

AUGMENTING LARGE AUDIO-LANGUAGE MODELS WITH FRAME-LEVEL GROUNDING FOR FINE-GRAINED TEMPORAL PERCEPTION

Yanfeng Shi¹, Yan Song¹, Junhui Li¹, Tinggan Huang¹, Wu Guo¹, Haoyu Song², Ian McLoughlin²

¹National Engineering Research Center of Speech and Language Information Processing,
University of Science and Technology of China, Hefei, China

²ICT Cluster, Singapore Institute of Technology, Singapore

ABSTRACT

Large Audio-Language Models (LALMs) have substantially advanced general audio understanding, yet they remain limited in fine-grained temporal perception, particularly in precise event localization. Existing approaches primarily post-train LALMs to predict event boundaries as timestamp tokens. However, this generative formulation lacks explicit correspondence between the timestamp predictions and fine-grained acoustic evidence, limiting the precision and reliability of temporal localization. To address this issue, we augment the LALM with a dedicated frame-level grounding model while leveraging its semantic modeling capability to represent the event query. Specifically, the frozen LALM encodes the event query with audio as context, and the grounding model combines these query representations with fine-grained audio features to localize the target event at the frame level. Extensive experiments across diverse temporal grounding benchmarks demonstrate strong and consistent improvements over existing methods. Further evaluation shows that the grounding model can provide temporal evidence to support downstream reasoning.¹

Index Terms— Large audio-language models, frame-level localization, temporal audio grounding

1. INTRODUCTION

Large Audio-Language Models (LALMs) have significantly advanced general audio understanding by integrating acoustic perception with the semantic knowledge and reasoning capabilities of large language models [1, 2, 3]. Through multimodal pretraining and instruction tuning, LALMs acquire the ability to interpret diverse acoustic content, including speech, environmental sounds, and music, guided by natural language instructions. However, existing LALMs remain limited in fine-grained temporal perception, particularly in the precise localization of audio events [4]. This limitation restricts their applicability to temporal tasks such as audio grounding [5] and can further undermine downstream reasoning that depends on reliable temporal evidence, such as determining which sound event lasts the longest.

Recent work has sought to address this limitation by enhancing the temporal localization capabilities of LALMs [6, 7, 8]. Temporal information is encoded using special timestamp tokens [4] or incorporated into audio feature sequences through temporal markers [9], offering temporal references for localization. Meanwhile, fine-grained annotations provide supervision for temporal grounding [7], while reinforcement learning further refines localization pre-

cision [8]. Collectively, these efforts have substantially improved the temporal grounding performance of LALMs.

Notably, existing methods largely follow a common paradigm in which the LALM is post-trained for fine-grained temporal localization, with event boundaries predicted as timestamp tokens through its text generation interface. LALMs are well suited to interpreting natural language queries through their semantic representations, whereas precise temporal localization requires resolving event activity based on fine-grained acoustic evidence. Under the generative formulation, the correspondence between temporal predictions and fine-grained acoustic cues is learned implicitly, which can compromise the precision and reliability of temporal localization. By contrast, frame-level prediction directly models event activity over the audio timeline and has long been adopted in sound event detection [10] for precise temporal localization [11, 12]. However, how to augment LALMs through frame-level prediction to improve their fine-grained temporal perception while leveraging their semantic modeling capability remains underexplored.

Motivated by these observations, we propose a framework that leverages a frozen LALM to represent queries in context and delegates fine-grained temporal localization to a dedicated frame-level grounding model. Specifically, given the audio and query prompt as input, we extract hidden states from the LALM as query representations for temporal grounding. The grounding model combines these representations with frame-level audio features from a pretrained audio encoder through adaptive frame-level alignment and fusion, and the resulting sequence is then temporally modeled to predict frame-level localization scores over the audio timeline. Extensive experiments demonstrate strong performance across diverse temporal grounding benchmarks, supporting the effectiveness of augmenting LALMs with frame-level grounding. Further, we employ the grounding model as an external tool and observe improvements on downstream temporal reasoning.

