

Acoustic Prompting via Stage-wise Modulation for Few-Shot Learning in Audio Language Models

Hyebin Cho, Jaehyuk Jang, Changick Kim**, Joon Son Chung**

Korea Advanced Institute of Science and Technology, South Korea

{hyebin.cho, jhyuk, changick, joonson}@kaist.ac.kr

Abstract

Audio-Language Models (ALMs) have shown remarkable success in zero-shot audio classification by aligning audio waveforms with text. Recent efforts to improve downstream performance focus on learning optimal text prompts. However, previous approaches focus on the text encoder, leaving the potential of learnable prompts within the audio encoder unexplored. In this paper, we propose a novel framework that introduces trainable prompts into the audio encoder to capture task-specific acoustic features. We demonstrate that integrating audio-side prompt learning with existing text-side approaches enhances few-shot adaptation. Through extensive experiments across 11 datasets show that integrating our method as a plug-and-play module alongside existing text prompt tuning generally leads to performance improvements. These findings suggest that explicitly modulating the audio representation space effectively complements text-only prompting approaches. The code is available at <https://github.com/hyebin-c/aspl>.

Index Terms: audio language model, prompt learning, few-shot learning

1. Introduction

Recent advances in large-scale multimodal foundation models have fundamentally transformed the landscape of audio understanding. In particular, Audio-Language Models (ALMs) have emerged as powerful tools, enabling highly effective open-vocabulary recognition across diverse acoustic domains. By aligning audio waveforms with natural language descriptions in a shared embedding space, these models, such as Contrastive Language–Audio Pretraining (CLAP) [1], AudioCLIP [2], and Wav2CLIP [3], have enabled open-vocabulary capabilities allowing for the classification of sounds without the need for task-specific retraining.

While generative ALMs like Qwen2-Audio [4] and SALMONN [5] have also shown promise in audio-to-text generation, contrastive learning-based models remain the cornerstone for discriminative tasks due to their cross-modal alignment. Trained on large-scale datasets [6], CLAP leverages the HTSAT audio encoder [7] to bridge diverse acoustic domains with their semantic captions, demonstrating state-of-the-art performance across various downstream applications [8, 9, 10, 11, 12, 13].

To adapt these large-scale foundation models to specific downstream tasks, prompt learning has emerged as a parameter-efficient alternative to full fine-tuning [14, 15]. By prepending learnable soft prompts to the input data, researchers have successfully steered models to perform better in few-shot and

zero-shot scenarios. However, existing prompt learning research for ALMs [16, 17, 18] has been heavily text-centric. This bias stems from the inherent stability of the text embedding space, where text serves as a discrete and semantically consistent anchor, making it a reliable target for optimization. In contrast, audio signals are continuous, high-dimensional, and often exhibit significant intra-class variance and background noise, leading many to believe that optimizing the audio encoder is less efficient or more prone to instability.

We argue that this text-centric bias introduces a fundamental bottleneck. Restricting adaptation to the text modality while keeping the audio representation fixed constrains the model’s ability to achieve optimal cross-modal alignment. In this context, mutual adaptation, in which both audio and text representations are jointly refined, can be beneficial for capturing the nuanced characteristics of task-specific audio data. Text-only prompts cannot fully bridge the acoustic gap caused by domain shifts between the massive pretraining corpus and diverse downstream datasets. To the best of our knowledge, the potential of learnable prompts within the audio encoder remains unexplored in the context of few-shot audio classification.

In this paper, we propose a novel plug-in framework for **Audio-Side Prompt Learning (ASPL)** that can be seamlessly integrated with existing text-side counterparts. Our approach injects learnable prompt tokens at three strategic locations within the audio pipeline to capture task-specific acoustic features: (1) the log-mel spectrogram transformation stage, (2) the stage immediately following patch embedding, and (3) the early layers of the Swin Transformer blocks. This multi-level prompting strategy allows the model to adjust its perception of sound at both the low-level acoustic and high-level structural stages.

