ZEBRA: Zero-Shot Entropy-Regularized Prompt Learning for Base-to-Novel Generalization in Audio-Language Models

Asif Hanif †, *Mohammad Yaqub*

Mohamed Bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE

asif.hanif@mbzuai.ac.ae, mohammad.yaqub@mbzuai.ac.ae

Abstract

Audio-Language Models (ALMs) achieve strong zero-shot performance by aligning audio with textual class descriptions. Although prompt learning improves accuracy on base classes through few-shot supervised adaptation, we observe a critical trade-off: it often degrades performance on novel classes, sometimes falling below zero-shot accuracy. This exposes a base-to-novel generalization gap in prompt learning for ALMs. To address this issue, we propose **ZEBRA** (Zero-shot Entropy-Regularized Prompt Learning for Base-to-Novel Generalization), a plug-and-play framework that fuses zero-shot logits with prompt-learning logits, and employs self-entropy regularization to reduce overfitting to base classes. Experiments across multiple audio classification datasets show that ZEBRA consistently improves novel-class performance while maintaining strong base accuracy, significantly reducing the base-to-novel gap compared to standard prompt learning. The code is available at: https://github.com/asif-hanif/zebra.

Index Terms: audio-language models, prompt learning, base-to-novel generalization, audio classification

1. Introduction

Recent advances in Vision-Language Models (VLMs) have inspired the development of Audio-Language Models (ALMs), which achieve strong performance on zero-shot audio recognition tasks [1, 2, 3, 4]. In the zero-shot setting, audio features are aligned with textual descriptions of class labels, enabling recognition without task-specific training. This paradigm provides flexibility and generalization to unseen categories while eliminating the need for large labeled datasets. Despite these advantages, zero-shot performance often falls short of fully supervised or few-shot adaptation. The primary limitation lies in the reliance on manually designed text prompts, which may not optimally align with the target dataset. Prompt learning addresses this limitation by replacing hand-crafted templates with learnable context tokens that are optimized using labeled data from base classes [5, 6]. By adapting prompts to the downstream task, prompt learning significantly improves performance over zero-shot baselines.

However, in the context of audio-language models, we observe a critical trade-off: while prompt learning consistently improves accuracy on base classes (seen during few-shot training), it often degrades performance on novel classes (unseen during few-shot training). In several cases, prompt learning performs even worse than the original zero-shot model on novel categories (refer to Figure 1 and Table 1). This

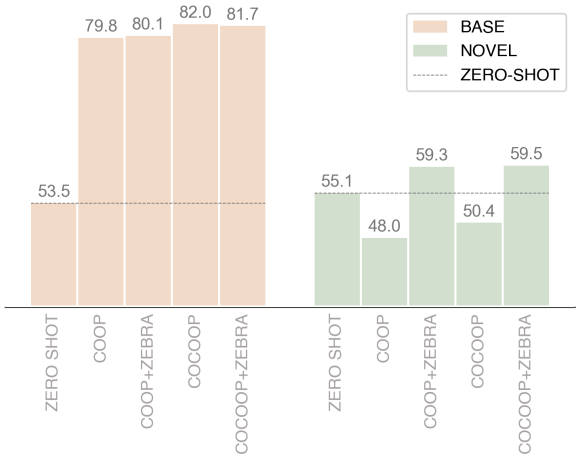

Figure 1: ***Comparison of Base and Novel Performance.*** *Existing prompt-learning methods improve accuracy on base classes but generalize poorly to novel classes, often performing even worse than zero-shot inference. In contrast, incorporating ZEBRA with these baselines consistently boosts novel-class accuracy while maintaining strong performance on base classes.*

reveals a fundamental base-to-novel generalization gap under prompt learning. Optimizing prompts solely using base-class supervision encourages overfitting to seen categories, distorting the semantic alignment learned during large-scale pre-training and weakening zero-shot transferability.

To address this challenge, we propose **ZEBRA**: **Z**ero-shot **E**ntropy-Regularized Prompt Learning for **B**ase-to-Novel Gene**RA**lization. ZEBRA is a plug-and-play framework that operates on top of existing prompt learning methods and is designed to preserve zero-shot generalization while still benefiting from supervised adaptation. ZEBRA introduces two complementary mechanisms: First, it fuses zero-shot logits with the prompt-learning logits, effectively anchoring adaptation to the original zero-shot decision space. Second, it introduces a self-entropy regularization term in the loss function to mitigate over-confidence in base classes during few-shot training. By discouraging over-confident predictions, self-entropy prevents excessive specialization to seen categories and promotes smoother decision boundaries, thereby improving generalization to unseen classes. ZEBRA (refer to Figure 2 for an overview) is lightweight, as it introduces no additional learnable parameters to existing prompt-learning baselines. The zero-shot logits are computed once and reused during both few-shot training and

† indicates the corresponding author.

inference, eliminating extra forward passes through the text encoder and resulting in negligible computational overhead. Our *contributions* are as follows:

