Listening Between the Frames: Bridging Temporal Gaps in Large Audio-Language Models

Hualei Wang[1,2]* Yiming Li[1,2]* Shuo Ma[1,2] Hong Liu[1] Xiangdong Wang[1]†

[1]Beijing Key Laboratory of Mobile Computing and Pervasive Device,
Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
[2]University of Chinese Academy of Sciences, Beijing, China
{wanghualei23s,liyiming22s,mashuo20g,hliu,xdwang}@ict.ac.cn

Abstract

Recent Large Audio-Language Models (LALMs) exhibit impressive capabilities in understanding audio content for conversational QA tasks. However, these models struggle to accurately understand timestamps for temporal localization (e.g., Temporal Audio Grounding) and are restricted to short audio perception, leading to constrained capabilities on fine-grained tasks. We identify three key aspects that limit their temporal localization and long audio understanding: (i) timestamp representation, (ii) architecture, and (iii) data. To address this, we introduce TimeAudio, a novel method that empowers LALMs to connect their understanding of audio content with precise temporal perception. Specifically, we incorporate unique temporal markers to improve time-sensitive reasoning and apply an absolute time-aware encoding that explicitly grounds the acoustic features with absolute time information. Moreover, to achieve end-to-end long audio understanding, we introduce a segment-level token merging module to substantially reduce audio token redundancy and enhance the efficiency of information extraction. Due to the lack of suitable datasets and evaluation metrics, we consolidate existing audio datasets into a new dataset focused on temporal tasks and establish a series of metrics to evaluate the fine-grained performance. Evaluations show strong performance across a variety of fine-grained tasks, such as dense captioning, temporal grounding, and timeline speech summarization, demonstrating TimeAudio's robust temporal localization and reasoning capabilities.

Code — https://github.com/lysanderism/TimeAudio

Introduction

Audio, mainly including speech and non-speech sounds, is fundamental to human life, helping us perceive our surroundings, gather crucial information, and interact with others. To automatically interpret acoustic content and map it to human cognition, models are specially designed with advanced neural architectures and learning schemes to align audio features with natural language. Due to the intrinsic linguistic nature, speech tasks, such as automatic speech recognition and spoken question answering, have been easily integrated with language models (Chuang et al. 2019; Zhang et al. 2023). For environmental sound, methods like Contrastive Language Audio Pre-training (CLAP) (Wu et al. 2023) has been proposed to embed audio and its corresponding caption into a shared latent space. However, these models are restricted to fixed task formulations or exclusive audio types (either speech or sound), limiting their ability to achieve human-like audio understanding.

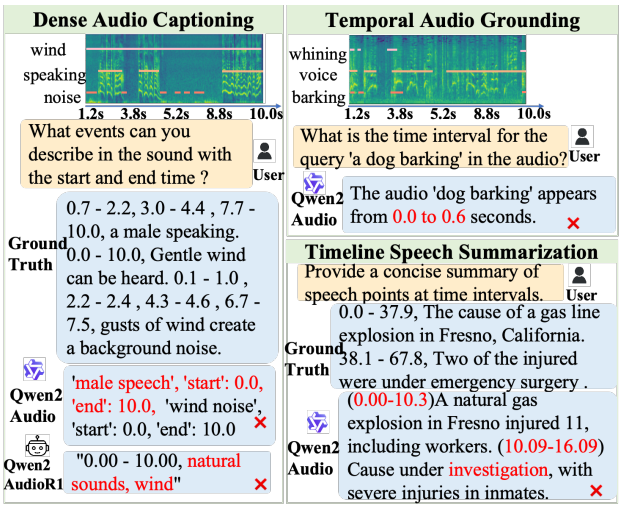

Figure 1: Example of failed cases by Qwen2-Audio and Qwen2-Audio-R1 on fine-grained tasks that require both semantics and timestamps as output.

Leveraging rich knowledge in large language models (LLMs) (Achiam et al. 2023), large audio language models (LALMs), which integrate audio encoders into pre-trained decoder-based LLMs, enable free-form audio question answering (AQA) (Lipping et al. 2022) and unified audio understanding. For example, Qwen2-Audio (Chu et al. 2024) conducts large-scale pre-training to align audio and text modalities, followed by tuning on instruction data to further enhance the command-following capability. It not only demonstrates competitive performance on specific tasks compared to expert models (e.g., whisper-large-v3 (Radford et al. 2023a)) but also offers more flexible and natural interactions.

*Equal contribution.
†Corresponding author.

Despite the notable achievements witnessed in LALMs, they still fall short in fine-grained audio understanding (Xu et al. 2024; Mesaros et al. 2021), especially when precise timestamp prediction is required. To demonstrate this, we evaluate Qwen2-Audio's capability to summarize sound events or speech content along with the corresponding onset/offset. As shown in Figure 1, it struggles to accurately link temporal locations with acoustic semantics or speech meanings. Moreover, it exhibits significant hallucination when processing long audio, as it lacks the capability for end-to-end comprehension of long-form audio. Similar poor performance from other LALMs can also be observed in the following experiments reported later (see Table 2). The underlying reasons may be two folds: (1) existing LALMs directly project audio features into the shared latent space without explicitly modeling detailed grounding information, making it difficult for the LLM decoder to predict precise timestamps; (2) the instruction tuning data and evaluation benchmarks (Sakshi et al. 2024) focus on general audio understanding rather than highlighting fine-grained temporal reasoning.

To address the above shortcomings, we propose TimeAudio, a comprehensive framework that incorporates fine-grained acoustic cues into LALMs with enhanced module designs and a specially curated dataset.