2. METHOD

In this section, we detail the proposed framework illustrated in Fig. 1, including its application to downstream temporal reasoning.

2.1. LALM Query Representation

Rather than adapting the LALM for fine-grained temporal localization, we keep it frozen and leverage its semantic representations for grounding. The LALM can contextualize the event query with the input audio and localization instruction. Accordingly, we use its hidden states to condition the subsequent frame-level grounding model.

¹Examples of temporal grounding and reasoning are available at <https://saber5203.github.io/frame-level-grounding/>.

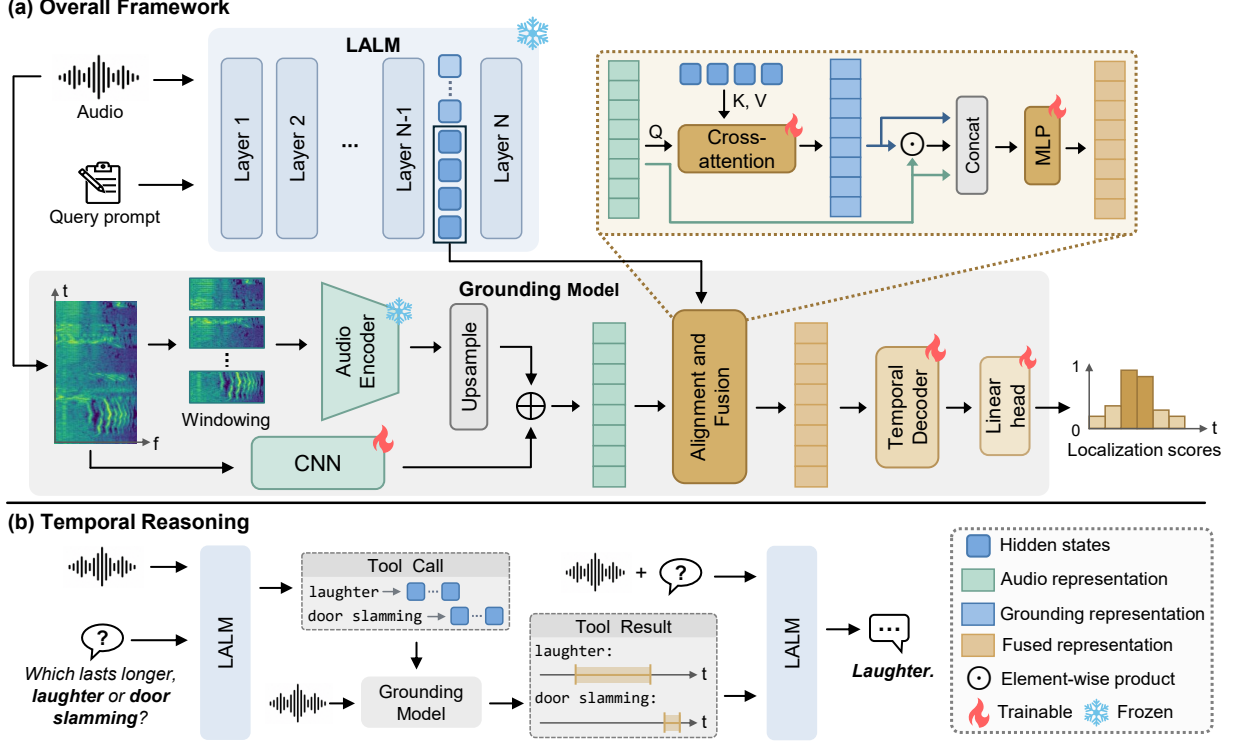


Fig. 1. The proposed framework is illustrated in (a), and (b) presents the grounding model as an external tool that provides temporal evidence for LALM reasoning.