2. Method

2.1. Preliminary: Audio-Text Alignment in ALMs

Pre-trained ALMs align modality-specific representations in a shared space via cosine similarity. Given an audio waveform x , the audio encoder extracts a global representation $\mathbf{z}_a = E_a(x)$. For a C -class task, textual descriptions $\{T_i\}_{i=1}^C$ are generated using a template (e.g., “A recording of a {class}”) and encoded into target embeddings $\mathbf{z}_{t,i} = E_t(T_i)$. The predicted class index \hat{y} is then determined by maximizing the cosine similarity between the audio and text embeddings:

$$\hat{y} = \operatorname{argmax}_{i \in \{1, \dots, C\}} \frac{\langle \mathbf{z}_a, \mathbf{z}_{t,i} \rangle}{\|\mathbf{z}_a\| \|\mathbf{z}_{t,i}\|}, \quad (1)$$

where $\langle \cdot, \cdot \rangle$ and $\|\cdot\|$ denote the inner product and L_2 norm, respectively.

**indicates the corresponding author.

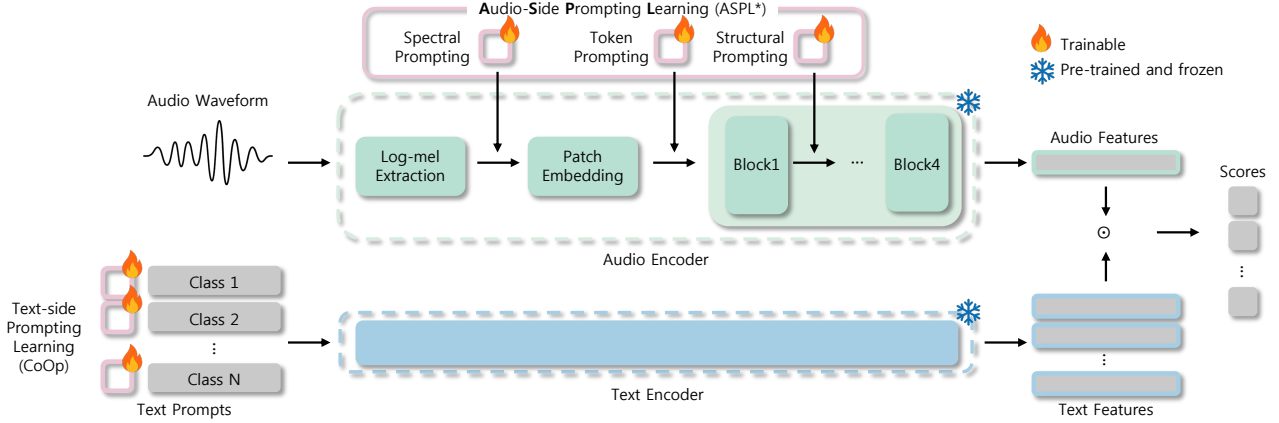


Figure 1: **Overview of the proposed ASPL framework.** Built upon the CLAP-HTSAT architecture, our method performs three-stage modulation on the audio encoder, representing the full ASPL* configuration. ASPL is designed as a plug-and-play framework compatible with various text-side methods; while the figure illustrates integration with CoOp, it can also be applied to CoCoOp or PALM. The fire (🔥) and ice (❄️) icons denote trainable parameters and frozen pre-trained encoders, respectively.

2.2. Text-side Prompt Learning

Previous research has predominantly focused on optimizing the text encoder to improve cross-modal alignment. For instance, **CoOp** [14] and **CoCoOp** [15] introduce static and dynamic learnable prompt tokens, respectively. More recently, **PALM** [18] expanded the definition of prompting beyond discrete token concatenation by demonstrating that lightweight affine transformations on text features act effectively as prompt learning. Despite their success, these methods exclusively refine text representations, leaving the potential for analogous learnable prompts within the *audio encoder* unexplored.