- At the module level, temporal markers are integrated into LALM's vocabulary to reduce the convergence burden of numerical regression, and absolute time-aware encoding is adopted to explicitly inject timestep information into audio embeddings. Additionally, to efficiently handle long audio inputs, we devise a novel token selection and merging strategy, that balances token length with information density.
- At the data level, we first design several novel timestamps-related understanding tasks, such as dense audio captioning and timeline speech summarization, that require both high-level semantic understanding and low-level temporal grounding. Based on these tasks, a new large-scale instruction dataset named FTAR is constructed. We also introduce well-established metrics to evaluate LALMs' fine-grained understanding performance.

By incorporating the above techniques and being further fine-tuned on the gathered FTAR dataset, our proposed TimeAudio demonstrates significant improvements in temporal localization and fine-grained understanding capability compared to prior LALMs.

Related Work

Fine-grained Temporal Audio Understanding

General audio understanding pays attention to overall content in the clip, for example, audio captioning (Wu, Dinkel, and Yu 2019) solely requires the model to output the contained sound events with simple temporal order descriptions. In contrast, fine-grained temporal understanding aims to grasp semantics along with their corresponding time intervals. For instance, audio grounding models (Xu et al. 2024) produce specific sound events with their timestamps, and some meeting summary tasks summarize key points with respective time spans(Hu et al. 2023). Fine-grained temporal understanding offers more traceable evidence and details for users, which is important to reduce hallucination and build responsible models. However, current LALMs perform unpromisingly regarding fine-grained temporal reasoning, while expert models (Wu et al. 2025; Li et al. 2024; Wang et al. 2023) lack zero-shot capabilities, which limits their broader applications. Moreover, the absence of suitable instruction tuning datasets for temporal tasks, coupled with extreme task imbalance in existing datasets, hinders model performance. To address this issue, we construct a new, diverse dataset and propose TimeAudio, a model featuring time-sensitive modules that enhance LALM performance on fine-grained tasks.

Large Audio Language Models

Vanilla LLMs have proven effective at leveraging their captured knowledge for zero-shot solutions to general challenges. Recent research has explored to extend their capabilities further by integrating information from other modalities like video (Liu et al. 2023) and audio (Gong et al. 2023). In the audio modality, LALMs project acoustic features into the embedding space of LLMs and leverage supervised finetuning to enhance the instruction following capability for audio inputs. The pioneer Pengi (Deshmukh et al. 2023) utilizes fixed templates as prompts to align audio and text modality, showing promising outcomes compared to the original CLAP. LTU (Gong et al. 2023) and GAMA (Ghosh et al. 2024) enable the model to answer free-form sound-related questions by incorporating diverse AQA pairs and robust feature encoders. SALMONN (Tang et al. 2023) and Qwen2-Audio (Chu et al. 2024) further extend the paradigm to more audio types and instruction forms, significantly improving LALM's scalability. However, the previously mentioned LALMs frequently cause hallucination issues if provided with longer input, as they struggle to perform end-to-end processing on full-length audio. To handle longer inputs, Audio Flamingo2 (Ghosh et al. 2025) proposes a sliding window mechanism with CLAP to achieve long environmental sound comprehension. Qwen2.5-Omni (Xu et al. 2025) adopts a streaming transcription strategy to recognize full-length speech content. These methods are not directly transferable, as they depend on extensive training or specialized adaptations of the model architecture.

Method

In this section, we present TimeAudio, an LALM that utilizes temporal markers and incorporating two key modules: absolute time-aware encoding and segment-level merging. These modules aim to enhance TimeAudio's temporal awareness and expand its capacity to understand and localize details in audio tasks. To further bridge the gap in fine-grained temporal reasoning and robust instruction following, we introduce the FTAR (fine-grained temporal audio reasoning) dataset – a comprehensive dataset built for instruction tuning on time-sensitive tasks. We then fine-tune our model on this dataset to fully unleash its capabilities.

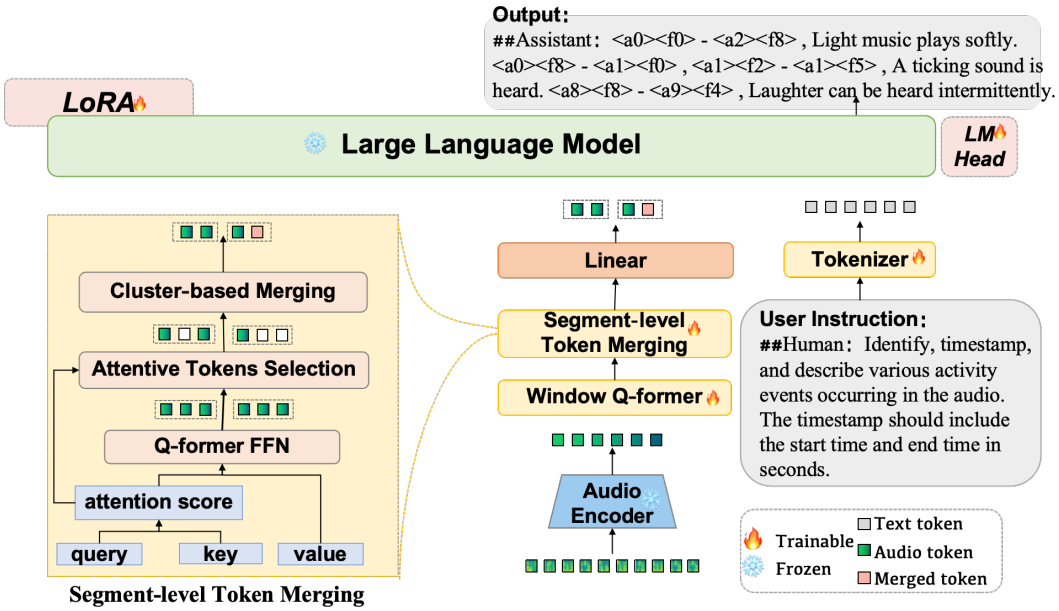

Figure 2: Overview of our TimeAudio method. The input audio is first split into segments and encoded into audio tokens. The window Q-former then projects these audio tokens into the language space and utilize a segment-level token merging to retain important semantic information along time. Timestamps are converted to special anchor and offset tokens.