- We identify and analyze a base-to-novel generalization gap in prompt learning for audio-language models, where improvements on base classes often come at the cost of degraded novel-class performance (see Figure 1 or Table 1).
- We propose ZEBRA (Zero-shot Entropy-Regularized Prompt Learning for Base-to-Novel Generalization), a plug-and-play framework that builds upon existing prompt-learning methods without introducing additional learnable parameters and with negligible computational overhead.
- ZEBRA integrates zero-shot logit fusion to anchor adaptation to the pre-trained decision space and employs self-entropy regularization to reduce overconfidence on base classes, enhancing generalization to unseen categories.
- We demonstrate consistent improvements in novel-class generalization across multiple audio classification datasets, significantly reducing the base-to-novel performance gap compared to vanilla baseline prompt learning methods.

2. Related Work

Inspired by the success of CLIP [7] in the image–language domain, many audio–language models have adopted a similar contrastive learning framework. CLAP [2] and AudioCLIP [3], for example, extend this paradigm to align audio and textual representations, enabling robust audio classification and cross-modal retrieval. Like CLIP, these models are trained to maximize similarity between matched audio–text pairs while minimizing similarity for mismatched pairs.

Few-shot adaptation has attracted significant attention in vision–language models, particularly for adapting CLIP [8]. Among these, prompt learning has emerged as a parameter-efficient alternative to full fine-tuning, enabling adaptation to downstream tasks by introducing a small set of learnable parameters while keeping the pretrained backbone frozen [5, 6]. This strategy substantially reduces computational and storage costs compared to full model updates. Prompt learning has been successfully applied across natural language processing [9], computer vision [8, 10, 11], and more recently, the audio-language domain [12, 13, 14]. In audio–language models, approaches such as PALM [12] demonstrate that learning only a small set of prompt vectors can effectively adapt frozen pretrained ALM to achieve strong audio classification performance.

3. Methodology

Zero-Shot Classification in ALM. In CLIP-style audio-language models (ALMs) [7], zero-shot classification is performed by measuring the similarity between the audio representation and a set of class-specific text descriptions. Let $\mathbf{x}$ denote an input audio waveform, and let $\mathbf{t} = \{t_1, t_2, \ldots, t_c\}$ represent the set of textual class descriptions for c classes. The prediction logits are computed as:

$$f_{\text{zs}}(\mathbf{x}) = \left\{ \text{sim}\left(f_A(\mathbf{x}),\ f_T(t_i)\right) \right\}_{i=1}^{c}, \tag{1}$$

where f_A and f_T denote the audio and text encoders, respectively, and $\text{sim}(\cdot)$ is a cosine similarity function. The predicted label corresponds to the class with the highest similarity score.

Base-to-Novel Generalization. In the base-to-novel setting, the full label space is defined as $\mathcal{C} = \{1, 2, \ldots, c\}$ and is partitioned into base classes $\mathcal{C}_B$ and novel classes $\mathcal{C}_N$, such that $\mathcal{C}_B \cup \mathcal{C}_N = \mathcal{C}$ and $\mathcal{C}_B \cap \mathcal{C}_N = \emptyset$. The model is trained using labeled data from $\mathcal{C}_B$ and is expected to generalize effectively to the unseen (novel) classes in $\mathcal{C}_N$. In the context of prompt learning, the prompts are optimized using base-class samples from the training set, and evaluation is conducted on both base and novel class samples from the test set to assess generalization performance.

Prompt Learning in ALM. Textual class descriptions play a central role in zero-shot inference for ALMs; however, manually designed prompts often introduce performance variability and sensitivity to wording. Prompt learning mitigates this issue by introducing auxiliary learnable parameters φ into the text encoder f_T. These parameters are optimized in a few-shot setting to adapt the frozen pre-trained model to downstream tasks while reducing reliance on manual prompt engineering [9]. Formally, prompt learning learns φ by minimizing a task-specific objective over a few-shot training dataset $\mathcal{D}$:

$$\underset{\varphi}{\text{minimize}} \sum_{(\mathbf{x},y)\in\mathcal{D}} \mathcal{L}\big(f_{\text{pr}}(\mathbf{x};\varphi), y\big), \tag{2}$$

where $(\mathbf{x}, y)$ denotes an audio-label pair from $\mathcal{D}$, $f_{\text{pr}}(\mathbf{x};\varphi)$ represents the model's prediction conditioned on the learnable prompt parameters φ, and $\mathcal{L}(\cdot)$ is the task-specific loss function, typically the cross-entropy loss. Two representative instances of $f_{\text{pr}}(\mathbf{x};\varphi)$ are COOP [5] and COCOOP [6]. In COOP, a set of shared learnable context tokens is optimized and prepended to each class name before being passed to text encoder. In COCOOP, the context tokens are dynamically generated conditioned on the audio feature vector, enabling instance-specific adaptation. Although these approaches improve performance on base classes, they optimize prompts solely using base-class supervision and do not explicitly preserve the original zero-shot alignment. As a result, the adapted representation space can become biased toward seen categories, leading to degraded performance on novel classes, as shown in Figure 1.