2.3. Audio-Side Prompt Learning (ASPL)

Inspired by prior approaches like PALM, which successfully established affine transformations as a parameter-efficient prompting mechanism, we propose Audio-Side Prompt Learning (ASPL). As illustrated in Fig. 1, ASPL functions as a plug-in framework that establishes structural symmetry with text-side prompting. ASPL’s core novelty lies in repurposing this affine mechanism as a stage-wise *acoustic prompt*. Rather than introducing heavily parameterized adapter modules or applying arbitrary network conditioning, we specifically target the early acoustic-to-semantic transition phases: log-mel extraction, patch embedding, and the initial Swin Transformer block. This targeted placement allows the model to recalibrate low-level acoustic variations (e.g., spectral shifts) at the perceptual level. By injecting task-specific inductive biases exactly where the audio encoder builds its abstractions, we effectively prevent physical nuances from disrupting the pre-trained semantic knowledge in deeper layers. Formally, for a 3D input feature X at these targeted stages, we apply a lightweight prompt-driven modulation:

$$X' = \gamma \odot X + \beta, \quad (2)$$

where γ and β are learnable 1D continuous prompt vectors broadcasted channel-wise across the remaining temporal or spatial dimensions. Unlike text-side methods such as PALM that require class-specific parameters, our acoustic prompt vectors are shared across all classes and instances. This architectural distinction ensures extreme parameter efficiency that remains strictly constant regardless of the dataset’s label space complex-

ity. Instead of acting merely as scaling factors, these parameters function as *continuous acoustic prompts*. By optimizing these vectors in a continuous latent space, they enable the model to recalibrate spectral and structural cues for downstream tasks with minimal parameter overhead.

2.3.1. Acoustic Prompting via Spectral Modulation

The first ASPL stage operates directly after log-mel spectrogram $X \in \mathbb{R}^{B \times F \times T}$ (batch, frequency, time). Instead of appending discrete tokens, we introduce continuous acoustic prompts $(\gamma_{spec}, \beta_{spec}) \in \mathbb{R}^F$, which are broadcasted over B and T to perform frequency-wise scaling. This mechanism functions as a learnable equalizer that prompts the model to adapt to variations in energy distribution and recording conditions. By doing so, it efficiently steers the acoustic representations to adapt to spectral variations and recording conditions without requiring deep learning of acoustic bias.

2.3.2. Token-Level Prompting at the Patch Embedding Interface

The second ASPL stage modulates the latent tokens $X \in \mathbb{R}^{B \times L \times C}$ (batch, token length, channel) generated by the Patch Embedding layer. By performing channel-wise recalibration via the learnable parameters $(\gamma_{tok}, \beta_{tok}) \in \mathbb{R}^C$ broadcasted over B and L , we adaptively emphasize task-specific acoustic patterns. This interface-level refinement seamlessly bridges the transition from local convolutional features to global attention mechanisms, potentially aligning the token space with the target domain.

2.3.3. Structural Prompting via Early-Layer Conditioning

The final structural modulation stage targets the output of the initial Swin Transformer block. By broadcasting parameters $(\gamma_{block1}, \beta_{block1}) \in \mathbb{R}^C$ across batch B and sequence length L , we condition representations immediately after the first local-global mixing. This task-specific bias stabilizes early features, adapting to acoustic nuances while preventing semantic drift in deeper layers.

Table 1: *Average performance comparison on 11 datasets. We report the average top-1 accuracy (%) in a 16-shot setting across three independent runs (seeds 0, 1, 2). **Bold** and underlined denote the best and second-best results, respectively. Our ASPL variants provide overall performance gains over text-side prompt learning baselines.*

Method	Beijing Opera	CRE MA-D	ESC50	ESC50 Actions	GT Music Genre	NS Instruments	RAVDESS	SESA	TUT 2017	Urban Sound	Vocal Sound	Average
CoOp	96.18	39.44	93.80	94.92	74.50	61.65	40.26	89.52	68.20	76.10	74.57	73.56
+ ASPL	96.91	35.12	94.10	96.25	76.17	63.68	42.09	88.89	74.90	76.30	74.11	<u>74.41</u>
+ ASPL*	96.75	37.75	94.35	95.50	78.00	65.68	42.77	91.11	74.38	77.40	77.28	75.54
CoCoOp	98.17	39.87	94.72	97.75	77.83	64.00	43.11	89.84	75.84	78.67	81.17	76.45
+ ASPL	98.60	40.45	95.37	97.83	79.67	64.76	46.64	92.38	78.68	79.65	82.27	77.85
+ ASPL*	97.88	38.06	95.32	97.42	79.83	64.58	43.72	93.02	78.76	79.72	82.13	<u>77.31</u>
PALM	98.74	36.15	96.33	98.25	80.33	66.24	46.16	90.79	79.80	80.97	82.64	77.86
+ ASPL	98.59	38.08	96.52	97.75	82.17	66.55	49.76	92.06	82.51	81.73	83.03	<u>78.98</u>
+ ASPL*	97.89	43.41	96.58	98.00	80.00	65.19	48.20	94.61	83.14	82.21	82.60	79.26