Overview of TimeAudio

TimeAudio is based on the fundamental architecture of SALMONN (Tang et al. 2023). Its overview is provided in Figure 2. Specifically, TimeAudio is consists of four components: a sliding audio encoder, a window Q-former, a segment-level token merging module, and an LLM to process raw audio. The sliding audio encoder first divides long audio into shorter segments and combines the BEATs (Chen et al. 2022) and the Whisper encoder (Radford et al. 2023b) to extract features for each segments independently. Then, the window Q-former projects these encoded audio tokens into the language space and applies a segment-level token merging mechanism based on attention scores to filter out unimportant acoustic information. Finally, the audio embeddings and the textual token embeddings of user prompts are fed into the LLM to generate response.

Temporal Markers

To improve LALM's comprehension and reduce its hallucination, it is promising to capture fine-grained temporal relationships instead of merely detecting the content of the audio. Previous studies (Sridhar, Guo, and Visser 2024; Huang et al. 2024) show that LLMs can inherently encode temporal information from sequential inputs, however, directly predicting precise timestamps over long audio remains challenging. In temporal audio understanding tasks, fine-grained audio comprehension requires temporal semantics across large time spans and during overlapping events. However, predicting timestamps with the simple number token hinder the language model to capture precise semantics. Furthermore, the relative time tokens (Wang et al. 2024; Bain et al. 2023) with fixed intervals (<0.2> <0.4> <0.6>...) imposes a heavy burden on the LLM's vocabulary and brings quantization errors in audio processing.

To resolve these issues, we introduce ***Temporal Markers***, which incorporate unique temporal tokens into the tokenizer to assist the LLM perceive specific timestamps. Given a fine-grained caption depicting a particular audio clip and its associated timestamps, we have designed anchor and offset tokens to convert continuous timestamps into a sequence of discrete temporal tokens. The anchor token grounds the prediction immediately while the offset tokens represent a fine-grained adjustment. This strategy reduces the total number of temporal tokens, maintaining constant precision regardless of audio length. We then convert the timestamp-related text into the unified temporal marker format. Textual and temporal tokens are mapped into a shared semantic space through the extended word embedding layer. An example input (containing male voice from 0.0s to 2.5s, 3.2s to 8.0s and soundtrack from 0.0s to 9.0s) is shown below:

<s><audio>$\mathbf{F}_{audio}$</audio> <a0><f0> - <a2><f5>, <a3><f2> - <a8><f0>, A male voice delivers a great performance. <a0><f0> - <a9><f0>, the soundtrack is filled with rich music.</s>

where <s> and </s> indicate the start and end of sequence, <audio> and </audio> indicate the start and end of encoded audio features. <a> and <f> indicate the anchor and offset token with different time. Despite solving quantization errors with temporal marker tokens, randomly initializing these tokens degrade the pretrained embedding space.To

tackle this issue, we transfer knowledge that implicitly contained in raw numeral to these anchor tokens, as temporal understanding is already implicitly contained within the LLM. Furthermore, we compute the embedding for each offset token by averaging the embeddings of its numeral tokens and the decimal-point token. For example, consider the time token $\langle a_0 \rangle$, $\langle f_0 \rangle$ and its embedding:

$$[\mathbf{W}_{\text{token}}]_{\text{ID}(\langle a_0 \rangle)} = [\mathbf{W}_{\text{token}}]_{\text{ID}(0)}$$
$$[\mathbf{W}_{\text{token}}]_{\text{ID}(\langle f_0 \rangle)} = ([\mathbf{W}_{\text{token}}]_{\text{ID}(0)} + [\mathbf{W}_{\text{token}}]_{\text{ID}(.)})/2 \quad (1)$$

where ID() denotes the token ID of the input token. To ensure knowledge is properly transferred from numeral tokens to temporal tokens at the final prediction stage, we apply the same initialization to the LLM's prediction head.

Absolute Time-aware Encoding

While recent multi-modal models (Xu et al. 2025; Guo et al. 2025) have demonstrated the efficiency of absolute temporal positions, LALMs still suffer degraded performance when understanding the exact temporal order of events. The considerable diversity in event and speech prosody makes it difficult for the model to accurately identify the true temporal locations and perform effective searching. To enhance temporal awareness of the audio feature, we introduce a time-aware encoding that explicitly grounds acoustic features to absolute timeline.

Given a long audio X, we divide it into N_s length segments, represented as $\mathbf{X}_a = \{x_i\}_{i=0}^{N_s-1}$. And each segment x_i is encoded to a contiguous audio embedding as:

$$\mathbf{W}_i = \text{Concat}(\mathcal{G}_{\text{Whisper}}(x_i), \mathcal{G}_{\text{BEATs}}(x_i)) \quad (2)$$

where the dual audio encoders $\mathcal{G}(\cdot)$ extract speech and audio feature independently, and these are concatenated frame-by-frame along the feature dimension, yielding the acoustic embeddings $\mathbf{W}_i$. Then, we construct learnable the absolute time embedding $\mathbf{W}_t$ to explicitly provide information for temporal grounding. This preserves the sequence embedding with the relative order while accurately reflecting the specific position of each time point in the audio sequence. The audio embedding is augmented with its corresponding absolute time embedding:

$$\mathbf{e}_t(t_i) = \mathbf{h}(j_i)^\top \mathbf{W}_t \; = \; [\mathbf{W}_t]_{j_i} \in \mathbb{R}^d, \quad (3)$$
$$\hat{\mathbf{W}}_i = \mathbf{W}_i + \mathbf{e}_t(t_i). \quad (4)$$

where t represents the absolute timestamp (in seconds) associated with the corresponding segments. The time embedding is selected by one-hot lookup $\mathbf{h}(j_i)^\top \mathbf{W}_t = [\mathbf{W}_t]j_i$, and j_i denotes the discretized time index. The time embedding $\mathbf{W}_t$ is zero-initialized to preserve the integrity of the pretrained audio encoders during the initial phases of training. We posit that absolute time-aware and positional encodings are orthogonal, a claim validated by our experimental evidence.