Specifically, the audio and query prompt are provided as input to the LALM: “<audio> When does “<query>” occur in the audio?”, where <query> is replaced by the target event description and <audio> by the audio embeddings produced by the LALM’s native audio encoder. We use hidden states from the penultimate layer, which have been reported to encode richer semantic information than those from the final layer [13]. From this sequence, only the hidden states corresponding to the textual tokens are retained. These states encode the query and localization instruction while incorporating the preceding audio context through self-attention. We discard the audio-side hidden states because frame-level acoustic information is provided separately by a pretrained audio encoder. The retained states are then projected to the shared feature space:

$$\tilde{H} = \text{Proj}_h(H_{\text{text}}) \in \mathbb{R}^{L \times d}, \quad (1)$$

where H_{text} denotes the extracted textual hidden states, L is the number of textual tokens, and d is the shared feature dimension.

2.2. Frame-Level Grounding

Audio Representation. Precise temporal localization requires audio features with strong event discrimination and fine temporal structure. Pretrained audio encoders such as BEATs [14] and EAT [15] provide strong representations for audio events, and ATST-Frame [16] is further tailored to frame-level tasks. We therefore employ a pretrained audio encoder to obtain frame-level audio representations and keep it frozen during training.

To match the input configuration of the pretrained audio encoder, recordings are divided into non-overlapping 10s windows, with the final window zero-padded to 10s when necessary. Each window

is encoded independently, and the resulting features are concatenated along the temporal dimension to form the encoder sequence for the complete recording as $X_{\text{enc}} = \text{Concat}_t(\mathcal{E}(m_1), \dots, \mathcal{E}(m_K))$, where $\mathcal{E}$ denotes the audio encoder and $\{m_k\}_{k=1}^K$ are the log-Mel windows. X_{enc} is then upsampled to the target temporal resolution, while a CNN branch operating on the log-Mel spectrogram supplements the upsampled features with fine-grained acoustic details [17], resulting in the final audio representation:

$$X = \text{Proj}_x(X_{\text{enc}}^\uparrow + wX_{\text{cnn}}) = \{x_t\}_{t=1}^T \in \mathbb{R}^{T \times d}, \quad (2)$$

where $X_{\text{enc}}^\uparrow$ denotes the upsampled encoder features, X_{cnn} denotes the CNN features, w is a learnable scalar, and T is the number of frames.

Frame-Adaptive Alignment and Fusion. Preserving $\tilde{H}$ at the token level is important, since a natural language query may contain multiple semantic components, such as event identities and acoustic attributes, while a single audio frame may correspond to only a subset of them. Aggregating $\tilde{H}$ into a global representation would provide the same semantic condition to all frames and may obscure fine-grained correspondences between acoustic content and individual tokens [22]. We therefore leverage cross-attention to allow each audio frame to adaptively retrieve relevant semantic information from the token-level sequence $\tilde{H}$. Specifically, the frame-level audio representations X serve as the queries, while $\tilde{H}$ provides the keys and values:

$$G = \text{MHA}(X, \tilde{H}, \tilde{H}) = \{g_t\}_{t=1}^T \in \mathbb{R}^{T \times d}, \quad (3)$$

where MHA denotes multi-head attention, and G denotes the resulting frame-level grounding representations.

Table 1. Main temporal grounding results. R@.5 and R@.7 denote recall computed at IoU thresholds of 0.5 and 0.7, respectively, and mIoU denotes mean IoU. Best and second-best results are highlighted in bold and underlined.

