pyannote.metrics` (Bredin, 2017), with a collar of 0.25 s. Speaker labels are optimally permuted between hypothesis and reference before scoring, so that a hypothesis with consistent but arbitrarily named speakers incurs no penalty from the labeling alone. DER aggregates missed speech, false alarms, and speaker confusion as a fraction of total reference speech duration.

Symmetric mIoU. For a set of reference intervals and a set of predicted intervals, we define the recall-side mean-best-IoU as the average over reference intervals of the maximum IoU with any predicted interval, and the precision-side mean-best-IoU symmetrically as the average over predicted intervals of the maximum IoU with any reference interval. The reported symmetric mIoU is the harmonic mean of these two quantities. The harmonic mean prevents over-prediction from inflating the score: emitting many spurious intervals raises recall-side IoU but lowers precision-side IoU, and the harmonic mean is dominated by the smaller of the two.

F1 at IoU threshold. For dense audio captioning and audio grounding, a predicted interval is a true positive if its IoU with a greedily-matched ground-truth interval is at least 0.5; precision, recall, and F1 are computed from the resulting true-positive, false-positive, and false-negative counts. For diarization, the speaker-labeled F1@IoU0.5 additionally requires that the predicted and ground-truth segments carry the same speaker label, so a temporally correct segment with the wrong speaker is not counted as a true positive.

Boundary MAE. For tasks with continuous time boundaries (ASR, grounding, dense captioning, music), we report the mean absolute error of segment onsets and offsets in seconds. The MAE is computed only over greedy-matched pairs with $\mathrm{IoU} \geq 0.3$, so that boundary error is measured only where the prediction and reference refer to the same underlying event.

METEOR. For dense audio captioning, caption quality is measured with METEOR computed over greedy-matched predicted and ground-truth segment pairs and averaged across matched pairs. Matching uses the same IoU-based greedy procedure as the other interval metrics.

Music-specific metrics. For chords, we report two frame-level accuracies sampled at 0.5 s resolution: *root accuracy* (fraction of sampled frames with the correct chord root) and *quality accuracy* (fraction with the correct chord quality label). We additionally report *chord F1@IoU0.5*, which greedily matches predicted and ground-truth chord segments requiring both $\mathrm{IoU} \geq 0.5$ and an exact chord-label match, and *chord mIoU*, the symmetric mIoU over all chord segments. For instruments, we report *instrument-name F1* (token-level F1 between the predicted and ground-truth instrument-name sets) and *instrument boundary MAE* (onset/offset error in seconds over matched instruments). *Tempo accuracy* is the fraction of examples whose predicted BPM falls within 4% of the ground truth.

C Per-Component SFT Ablation

To isolate the contribution of each SFT innovation, we train four variants on a controlled subset of 2,000 real-world examples: (i) *Naive*, with neither the distance-aware loss nor the time-aware projector; (ii) + *Loss*, adding only the distance-aware Gaussian loss; (iii) + *Projector*, adding only the time-aware multi-modal projector; and (iv) + *Both*, the full SFT recipe with both components. Due to compute constraints, we report three representative tasks spanning speech and audio (diarization, dense audio captioning, and audio temporal grounding) in Table 3.

An interesting pattern emerges. The distance-aware loss, in isolation, provides large improvements across all three tasks, including a 14.6-point gain in dense captioning event F1 and a corresponding gain in diarization speaker-labeled F1 and grounding F1. In contrast, the time-aware projector in isolation is not useful, suggesting that wall-clock positional information requires a corresponding learning signal to be effective. However, when paired with the loss, the projector contributes a clear additional benefit on the two tasks with the densest temporal structure: diarization (DER improves from 86.2% to 79.3%) and audio grounding (F1 from 38.0% to 38.7%, MAE from 1.73 s to 1.69 s, mIoU from 41.9% to 42.5%).

D Potential Risks

We do not foresee significant risks specific to this work beyond those already inherent to existing large audio-language models and speech recognition systems. TEMPO performs temporal localization of events in audio and does not generate speech, identify individual speakers by name, or enable capabilities not already present in publicly available audio understanding systems. All training and evaluation data are drawn from established, publicly released research corpora.

E License for Artifacts

We use a number of publicly released scientific artifacts in this work. Below, we summarize the licenses under which they are distributed and confirm that our use is consistent with their intended terms.

Datasets. AMI and ICSI are released under the CC BY 4.0 license for research use. Switchboard is distributed by the Linguistic Data Consortium (LDC) under a research license. AudioSet and AudioSet Strong are released under CC BY 4.0. TACOS is released under CC BY 4.0. Slakh2100 is released under CC BY 4.0. LibriSpeech is released under CC BY 4.0. ESC-50 is released under the CC BY-NC 3.0 license, which permits non-commercial research use.

Models. Audio Flamingo 3, our base model, is released under a research-only license by NVIDIA. Baselines we evaluate against (Audio Flamingo Next, Qwen3-Omni, TimeAudio) are accessed through their official releases under their respective research licenses.

Compatibility with intended use. All artifacts are used for non-commercial academic research, which is consistent with the intended use specified by each provider.

F Computational Resources

TEMPO is built on Audio Flamingo 3, which consists of a frozen Whisper-large audio encoder (approximately 0.6B parameters) and a Qwen2-7B language model (approximately 7B parameters), connected by a two-layer MLP projector. The audio encoder remains frozen throughout training. All experiments are run on NVIDIA A6000 GPUs. SFT runs on 8 GPUs on a single node, with Stage 1 covering 51,512 synthetic examples over two epochs and Stage 2 covering 32,726 real examples over two additional epochs. GRPO training (1,000 steps with 8 completions per prompt) runs on 8 GPUs distributed across 2 nodes (4 GPUs per node) using DDP via HuggingFace Accelerate. Evaluation

Table 3: SFT component ablation on a controlled 2,000-example subset, reported on three representative tasks. Loss: distance-aware Gaussian timestamp loss. Proj: time-aware multi-modal projector. DER, eF1, F1, mIoU, and MET (METEOR) are reported as percentages; MAE in seconds. ↓: lower is better; ↑: higher is better. Best result per column in **bold**.

Variant	**Diarization**		**Dense Captioning**			**Audio Grounding**		
	DER↓	mIoU↑	eF1↑	mIoU↑	MET↑	F1↑	MAE↓	mIoU↑
Naive	98.4	23.8	35.3	43.3	20.5	33.9	2.05	36.1
+ Loss	86.2	**37.5**	**49.9**	**60.2**	21.5	38.0	1.73	41.9
+ Projector	104.2	22.2	34.2	42.8	20.5	34.0	2.02	36.4
+ Both (Ours)	**79.3**	**37.5**	49.7	60.1	**21.9**	**38.7**	**1.69**	**42.5**

across all five tasks (10,513 question-answer pairs) is run on a single A6000.

G Use of AI Assistants

We utilized AI assistants to help clarify explanations, suggest concise phrasing, and organize text for readability. These tools were used exclusively for linguistic support and were not used to generate scientific results or formulate claims.