Segment-level Token Merging

After obtaining the audio tokens with temporal feature, we apply the Q-former to project T-frames audio into L semantic tokens. Although Q-Former uses a fixed number of queries for alignment, handling long audio is still costly in computation. Some existing methods typically compress speech feature through higher-level projections (Kang and Roy 2024) or temporal down-sampling (Shang et al. 2024). Inspired by VisionZip (Yang et al. 2025), we incorporate a segment-level token merging to prune the redundant tokens in the end-to-end structure. Our method reuses Q-former's attention information, avoiding extra computation and memory overhead. We discuss the details of the segment token merging as follows.

Attentive Tokens Selection. In speech summarization, the particular concern is that long speech usually contains extensive transcripts with dispersed information. Therefore, our goal is to adaptively select important audio tokens from each audio segment and merge the redundant tokens into contextual feature. In order to evaluate the relative importance of each audio token, we investigate the attention scores computed within the Q-Former. Specifically, we compute the attention matrix:

$$\mathbf{A} = \text{Softmax}\big(\mathbf{Q}\mathbf{K}^\mathsf{T}/\sqrt{D}\big) \; \in \; \mathbb{R}^{B \times N_q \times N_q} \quad (5)$$

where $\mathbf{A}$ represents the multi-head attention scores, D is the state dimension, and $\mathbf{Q}$ and $\mathbf{K}$ represent query and key from $\mathbf{W}_i$, respectively. We identify the most salient tokens by averaging the attention scores across all heads to produce a unified score matrix. Tokens that receive higher average attention from all other tokens in the sequence are more significant and are therefore preserved.

Cluster-based Tokens Merging. After selecting the attentive audio tokens, we introduce a cluster-based merging process to retain information from the remaining tokens. This approach is based on the insight that the attention key vectors already capture the salient content of each token. To guide the merging, we first uniformly split the remaining tokens into target tokens and merge candidates. Then, we compute a similarity metric between the key vectors of the target and candidate tokens:

$$\mathbf{Sim}(\mathbf{h}_i, \mathbf{h}_j) = \mathbf{k}_i \, \mathbf{k}_j^\mathsf{T}, \quad (6)$$

where $\mathbf{Sim}(\mathbf{h}_i, \mathbf{h}_j)$ is the similarity obtained by the dot product of the key vectors; $\mathbf{h}_{i,j \in \{1,\ldots,n\}}$ denotes the remaining audio tokens. The candidate token assigned to the centroid to which it exhibits the highest semantic similarity. Finally, the most semantically-related tokens are fused together, yielding a set of contextual tokens.

Training Strategy

To develop an efficient LALM, we utilize a two-stage training pipeline as follows:

Stage-1: Temporal Token Alignment. In the first stage, we continue pre-training the model using checkpoints from well-trained LALMs to align fine-grained audio features with temporal information. We collect a wide range of grounding datasets, focusing on tasks such as Temporal Audio Grounding, Dense Audio Captioning, and Timeline Speech Summarization, that enable the model to search for and localize temporal information effectively. To facilitate

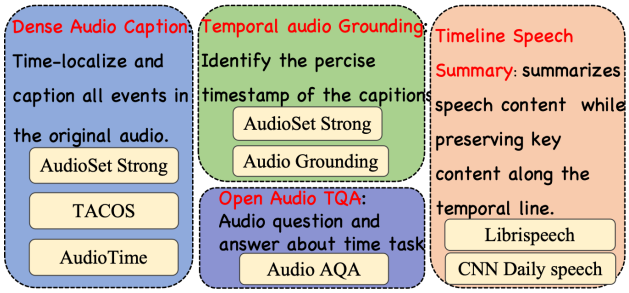

Figure 3: Involved tasks and datasets in the time-aware instruction tuning dataset.

this, we make the LoRA adapters, window Q-former, and the absolute time and special text embeddings trainable, while keeping the audio encoder and LLM frozen.

Stage-2: Long Audio Instruction Tuning. While the initial pre-training stage equips the model with a foundational capability for temporal reasoning, a significant mismatch occurs when it confronts long audio sequences, often resulting in semantic misalignment. To bridge this gap, the second stage focuses on enhancing its ability to understand long speech and respond to diverse instructions. This is achieved by fine-tuning the window Q-former and LoRA on a small set of instruction data, while keeping all other components frozen. This stage improves long audio comprehension while aligning both the temporal knowledge and the LLM's semantic space through diverse instructions.

FTAR Dataset Construction

In this section, we introduce the FTAR, a dataset comprising 260K publicly available audio-text pairs. It is composed of three audio tasks centered on temporal reasoning, drawn from diverse domains as detailed in Figure 3. The FTAR Dataset is specifically designed to help users obtain fine-grained content during interactions with AI assistants. Additional details are available in Appendix A.