Method	AudioGrounding			DESED			UnAV-100 subset			TACOS			Clotho-Moment		
	R@.5	R@.7	mIoU	R@.5	R@.7	mIoU	R@.5	R@.7	mIoU	R@.5	R@.7	mIoU	R@.5	R@.7	mIoU
AM-DETR [18]	15.4	5.7	29.6	20.3	9.6	33.7	57.0	37.0	50.3	7.8	3.4	21.6	85.8	80.3	79.5
FineLAP [19]	65.8	48.0	61.5	74.2	54.3	68.4	43.0	26.0	44.4	36.2	27.2	39.3	42.6	30.7	43.3
DASM [17]	68.8	48.8	64.0	81.0	64.7	73.9	32.0	11.0	39.1	42.4	32.0	46.7	18.9	5.9	32.8
MOSS-Audio [9]	37.1	19.8	42.0	43.8	20.5	45.2	43.0	32.0	45.3	25.6	11.5	30.3	26.4	13.7	31.5
Audio Flamingo 3 [20]	41.3	25.3	45.8	46.8	28.9	49.4	19.0	10.0	24.7	34.6	28.1	39.8	17.9	7.4	21.1
Qwen3-Omni [21]	66.6	45.2	60.9	72.8	52.0	65.1	75.0	53.0	66.2	47.1	38.5	49.6	46.9	37.1	50.6
TimeAudio [4]	75.3	60.3	67.2	70.6	57.7	68.0	18.0	8.0	23.0	45.4	38.7	49.6	30.3	12.0	29.6
SpotSound [7]	77.6	60.7	71.4	58.6	42.6	56.0	<u>81.0</u>	<u>67.0</u>	72.3	48.5	39.0	51.0	<u>91.8</u>	85.5	86.0
Ours (Audio Flamingo 3)	<u>84.4</u>	<u>67.7</u>	<u>76.1</u>	85.2	<u>72.4</u>	78.4	78.0	63.0	<u>72.7</u>	<u>53.2</u>	<u>41.9</u>	<u>54.9</u>	91.4	<u>86.7</u>	<u>87.1</u>
Ours (Qwen3-Omni)	85.3	68.8	76.4	<u>85.0</u>	72.8	<u>77.8</u>	83.0	70.0	74.1	53.7	42.7	55.6	92.8	88.5	88.9

To retain both representations while further modeling their interactions, the audio and grounding representations are fused at each frame as $z_t = \phi([x_t; g_t; x_t \odot g_t])$, forming the fused representations $Z = \{z_t\}_{t=1}^T \in \mathbb{R}^{T \times d}$, where $\odot$ denotes the element-wise product, $[\cdot; \cdot]$ denotes feature concatenation, and $\phi(\cdot)$ denotes the MLP used for feature fusion. The fused sequence Z is then processed by a temporal decoder to model dependencies across frames. The decoder outputs are mapped by a linear head to frame-level localization scores, which are subsequently converted into temporal intervals.

2.3. Temporal Reasoning

Beyond temporal grounding, we further explore whether the trained grounding model can provide temporal evidence for downstream audio reasoning. To this end, we use the grounding model as an external tool that the LALM can invoke during inference. Specifically, the LALM first identifies the temporal operation required by the question, which guides its subsequent use of the grounding model. When temporal evidence is useful, it formulates one or more concrete event queries and invokes the grounding model accordingly. For each tool call, the generated event queries are provided together with the audio to the LALM, from which the corresponding textual hidden states are extracted. These hidden states then condition the grounding model to predict the temporal intervals of the queried events. The resulting intervals are organized into a structured tool output together with derived temporal attributes, including duration, first onset, and last offset. The tool output is then incorporated into the LALM context as explicit temporal evidence for subsequent reasoning. If further evidence is required, the LALM can formulate additional event queries and invoke the grounding model again before producing the final answer.

3. EXPERIMENTS

3.1. Experimental Setup

Datasets. We conduct temporal grounding experiments on AudioGrounding (AG) [5], DESED [23], UnAV-100 [24], TACOS [25], Clotho-Moment (Clotho-M) [18], and FTAR [4]. This task aims to localize one or more temporal intervals of a target sound event in an audio recording. The query is provided as a natural-language event description, while event labels directly serve as queries in label-based datasets. For training, the corresponding training splits are

used, with DESED restricted to real recordings and only the temporal audio grounding portion of FTAR included. For evaluation, we use the test sets of AG, TACOS, and Clotho-M, the real test set of DESED, and the UnAV-100 subset [18]. The overall training and evaluation sets contain 98.1k and 14.5k samples, respectively.