2.3.4. Multi-modal Alignment and Practical Scalability

Together, these three prompting stages systematically modulate distinct levels of acoustic abstraction—from raw spectral features to deeper structural representations. When integrated with existing text-side prompting methods (e.g., CoOp, PALM), as a plug-and-play module, ASPL forms a comprehensive dual-sided prompting framework. This jointly refines both modalities, promoting balanced cross-modal alignment rather than forcing the text encoder to adapt to a frozen, static audio space. Consequently, this prompt-driven joint optimization enhances few-shot classification, while its extreme parameter efficiency provides a scalable alternative to full fine-tuning for large foundation models.

3. Experiments

3.1. Datasets

To comprehensively evaluate the effectiveness of our proposed method, we conducted experiments on 11 diverse audio classification datasets, following the benchmark suite and evaluation protocol established in PALM. The datasets cover a wide range of tasks, including: instrument classification (Beijing-Opera [19], NS-Instruments [20]); sound event classification (ESC-50 [21], ESC-50-Action [21], UrbanSound8K [22]); emotion recognition (CREMA-D [23], RAVDESS [24]); vocal sound classification (VocalSound [25]); surveillance event classification (SESA [26]); acoustic scene classification (TUT2017 [27]); and music analysis (GT-Music-Genre [28]). We utilized the official train-test splits where available, and for datasets without official splits, we employed a multi-fold cross-validation strategy, reporting the average accuracy.

3.2. Implementation Details

Baselines and Architecture. We compared our proposed ASPL against state-of-the-art text-side prompt learning methods: CoOp, CoCoOp, and PALM. Following PALM [18], we utilized CLAP-HTSAT as the audio encoder and the CLIP text encoder [29] from pretrained PENGI [30]. During training, both encoders were frozen, updating only the prompt parameters and ASPL modules.

Training Settings. All experiments followed a standard few-shot protocol, where $K \in \{1, 2, 4, 8, 16\}$ training samples were randomly selected per class. We fixed random seeds to

$\{0, 1, 2\}$ for three independent runs and reported the average Top-1 accuracy. On the text-side, a single prompt context was utilized without ensembling. All models were trained for 100 epochs using cross-entropy loss and SGD (learning rate: 0.01, momentum: 0.9, weight decay: 0) with a batch size of 16 on an NVIDIA RTX A5000. Both audio and text embeddings were L_2 -normalized, and a fixed temperature $\tau = 0.01$ was applied to the cosine similarity scores prior to the softmax layer, following the established prompt optimization protocol of PALM.

3.3. Main Results

Table 1 presents the quantitative comparison across 11 datasets. We designate our approach utilizing spectral and token modulations as **ASPL**, while **ASPL*** refers to the extended configuration incorporating early-layer structural conditioning. To ensure a fair comparison, our baselines focus exclusively on training-based adaptation methods.

The results demonstrate that our audio-side modulation provides a highly favorable performance-efficiency trade-off. Despite minor fluctuations per dataset, it achieves an average gain of $\sim 1.1\%$ with a minimal parameter overhead. Notably, the optimal modulation depth correlates with the adaptability of the base text encoder:

- **Synergy with Dynamic Prompts (CoCoOp):** For CoCoOp, which generates dynamic instance-conditioned text prompts, the lighter ASPL configuration achieves state-of-the-art performance. Since CoCoOp already handles instance-level variance, the focused spectral and token modulation of ASPL provides sufficient acoustic alignment without introducing redundancy.
- **Synergy with Static/Class Prompts (CoOp, PALM):** Conversely, for CoOp and PALM, which rely on static or class-specific prompts, the deeper structural modulation of ASPL* proves essential. These methods benefit significantly from the additional adaptability provided by early-layer conditioning, which compensates for the rigidity of the text representation.