- **Dense Audio Captioning:** The task of dense audio captioning requires generating event descriptions with their respective start and end times, formatted as: `<start>-<end>, captions`. We employ the Qwen2.5 model (Team 2024) to refines coarse-grained event labels into rich, fine-grained descriptions. We utilize real-world AudioSet-Strong (Hershey et al. 2021), TACOS (Primus, Schmid, and Widmer 2025), AudioTime (Xie et al. 2025) dataset to construct the task.
- **Temporal Audio Grounding:** The temporal audio grounding task is defined as localizing a specific event within an audio based on a descriptive sentence. For this task, we treat the event caption as a query and output the corresponding start and end times in the format `<start>-<end>`. We collect AudioSet-Strong, AudioGrounding (Xu et al. 2024) and formulate them.
- **Timeline Speech Summarization:** For the speech summarization task, it aims to condense spoken content while preserving key information along the temporal line, emphasizing both high-level semantic extraction and low-level grounding ability. We utilize F5-TTS (Chen et al. 2024) to synthesize speech from CNN/DailyMail summarization (Nallapati et al. 2016). Beyond synthetic data, We further collect Librispeech (Panayotov et al. 2015) segments for the task.

Tasks	# Sub-data	# Samples	Avg. len
Dense Audio Captioning	3	110K	11.3s
Temporal Audio Grounding	2	100K	9.8s
Timeline Speech Summarization	2	42K	81.7s
Audio TQA	1	15K	10.0s

Table 1: Datasets Used in Training for Various Tasks.

To enhance a rich diversity of temporal instructions, the Audio Temporal Question Answer (TQA) task is constructed from the OpenAQA dataset (Gong et al. 2023), which consists of free-form and diverse audio question-answer pairs on counting, duration, and time sequence. We also include audio captioning and speech recognition data to retain model's general audio understanding capabilities.

Table 1 presents the instruction tuning data across different tasks, highlighting the broad coverage in data scale, task coverage, and audio durations.

Experiments

Evaluation Setups

Implementation Details. We use 7B SALMONN as the base LALM for experiments (Tang et al. 2023). The window Q-former module and the weights of sequence embedding are initialized using the pre-trained checkpoint. We augment its vocabulary with M=20 specialized temporal tokens. The learnable time embedding is configured with a maximum of 768 positions. During data pre-processing, each audio is split into a sequence of 30-second intervals and the max number of segments is 5. Furthermore, to improve computational efficiency, we retain only 22 attentive tokens and 4 contextual tokens, effectively pruning 75% of redundant information. Our two-stage training process involves 10 epochs of continual pre-training using a learning rate of 1e-5 , followed by 5 epochs of instruction tuning with a learning rate of 2e-6. More details can be found in Appendix B.

Evaluation Tasks and Metrics. For dense audio captioning, we test on the AudioSet-Strong evaluation set (Hershey et al. 2021) with caption less than 100 words. Metrics such as METEOR score (Banerjee and Lavie 2005), event-based measures (Eb) and clip-level macro F1 score (At) (Mesaros, Heittola, and Virtanen 2016) are applied to evaluate the time perceptive ability and the diversity of descriptions between the generated events and the ground-truth. For temporal audio grounding, we evaluate on the AudioGrounding test data and report the mean Intersection over Union (mIoU) and Recall at IoU thresholds of 0.5, 0.7, 0.9 between predicted and ground-truth timestamps. To evaluate timeline speech summarization, we follow the protocol of (Kang and Roy 2024)

Model	Scale	Dense Audio Captioning			Temporal Audio Grounding				Timeline Speech Summarization		
		METEOR	Eb-F1	At-F1	R@0.5	R@0.7	R@0.9	mIoU	ROUGE-1	ROUGE-L	mIoU
Zero-shot LALMs											
Qwen2-Audio (Chu et al. 2024)	7B	6.7	9.8	50.3	32.1	18.7	10.8	20.5	17.4	12.3	13.3
Qwen2.5 omni (Xu et al. 2025)	7B	3.4	9.6	36.0	17.9	10.9	6.5	11.8	15.6	11.7	12.8
Qwen2-Audio-R1 (Li et al. 2025)	7B	3.7	8.5	40.6	29.3	17.3	8.8	18.5	15.2	11.5	16.2
Hubert-MiniChat (Kang and Roy 2024)	3B	-	-	-	-	-	-	-	33.6	22.3	-
FTAR-Tuned LALMs											
GAMA (Ghosh et al. 2024)	7B	19.7	11.0	67.8	31.0	21.1	14.8	22.3	-	-	-
Qwen audio (Chu et al. 2023)	7B	10.6	13.5	43.6	53.4	32.2	23.7	36.4	15.9	20.5	76.8
Qwen2-Audio (Chu et al. 2024)	7B	**22.4**	36.5	67.8	72.8	55.4	26.8	51.7	40.0	28.5	85.2
SALMONN-7B (Tang et al. 2023)	7B	19.9	32.4	68.0	71.4	55.8	28.6	51.9	39.5	28.7	84.3
SALMONN-13B (Tang et al. 2023)	13B	20.2	32.0	67.6	69.2	53.5	28.3	50.3	40.2	29.8	88.2
TimeAudio (ours)	7B	20.4	**37.4**	**70.5**	**75.7**	**61.2**	**36.5**	**57.8**	**42.4**	**30.8**	**94.2**

Table 2: Comparison of performance on the fine-grained temporal task with other LALMs' methods. The **bold** item denotes the best result.

and use a subset of the CNN/DailyMail test set, including only articles under 1600 characters. We report standard content quality metrics (ROUGE (Lin 2004)) and supplement them with a mean Intersection over Union (mIoU) score to measure how accurately the summary is grounded to the audio timeline. More details are available in Appendix C.