Implementation Details. We use Audio Flamingo 3 [20] and Qwen3-Omni [21] as the LALM backbones and ATST-Frame [16] as the audio encoder. The shared feature dimension is set to 768, and the audio encoder features are upsampled to 50 Hz, with a four-layer CNN branch providing fine-grained acoustic details. The fusion weight w between the upsampled encoder features and CNN features is initialized to 0.5. The cross-attention uses 12 attention heads, and the temporal decoder consists of five Conformer [26] layers with a hidden dimension of 768. The grounding model is trained using frame-level binary cross-entropy loss, with binary frame labels derived from the annotated temporal intervals. During training, the LALM and ATST-Frame remain frozen, while the remaining parameters are optimized for 20 epochs using AdamW with a learning rate of 5×10^{-5} and a batch size of 32. At inference, the frame-level localization scores are smoothed with a 15-frame median filter and thresholded at 0.3, after which consecutive positive frames are merged into temporal intervals. We evaluate temporal grounding using IoU-based metrics.

3.2. Main Grounding Results

Table 1 presents the main temporal grounding results. All baselines are evaluated on the same test sets using their publicly released checkpoints. For FineLAP and DASM, longer recordings are split into non-overlapping 10s windows to match their input length. The strong results of TimeAudio and SpotSound confirm that task-specific post-training can effectively improve the temporal localization capability of LALMs. Our framework further improves performance with both LALM backbones, particularly at the stricter R@.7 threshold, with gains of 15.1 points on DESED and 8.1 points on AG over the best LALM baselines. The proposed framework also performs strongly against specialized grounding models, with pronounced gains on TACOS and Clotho-M, which involve richer natural language queries, achieving 55.6 versus 46.7 and 88.9 versus 79.5 mIoU, respectively. Overall, our framework consistently improves mIoU across all benchmarks, demonstrating the effectiveness of augmenting LALMs with frame-level grounding.

Table 2. Ablation study of key design choices in our framework. Results are reported in mIoU. Qwen3-Omni[†] refers to audio representations from its native audio encoder. Bold variants indicate the settings used for the main results.

Variant	AG	DESED	UnAV	TACOS	Clotho-M
<i>Query Representation</i>					
MGA-CLAP [27]	74.9	75.7	71.9	50.2	85.6
Qwen3-Omni (w/o audio context)	75.3	74.4	72.4	53.7	86.9
Qwen3-Omni	76.4	77.8	74.1	55.6	88.9
<i>Audio Representation</i>					
Qwen3-Omni [†]	73.9	75.9	72.2	52.7	87.8
ATST-Frame (w/o CNN branch)	75.2	76.5	74.5	54.3	88.1
ATST-Frame	76.4	77.8	74.1	55.6	88.9
<i>Query-to-Audio Alignment</i>					
Global Pooling	75.0	76.8	73.8	54.0	87.3
Cross-Attention	76.4	77.8	74.1	55.6	88.9

3.3. Ablation Studies

Table 2 summarizes the ablation results using Qwen3-Omni as the LALM backbone. To examine the role of LALM-based query modeling in our framework, we compare alternative query representations for conditioning the grounding model while keeping the remaining setup unchanged. MGA-CLAP serves as a standalone query encoder in place of the LALM, as it is pretrained via audio-text contrastive learning and provides word-level features [27]. To isolate the contribution of audio contextualization within the LALM, we further remove the audio input to Qwen3-Omni while retaining the same textual input. The results show that the w/o audio context variant remains competitive with MGA-CLAP, while introducing audio context further improves performance. This highlights the value of audio context in refining LALM-based query representations for temporal grounding.

We then investigate the effect of audio representations in our framework. Using the native audio encoder of Qwen3-Omni in place of ATST-Frame leads to consistently lower performance across the benchmarks, indicating that representations from a specialized audio encoder are better suited to the fine temporal discrimination required for precise localization. We further evaluate the contribution of the CNN branch by removing it from the audio representation. ATST-Frame alone already provides strong performance, and incorporating the CNN features brings further improvements overall. These results support the use of a specialized audio encoder, with the CNN branch contributing complementary fine-grained acoustic information.

Further, we evaluate the importance of aligning the token-level hidden states with audio representations at the frame level. As a global alternative, we mean-pool the textual hidden states into a single query representation that is shared across all audio frames in the subsequent fusion. In contrast, cross-attention performs frame-level alignment between hidden states and audio representations. The consistent advantage of cross-attention over global pooling indicates that temporal grounding benefits from frame-specific alignment with token-level query representations rather than using a single global semantic condition across all frames.