In summary, our framework functions as a flexible plug-and-play module that effectively bridges the modality gap, with the deeper ASPL* serving as a powerful compensator for less flexible text backbones.

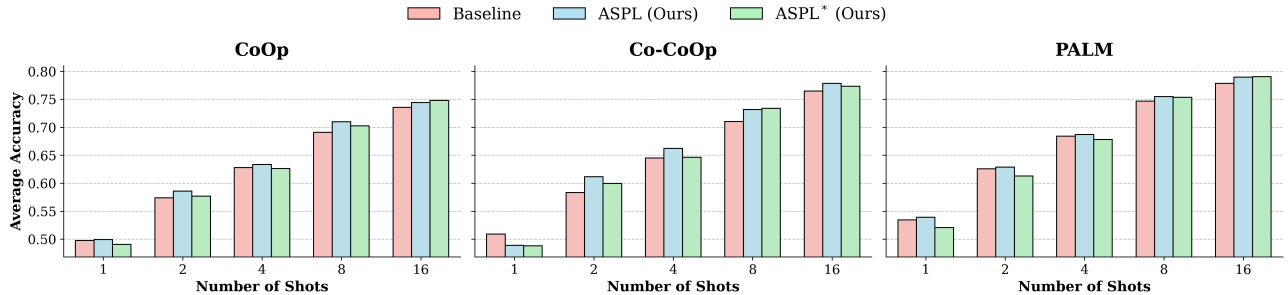


Figure 2: **Ablation study on the number of shots.** Comparison of average accuracy on 11 datasets for CoOp, CoCoOp, and PALM backbones. Our proposed ASPL variants (solid bars) consistently outperform the baselines (pink bars) across most few-shot settings, demonstrating robust adaptation capabilities.

Table 2: **Complexity and Performance Comparison.** Comparison of trainable parameters, inference latency (measured on CREMA-D), and average accuracy using the PALM baseline. Integrating ASPL and ASPL* improves overall accuracy while adding minimal parameter overhead and marginal latency variations (within the noise floor across 15 trials).

Method	# Params	Latency (ms)	Avg. Acc. (%)
CoOp	8,192	2.60	73.56
CoCoOp	107,072	14.76	76.45
PALM	4,100	2.60	77.86
+ ASPL	4,420	3.09	78.98
+ ASPL*	4,804	2.97	79.26

3.4. Ablation Study

Impact of Shot Availability. Figure 2 illustrates the performance of ASPL across varying shot counts. While we observe a slight performance degradation in the 1-shot setting—likely because a single example provides insufficient supervision to robustly optimize the continuous acoustic prompts without minor overfitting—our method consistently outperforms text-only baselines from the 2-shot setting onwards. This indicates that once a minimal threshold of data is available, our hierarchical prompting strategy effectively leverages the supervision to refine acoustic representations, yielding a steadily higher performance ceiling across 4, 8, and 16-shot setups.

Parameter Efficiency. A key advantage of our approach is the extreme parameter efficiency shown in Table 2. While baseline methods can introduce up to 107,072 additional parameters scaling with the number of classes, our method adds a negligible and fixed number of parameters regardless of dataset complexity: only 320 for ASPL and 704 for ASPL*. This extreme efficiency is achieved by deliberately targeting the early, low-dimensional stages of the audio encoder. Specifically, the parameter count directly corresponds to the affine pairs applied to 64-mel bins (128 parameters), the 96-dimensional initial token projection (192 parameters), and for ASPL*, the 192-dimensional first Transformer block (384 parameters). Remarkably, despite the negligible increase in parameters and inference latency, our method consistently yields an absolute accuracy gain of over 1.4%, proving highly suitable for resource-constrained, real-time audio applications.

Effect of Modulation Location. Table 3 demonstrates that our performance gains are rooted in deliberate architectural choices rather than arbitrary prompt insertion. While individual modulations in rows 2–3 cause slight degradation—suggesting that

Table 3: **Ablation study on modulation locations** (Baseline: CoOp). Combining spectral, token, and structural modulations yields the optimal synergy for audio-text alignment.