Compared Methods. For our baseline evaluation, we first represent general-purpose LALMs used in a zero-shot inference performance. We select Qwen2-Audio (Chu et al. 2024), Qwen2-Audio-R1 (Li et al. 2025), and Qwen2.5 Omni (Xu et al. 2025). Second, for models specifically adapted for fine-grained tasks, we evaluate several prominent audio LLMs, including Qwen Audio (Chu et al. 2023), Qwen2-Audio, GAMA (Ghosh et al. 2024), and SALMONN (Tang et al. 2023), all of which undergo parameter-efficient fine-tuning.

Main Results

Dense Audio Captioning. This task involves accurately capturing the temporal locations of all sound events within an audio clip, alongside providing faithful descriptions that match the underlying time. As shown in Table 2, existing LALMs exhibit significant limitations in precise temporal localization under zero-shot conditions, a fact underscored by the top-performing Qwen2-Audio, which achieves an Eb-F1 score of 9.8. The inaccurate event localization directly impacts the captioning evaluation, such as the METEOR metrics. Our method exhibits a clear strength in temporal accuracy, achieving a remarkable performance gain over fine-tuned Qwen2-Audio with +0.9 Eb-F1 and +2.7 At-F1 scores. This demonstrates that TimeAudio effectively process audio with precise event localization. As for the METEOR score, the fine-tuned Qwen2-Audio model achieves even stronger performance. We speculate that this is due to the pre-trained checkpoint used for initialization, which not being trained on sufficiently diverse audio captioning.

Temporal Audio Grounding. This task directly reflects the ability to precisely localize time interval with a given query event. Results show that TimeAudio achieves 57.8 score on mIoU of the AudioGrounding dataset, which surpasses the fine-tuned LALMs, i.e. SALMONN-7B, by a substantial 11.4% gain in mIoU. This demonstrates a superior overall accuracy in fine-grained temporal perception for given textual descriptions. It is particularly noteworthy that TimeAudio achieves its greatest performance gains on the temporal audio grounding task. We argue that the substantial improvement, centered on a task that explicitly measures temporal localization, demonstrates the effectiveness of our proposed methods.

Timeline Speech Summarization. While the dense audio captioning and temporal audio grounding focus on audio tasks, this task aims to enhance fine-grained speech understanding at the segment level. Overall, our model achieves a 42.4 ROUGE-1 and 30.8 ROUGE-L on CNN/-DailyMail speech, outperforming other LALMs with speech summarization capabilities. This highlights the contribution of our segment-level token merging in retain the important semantics of each audio segment. Besides, TimeAudio surpasses even specialized, task-specific Hubert-MiniChat, which demonstrates both the challenging nature of the task and the superiority of our model in processing long audio.

Ablation Study

In this section, we systematically conduct ablation studies on TimeAudio. Specifically, we evaluate the effectiveness of our proposed module (excluding data effects) in Table 3 and analyze the impact of the token merging in Table 4.

Multi-Task Ablation. The TM-only setup yields substantial improvements on time-sensitive tasks, with gains of +3.6 Eb-F1 on AudioSet-Strong, +3.0 mIoU on AudioGrounding and +4.4 mIoU on CNN/DailyMail. These findings suggest that using time tokens significantly enhanced temporal understanding to accurately locate the timestamps. In the case of adding TM and ATE, the model's ability to temporally describe audio is enhanced, as indicated by an increase of 0.9 in METEOR and 5.4 in the Eb-F1. Adding

Method	Dense Audio Captioning			Temporal Audio Grounding				Timeline Speech Summarization		
	METEOR	Eb-F1	At-F1	R@0.5	R@0.7	R@0.9	mIoU	ROUGE-1	ROUGE-L	mIoU
SALMONN (FTAR-Tuned)	19.9	32.4	68.0	71.4	55.8	28.6	51.9	39.5	28.7	84.3
+w/ TM (not initialize)	19.5	33.5	69.2	71.7	56.2	33.5	52.3	40.0	28.8	87.5
+w/ TM	20.5	36.0	70.8	72.7	57.8	34.2	54.9	41.0	29.8	88.7
+w/ ATE	20.3	35.8	69.5	72.0	56.9	32.5	53.8	40.6	30.1	86.6
+w/ TM + ATE	**20.8**	**37.8**	**71.4**	73.7	58.6	35.5	56.0	41.4	29.8	90.4
+w/ TM + ATE + SEM	20.4	37.4	70.5	**75.7**	**61.2**	**36.5**	**57.8**	**42.4**	**30.8**	**94.2**

Table 3: Ablation study on different module components. We analyze and compare the effects of using time-sensitive modules for TimeAudio. TM indicates Temporal Marker, ATE indicates Absolute Time-aware Encoding and SEM indicates Segment-level Token Merging. The **bold** item denotes the best result.

Retained Ratio	Timeline Speech Summarization		
	ROUGE-1	ROUGE-L	mIoU
0.10	24.3	18.4	72.9
0.15	31.6	20.5	73.2
0.20	40.2	29.7	84.8
0.25	42.4	30.8	**94.2**
0.30	42.5	31.1	94.0

Table 4: Ablation for dominant audio tokens retained ratio.

segment-level token merging introduces expected trade-offs: while it improves timeline speech summarization accuracy on ROUGE-1 and mIOU, it leads to a moderate decline in dense captioning precision. The STM method saves more semantic information in long-form audio and improve alignment between the summarization and the audio content. These results highlight the effectiveness of our novel modules in the TimeAudio architecture.

Performance with More Attentive Tokens. As illustrated in Table 4, we present the results of token merging method under different ratio of retained tokens. A smaller ratio of retained tokens fails to capture sufficient semantic information across long audio, resulting in suboptimal performance in both aspects. Conversely, raising the attentive-token count offers marginal improvements in semantic understanding before the gains level off. We finally set the ratio of attentive tokens to 0.25 as a balanced trade-off between performance and computational efficiency.