Table 3. Temporal reasoning performance of direct LALM inference and using the grounding model as an external tool. #Q denotes the number of questions in each category. The better result in each category is highlighted in bold.

Question Type	#Q	Accuracy (%)	
		Direct	w/ Grounding Model
Onset/Offset	144	53.47	82.64
Duration	106	50.94	83.02
Ordering	151	71.52	74.17
Counting	114	43.86	45.61
Other	94	67.02	67.02
Overall	609	57.80	71.26

3.4. Temporal Reasoning

We evaluate downstream temporal reasoning on the development split of the temporal soundscapes QA subset in DCASE 2025 Task 5 [28]. This benchmark contains multiple-choice questions designed to assess audio temporal reasoning. For analysis, we group the questions according to the required temporal operation into Onset/Offset, Duration, Ordering, Counting, and Other categories. We compare direct Qwen3-Omni inference with the same LALM using the trained grounding model as an external tool.

As shown in Table 3, incorporating the grounding model improves the overall accuracy from 57.80% to 71.26%. The largest improvements are observed for Duration and Onset/Offset, with gains of 32.08 and 29.17 percentage points, since these questions depend on explicit temporal information such as event boundaries or durations, which can be obtained directly from the grounding results. For Ordering questions, whose answers are determined by the occurrence order of sound events, Qwen3-Omni already achieves relatively strong performance, while the grounding model provides a modest improvement. The gains for Counting are also limited, as the frame-level localization objective does not explicitly supervise the number of event occurrences, and merged or fragmented intervals can lead to inaccurate counts. Overall, these results highlight the value of frame-level grounding in providing temporal evidence for downstream reasoning.

4. CONCLUSION

In this paper, we presented a framework that augments LALMs with frame-level grounding to strengthen their fine-grained temporal perception. Our framework leverages semantic representations from a frozen LALM to encode the event query with audio as context and combines them with fine-grained audio features in a dedicated grounding model for direct frame-level localization. Extensive experiments demonstrate strong and consistent improvements across diverse grounding benchmarks, with particularly pronounced gains under stricter localization criteria. Further, improvements in downstream temporal reasoning indicate that fine-grained localization can provide explicit temporal evidence for subsequent reasoning. These findings suggest that frame-level grounding offers a complementary route toward fine-grained temporal perception in LALMs.