	Spectral	Token	Structural	Avg. Acc. (%)
1				73.56
2	✓			73.35
3		✓		72.96
4			Output Space	72.94
5	✓	✓		74.41
6	✓	✓	Output Space	73.90
7	✓	✓	Late Block	74.09
8	✓	✓	Early Block	75.54

adapting isolated components abruptly disrupts the internal representational flow of the frozen encoder—our integrated ASPL (Row 5) achieves a synergistic gain. This indicates that a coordinated, multi-stage recalibration is essential to smoothly bridge acoustic variations without causing internal covariate shifts. Importantly, simply replicating the output-space modulation used in PALM for the audio encoder (Rows 4 and 6) proves ineffective for audio signals, confirming that our success is not a byproduct of naive adaptation of text-centric methods. Furthermore, the effectiveness of early-block modulation in ASPL* (Row 8) over late-block modulation (Row 7) reinforces our core design principle that complex audio signals, unlike text, necessitate early-stage acoustic conditioning to address domain shifts while preserving pre-trained cross-modal knowledge.

4. Conclusion

In this paper, we introduced Audio-Side Prompt Learning (ASPL), a *plug-and-play* framework designed to overcome the text-centric limitations of prompt learning in current audio-language models. By strategically injecting lightweight, coordinated modulations at the early spectral, token, and structural levels, our approach facilitates a balanced mutual adaptation without disrupting the pre-trained representational flow. Extensive evaluations across 11 diverse datasets demonstrate that ASPL improves the overall average accuracy compared to text-only baselines while maintaining remarkable parameter efficiency. These results underscore that unlocking the plasticity of the audio encoder’s early stages is a vital, yet often overlooked, component for effective few-shot recognition. Moving forward, we aim to explore dynamic gating mechanisms to automatically calibrate modulation depth, further advancing the development of adaptive multi-modal foundation models.

5. Acknowledgements

This work was supported by IITP grants funded by the Korean government (MSIT, RS-2025-02263169, Detection and Prediction of Emerging and Undiscovered Voice Phishing).

6. Generative AI Use Disclosure

The authors acknowledge the use of Google's Gemini during the drafting of this paper for the purpose of language polishing and grammatical correction. This tool was employed exclusively to improve the linguistic quality of the text and played no role in formulating the research methodology, conducting experiments, or deriving scientific insights. The authors have thoroughly validated all phrasing and hold sole responsibility for the integrity and accuracy of the submitted manuscript.