Qualitative Performance

In Figure 4, we compare our method with ground-truth and Qwen2-Audio to highlight TimeAudio's superior temporal awareness and understanding. The first example presents a multi-event changes between short audio. While Qwen2-Audio model exhibits a tendency to predict delayed start times and struggle to identify the complete time interval, our method demonstrates superior precision by accurately capturing the correct relative boundaries of the event. The second case features a long audio characterized by events that happen in a specific order on CNN/DailyMail speech. A key limitation of the Qwen2-Audio model is its short context input, which necessitates the ability of LLM to guess contents

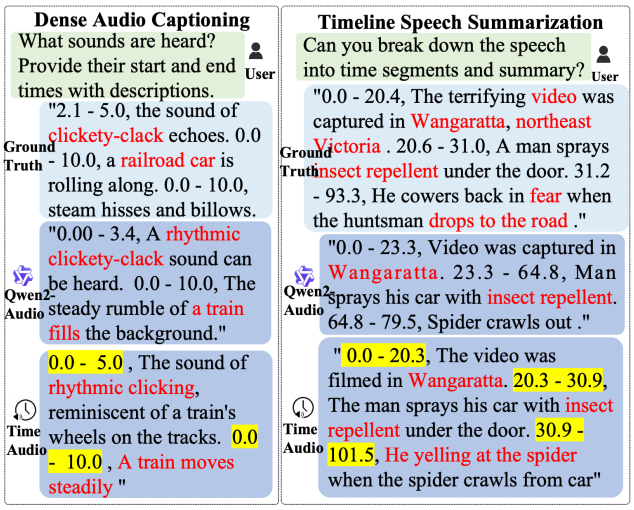

Figure 4: Qualitative results between models on the dense audio captioning task and timeline speech summarization task. Yellow denotes the time interval and red marks the audio content.

in longer audio segments. These qualitative examples underscore the robustness and precision of our method in scenarios that are especially challenging for other base methods. We provide additional cases on these tasks in Appendix D.

Conclusion

In this work, we propose TimeAudio, a time-sensitive LALM capable of fine-grained perception and understanding. This is enabled by a novel model design that incorporates temporal markers and absolute time-aware encoding for effective temporal modeling. To ensure effective compression in long audio, we introduce a segment-level token merging method to progressively preserves dominant information. Additionally, we construct the FTAR dataset to further strengthen the model's temporal reasoning. Extensive experiments demonstrate that TimeAudio significantly improves performance in time-centric scenarios, outperforming baselines on downstream temporal audio tasks while retaining general audio comprehension.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (62276250), the National Key R&D Program of China (2022YFF1203303) and key R&D program of Ningxia Autonomous Region (2024FRD05068). And also supported by the Major Project of the National Social Science Foundation of China (21&ZD292) and sponsored by Doubao Fund.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Bain, M.; Huh, J.; Han, T.; and Zisserman, A. 2023. Whisperx: Time-accurate speech transcription of long-form audio. *arXiv preprint arXiv:2303.00747*.

Banerjee, S.; and Lavie, A. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 65–72.

Chen, S.; Wu, Y.; Wang, C.; Liu, S.; Tompkins, D.; Chen, Z.; and Wei, F. 2022. Beats: Audio pre-training with acoustic tokenizers. *arXiv preprint arXiv:2212.09058*.

Chen, Y.; Niu, Z.; Ma, Z.; Deng, K.; Wang, C.; Zhao, J.; Yu, K.; and Chen, X. 2024. F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching. *arXiv preprint arXiv:2410.06885*.

Chu, Y.; Xu, J.; Yang, Q.; Wei, H.; Wei, X.; Guo, Z.; Leng, Y.; Lv, Y.; He, J.; Lin, J.; et al. 2024. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*.

Chu, Y.; Xu, J.; Zhou, X.; Yang, Q.; Zhang, S.; Yan, Z.; Zhou, C.; and Zhou, J. 2023. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*.

Chuang, Y.-S.; Liu, C.-L.; Lee, H.-Y.; and Lee, L.-s. 2019. Speechbert: An audio-and-text jointly learned language model for end-to-end spoken question answering. *arXiv preprint arXiv:1910.11559*.

Deshmukh, S.; Elizalde, B.; Singh, R.; and Wang, H. 2023. Pengi: An audio language model for audio tasks. *Advances in Neural Information Processing Systems*, 36: 18090–18108.

Ghosh, S.; Kong, Z.; Kumar, S.; Sakshi, S.; Kim, J.; Ping, W.; Valle, R.; Manocha, D.; and Catanzaro, B. 2025. Audio flamingo 2: An audio-language model with long-audio understanding and expert reasoning abilities. *arXiv preprint arXiv:2503.03983*.

Ghosh, S.; Kumar, S.; Seth, A.; Evuru, C. K. R.; Tyagi, U.; Sakshi, S.; Nieto, O.; Duraiswami, R.; and Manocha, D. 2024. Gama: A large audio-language model with advanced audio understanding and complex reasoning abilities. arXiv preprint arXiv: 240611768. *arXiv preprint arXiv:2406.11768*.

Gong, Y.; Luo, H.; Liu, A. H.; Karlinsky, L.; and Glass, J. 2023. Listen, think, and understand. *arXiv preprint arXiv:2305.10790*.

Guo, Y.; Liu, J.; Li, M.; Cheng, D.; Tang, X.; Sui, D.; Liu, Q.; Chen, X.; and Zhao, K. 2025. Vtg-llm: Integrating timestamp knowledge into video llms for enhanced video temporal grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 3, 3302–3310.

Hershey, S.; Ellis, D. P.; Fonseca, E.; Jansen, A.; Liu, C.; Moore, R. C.; and Plakal, M. 2021. The benefit of temporally-strong labels in audio event classification. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 366–370. IEEE.