5. REFERENCES

- [1] Y. Gong, H. Luo, A. H. Liu, L. Karlinsky, and J. R. Glass, “Listen, think, and understand,” in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2024.
- [2] C. Tang et al., “SALMONN: Towards generic hearing abilities for large language models,” in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2024.
- [3] Z. Kong, A. Goel, R. Badlani, W. Ping, R. Valle, and B. Catanzaro, “Audio Flamingo: A novel audio language model with few-shot learning and dialogue abilities,” in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2024, pp. 25125–25148.
- [4] H. Wang, Y. Li, S. Ma, H. Liu, and X. Wang, “Listening between the frames: Bridging temporal gaps in large audio-language models,” in *Proc. AAAI Conf. Artif. Intell.*, 2026, pp. 26233–26241.
- [5] X. Xu, H. Dinkel, M. Wu, and K. Yu, “Text-to-audio grounding: Building correspondence between captions and sound events,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2021, pp. 606–610.
- [6] J. An, P. Keung, J. Wang, O. Ahia, and N. A. Smith, “Encode once, decode never: Reusing audio LM internals for efficient temporal localization,” in *Proc. Conf. Lang. Model. (COLM)*, 2026, to appear.
- [7] L. Sun, X. Zhou, Z. Li, Y. Zhang, Y. Wang, and W. Xie, “Spot-Sound: Enhancing large audio-language models with fine-grained temporal grounding,” in *Proc. ACM Int. Conf. Multimedia (ACM MM)*, 2026, to appear.
- [8] Y. Shi et al., “Towards fine-grained temporal perception: Post-training large audio-language models with audio-side time prompt,” in *Proc. Interspeech*, 2026, to appear.
- [9] C. Yang et al., “MOSS-Audio technical report,” *arXiv preprint arXiv:2606.01802*, 2026.
- [10] A. Mesaros, T. Heittola, T. Virtanen, and M. D. Plumbley, “Sound event detection: A tutorial,” *IEEE Signal Process. Mag.*, vol. 38, no. 5, pp. 67–83, 2021.
- [11] E. Cakir, G. Parascandolo, T. Heittola, H. Huttunen, and T. Virtanen, “Convolutional recurrent neural networks for polyphonic sound event detection,” *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 25, no. 6, pp. 1291–1303, 2017.
- [12] N. Shao, X. Li, and X. Li, “Fine-tune the pretrained ATST model for sound event detection,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2024, pp. 911–915.
- [13] Z. Tian et al., “Audio-Omni: Extending multi-modal understanding to versatile audio generation and editing,” in *Proc. ACM SIGGRAPH Conf. Papers*, 2026, pp. 1–10.
- [14] S. Chen et al., “BEATs: Audio pre-training with acoustic tokenizers,” in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2023, pp. 5178–5193.
- [15] W. Chen, Y. Liang, Z. Ma, Z. Zheng, and X. Chen, “EAT: Self-supervised pre-training with efficient audio transformer,” in *Proc. Int. Joint Conf. Artif. Intell. (IJCAI)*, 2024, pp. 3807–3815.
- [16] X. Li, N. Shao, and X. Li, “Self-supervised audio teacher-student transformer for both clip-level and frame-level tasks,” *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 32, pp. 1336–1351, 2024.
- [17] P. Cai, Y. Song, Q. Gu, N. Jiang, H. Song, and I. McLoughlin, “Detect any sound: Open-vocabulary sound event detection with multi-modal queries,” in *Proc. ACM Int. Conf. Multimedia (ACM MM)*, 2025, pp. 582–591.
- [18] H. Munakata, T. Nishimura, S. Nakada, and T. Komatsu, “Language-based audio moment retrieval,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2025, pp. 1–5.
- [19] X. Li et al., “FineLAP: Taming heterogeneous supervision for fine-grained language-audio pretraining,” in *Proc. Annu. Meeting Assoc. Comput. Linguistics (ACL)*, 2026, pp. 10393–10408.
- [20] S. Ghosh et al., “Audio Flamingo 3: Advancing audio intelligence with fully open large audio language models,” in *Adv. Neural Inf. Process. Syst.*, 2025, vol. 38, pp. 41819–41886.
- [21] J. Xu et al., “Qwen3-Omni technical report,” *arXiv preprint arXiv:2509.17765*, 2025.
- [22] H. Tang et al., “Listen as you wish: Fusion of audio and text for cross-modal event detection in smart cities,” *Inf. Fusion*, vol. 110, 2024, Art. no. 102460.
- [23] N. Turpault, R. Serizel, A. P. Shah, and J. Salamon, “Sound event detection in domestic environments with weakly labeled data and soundscape synthesis,” in *Proc. Detection Classification Acoust. Scenes Events Workshop (DCASE)*, 2019, pp. 253–257.
- [24] T. Geng, T. Wang, J. Duan, R. Cong, and F. Zheng, “Dense-localizing audio-visual events in untrimmed videos: A large-scale benchmark and baseline,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 22942–22951.
- [25] P. Primus, F. Schmid, and G. Widmer, “TACOS: Temporally-aligned audio CaptiOnS for language-audio pretraining,” in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust. (WASPAA)*, 2025, pp. 1–5.
- [26] A. Gulati et al., “Conformer: Convolution-augmented transformer for speech recognition,” in *Proc. Interspeech*, 2020, pp. 5036–5040.
- [27] Y. Li, Z. Guo, X. Wang, and H. Liu, “Advancing multi-grained alignment for contrastive language-audio pre-training,” in *Proc. ACM Int. Conf. Multimedia (ACM MM)*, 2024, pp. 7356–7365.
- [28] C.-H. H. Yang et al., “Multi-domain audio question answering benchmark toward acoustic content reasoning,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2026, pp. 22517–22521.