7. References

- [1] B. Elizalde, S. Deshmukh, M. Al Ismail, and H. Wang, "Clap learning audio concepts from natural language supervision," in *Proc. ICASSP*. IEEE, 2023, pp. 1–5.
- [2] A. Guzhov, F. Raue, J. Hees, and A. Dengel, "Audioclip: Extending clip to image, text and audio," in *Proc. ICASSP*. IEEE, 2022, pp. 976–980.
- [3] H.-H. Wu, P. Seetharaman, K. Kumar, and J. P. Bello, "Wav2clip: Learning robust audio representations from clip," in *Proc. ICASSP*. IEEE, 2022, pp. 4563–4567.
- [4] Y. Chu, J. Xu, Q. Yang, H. Wei, X. Wei, Z. Guo, Y. Leng, Y. Lv, J. He, J. Lin *et al.*, "Qwen2-audio technical report," *arXiv preprint arXiv:2407.10759*, 2024.
- [5] C. Tang, W. Yu, G. Sun, X. Chen, T. Tan, W. Li, L. Lu, Z. Ma, and C. Zhang, "Salmonn: Towards generic hearing abilities for large language models," *arXiv preprint arXiv:2310.13289*, 2023.
- [6] Y. Wu, K. Chen, T. Zhang, Y. Hui, T. Berg-Kirkpatrick, and S. Dubnov, "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation," in *Proc. ICASSP*. IEEE, 2023, pp. 1–5.
- [7] K. Chen, X. Du, B. Zhu, Z. Ma, T. Berg-Kirkpatrick, and S. Dubnov, "Hts-at: A hierarchical token-semantic audio transformer for sound classification and detection," in *Proc. ICASSP*. IEEE, 2022.
- [8] J. Kim, M. Jeon, J. Jung, S. H. Woo, and J. Lee, "Enclap++: Analyzing the enclap framework for optimizing automated audio captioning performance," *Proc. DCASE workshop*, 2024.
- [9] D. Takeuchi, B. T. Nguyen, M. Yasuda, Y. Ohishi, D. Niizumi, and N. Harada, "Clap-art: Automated audio captioning with semantic-rich audio representation tokenizer," *Proc. INTERSPEECH*, 2025.
- [10] R. Yang, M. Rao, H. Sundar, A. Raju, A. Khare, S. Tankasala, D. He, and V. Ravichandran, "On retrieval of long audios with complex text queries," in *Proc. INTERSPEECH*, 2025, pp. 2660–2664.
- [11] H. Wang, J. Mao, Z. Guo, J. Wan, H. Liu, and X. Wang, "Leveraging language model capabilities for sound event detection," *Proc. INTERSPEECH*, 2024.
- [12] Y. Xin, X. Peng, and Y. Lu, "Masked Audio Modeling with CLAP and Multi-Objective Learning," in *Proc. INTERSPEECH*, 2023, pp. 2763–2767.
- [13] X. Jing, A. Triantafyllopoulos, and B. Schuller, "ParaCLAP – Towards a general language-audio model for computational paralinguistic tasks," in *Proc. INTERSPEECH*, 2024, pp. 1155–1159.
- [14] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *Intl. Journal of computer vision*, vol. 130, no. 9, pp. 2337–2348, 2022.
- [15] —, "Conditional prompt learning for vision-language models," in *Proc. CVPR*, 2022, pp. 16 816–16 825.
- [16] J. Shi, Y. Zhang, and Y. Gao, "CLEP-DG: Contrastive Learning for Speech Emotion Domain Generalization via Soft Prompt Tuning," in *Proc. INTERSPEECH*, 2025, pp. 4498–4502.
- [17] Y. Li, X. Wang, and H. Liu, "Audio-free prompt tuning for language-audio models," in *Proc. ICASSP*. IEEE, 2024, pp. 491–495.
- [18] A. Hanif, M. T. Agro, M. A. Qazi, and H. Aldarmaki, "Palm: Few-shot prompt learning for audio language models," in *Proc. EMNLP*, 2024, pp. 18 527–18 536.
- [19] M. Tian, A. Srinivasamurthy, M. Sandler, and X. Serra, "A study of instrument-wise onset detection in beijing opera percussion ensembles," in *Proc. ICASSP*. IEEE, 2014, pp. 2159–2163.
- [20] J. Engel, C. Resnick, A. Roberts, S. Dieleman, M. Norouzi, D. Eck, and K. Simonyan, "Neural audio synthesis of musical notes with wavenet autoencoders," in *Proc. ICML*. PMLR, 2017, pp. 1068–1077.
- [21] K. J. Piczak, "Esc: Dataset for environmental sound classification," in *Proc. ACM MM*, 2015, pp. 1015–1018.
- [22] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *Proc. ACM MM*, 2014, pp. 1041–1044.
- [23] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "Crema-d: Crowd-sourced emotional multimodal actors dataset," *IEEE Trans. on Affective Computing*, vol. 5, no. 4, pp. 377–390, 2014.
- [24] S. R. Livingstone and F. A. Russo, "The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english," *PLoS one*, vol. 13, no. 5, p. e0196391, 2018.
- [25] Y. Gong, Y.-A. Chung, and J. Glass, "Psla: Improving audio tagging with pretraining, sampling, labeling, and aggregation," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 29, pp. 3292–3306, 2021.
- [26] T. Spadini, "Sound events for surveillance applications," Nov. 2019. [Online]. Available: <https://doi.org/10.5281/zenodo.3519845>
- [27] A. Mesaros, T. Heittola, and T. Virtanen, "Tut acoustic scenes 2017, development dataset," *Proc. DCASE workshop*, 2017.
- [28] B. L. Sturm, "An analysis of the gtzan music genre dataset," in *Proc. ACM MIRUM*, 2012, pp. 7–12.
- [29] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *Proc. ICML*. PMLR, 2021, pp. 8748–8763.
- [30] S. Deshmukh, B. Elizalde, R. Singh, and H. Wang, "Pengi: An audio language model for audio tasks," *NeurIPS*, vol. 36, pp. 18 090–18 108, 2023.