Hu, Y.; Ganter, T.; Deilamsalehy, H.; Dernoncourt, F.; Foroosh, H.; and Liu, F. 2023. MeetingBank: A benchmark dataset for meeting summarization. *arXiv preprint arXiv:2305.17529*.

Huang, B.; Wang, X.; Chen, H.; Song, Z.; and Zhu, W. 2024. Vtimellm: Empower llm to grasp video moments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14271–14280.

Kang, W.; and Roy, D. 2024. Prompting large language models with audio for general-purpose speech summarization. *arXiv preprint arXiv:2406.05968*.

Li, G.; Liu, J.; Dinkel, H.; Niu, Y.; Zhang, J.; and Luan, J. 2025. Reinforcement learning outperforms supervised fine-tuning: A case study on audio question answering. *arXiv preprint arXiv:2503.11197*.

Li, Y.; Guo, Z.; Wang, X.; and Liu, H. 2024. Advancing multi-grained alignment for contrastive language-audio pre-training. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 7356–7365.

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, 74–81.

Lipping, S.; Sudarsanam, P.; Drossos, K.; and Virtanen, T. 2022. Clotho-aqa: A crowdsourced dataset for audio question answering. In *2022 30th European Signal Processing Conference (EUSIPCO)*, 1140–1144. IEEE.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36: 34892–34916.

Mesaros, A.; Heittola, T.; and Virtanen, T. 2016. Metrics for polyphonic sound event detection. *Applied Sciences*, 6(6): 162.

Mesaros, A.; Heittola, T.; Virtanen, T.; and Plumbley, M. D. 2021. Sound event detection: A tutorial. *IEEE Signal Processing Magazine*, 38(5): 67–83.

Nallapati, R.; Zhou, B.; Gulcehre, C.; Xiang, B.; et al. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. *arXiv preprint arXiv:1602.06023*.

Panayotov, V.; Chen, G.; Povey, D.; and Khudanpur, S. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 5206–5210. IEEE.

Primus, P.; Schmid, F.; and Widmer, G. 2025. TACOS: Temporally-aligned Audio CaptiOnS for Language-Audio Pretraining. *arXiv preprint arXiv:2505.07609*.

Radford, A.; Kim, J. W.; Xu, T.; Brockman, G.; McLeavey, C.; and Sutskever, I. 2023a. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, 28492–28518. PMLR.

Radford, A.; Kim, J. W.; Xu, T.; Brockman, G.; McLeavey, C.; and Sutskever, I. 2023b. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, 28492–28518. PMLR.

Sakshi, S.; Tyagi, U.; Kumar, S.; Seth, A.; Selvakumar, R.; Nieto, O.; Duraiswami, R.; Ghosh, S.; and Manocha, D. 2024. Mmau: A massive multi-task audio understanding and reasoning benchmark. *arXiv preprint arXiv:2410.19168*.

Shang, H.; Li, Z.; Guo, J.; Li, S.; Rao, Z.; Luo, Y.; Wei, D.; and Yang, H. 2024. An end-to-end speech summarization using large language model. *arXiv preprint arXiv:2407.02005*.

Sridhar, A. K.; Guo, Y.; and Visser, E. 2024. Enhancing temporal understanding in audio question answering for large audio language models. *arXiv preprint arXiv:2409.06223*.

Tang, C.; Yu, W.; Sun, G.; Chen, X.; Tan, T.; Li, W.; Lu, L.; Ma, Z.; and Zhang, C. 2023. Salmonn: Towards generic hearing abilities for large language models. *arXiv preprint arXiv:2310.13289*.

Team, Q. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Wang, H.; Mao, J.; Guo, Z.; Wan, J.; Liu, H.; and Wang, X. 2023. Leveraging Language Model Capabilities for Sound Event Detection. *arXiv preprint arXiv:2308.11530*.

Wang, H.; Xu, Z.; Cheng, Y.; Diao, S.; Zhou, Y.; Cao, Y.; Wang, Q.; Ge, W.; and Huang, L. 2024. Grounded-videollm: Sharpening fine-grained temporal grounding in video large language models. *arXiv preprint arXiv:2410.03290*.

Wu, M.; Dinkel, H.; and Yu, K. 2019. Audio caption: Listen and tell. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 830–834. IEEE.

Wu, Y.; Chen, K.; Zhang, T.; Hui, Y.; Berg-Kirkpatrick, T.; and Dubnov, S. 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)*, 1–5. IEEE.

Wu, Y.; Tsirigotis, C.; Chen, K.; Huang, C.-Z. A.; Courville, A.; Nieto, O.; Seetharaman, P.; and Salamon, J. 2025. Flam: Frame-wise language-audio modeling. *arXiv preprint arXiv:2505.05335*.

Xie, Z.; Xu, X.; Wu, Z.; and Wu, M. 2025. Audiotime: A temporally-aligned audio-text benchmark dataset. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

Xu, J.; Guo, Z.; He, J.; Hu, H.; He, T.; Bai, S.; Chen, K.; Wang, J.; Fan, Y.; Dang, K.; et al. 2025. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*.

Xu, X.; Ma, Z.; Wu, M.; and Yu, K. 2024. Towards weakly supervised text-to-audio grounding. *IEEE Transactions on Multimedia*.

Yang, S.; Chen, Y.; Tian, Z.; Wang, C.; Li, J.; Yu, B.; and Jia, J. 2025. Visionzip: Longer is better but not necessary in vision language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 19792–19802.

Zhang, Y.; Han, W.; Qin, J.; Wang, Y.; Bapna, A.; Chen, Z.; Chen, N.; Li, B.; Axelrod, V.; Wang, G.; et al. 2023. Google usm: Scaling automatic speech recognition beyond 100 languages. *arXiv preprint arXiv:2303.01037*.