ZEBRA: Zero-Shot Entropy-Regularized Base-to-Novel Generalization. To mitigate the base-to-novel generalization gap, we propose ZEBRA, a plug-and-play framework that operates on top of existing prompt learning methods. ZEBRA preserves zero-shot transferability while benefiting from supervised adaptation through two complementary mechanisms.

(i) Zero-shot Logit Fusion: Instead of relying solely on prompt-learning logits, we fuse them with the original zero-shot logits during both few-shot training and inference:

$$f_{\text{zebra}}(\mathbf{x};\varphi) = \lambda_{\text{zs}} \cdot f_{\text{zs}}(\mathbf{x}) + \lambda_{\text{pr}} \cdot f_{\text{pr}}(\mathbf{x};\varphi), \tag{3}$$

where $f_{\text{zs}} \in \mathbb{R}^c$ denotes logits from zero-shot setting, $f_{\text{pr}} \in \mathbb{R}^c$ denotes logits from any prompt-learning method (e.g., COOP or COCOOP) and $\lambda_{\text{zs}}, \lambda_{\text{pr}}$ control the contribution of the zero-shot and prompt learning components, respectively. This fusion anchors adaptation to the pre-trained decision space.

(ii) Self-Entropy Regularization: To prevent overconfidence on base classes, we introduce a self-entropy regularization term:

$$\mathcal{L}_{\text{ent}} = -\sum_i \mathbf{p}^i(\mathbf{x}) \log \mathbf{p}^i(\mathbf{x}), \tag{4}$$

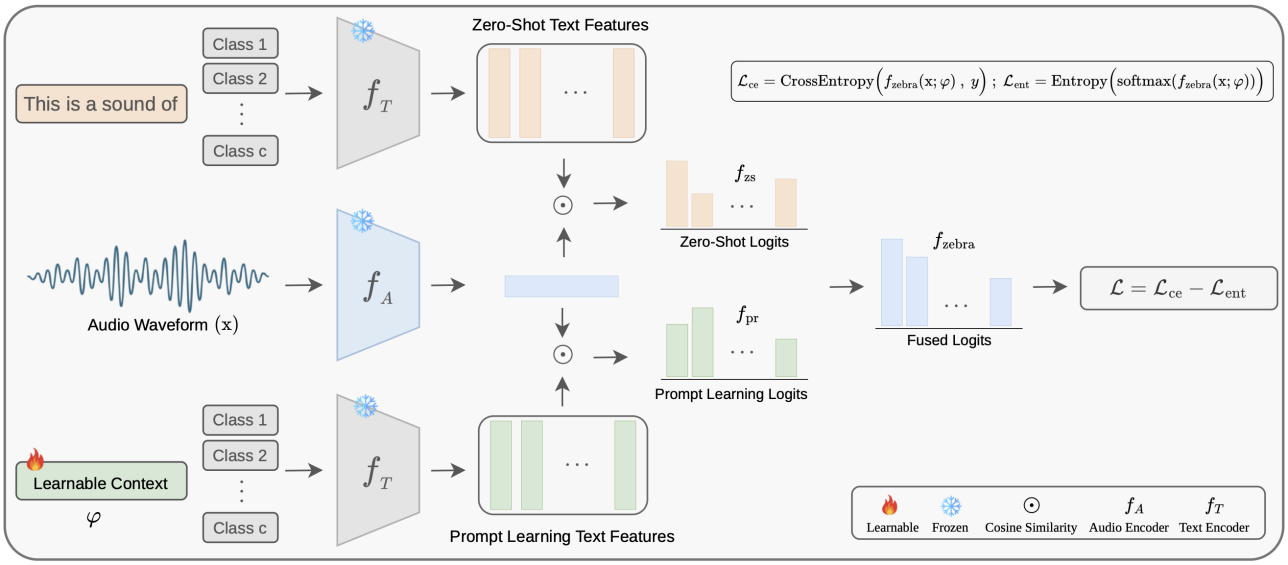

Figure 2: ZEBRA *approach operates on top of existing prompt learning methods to bridge the base-to-novel generalization gap, preserving zero-shot transferability while benefiting from supervised adaptation through few-shot prompt learning. ZEBRA introduces no additional learnable parameters to existing prompt learning methods and incurs negligible computational overhead.*

where $\mathbf{p}(\mathbf{x})$ is the softmax probability derived from $f_{\text{zebra}}(\mathbf{x};\varphi) \in \mathbb{R}^c$. The final training objective becomes:

$$\mathcal{L}_{\text{zebra}}\big(\mathbf{x}, y; \varphi\big) = \mathcal{L}_{\text{ce}}\big(f_{\text{zebra}}, y\big) - \mathcal{L}_{\text{ent}}(\text{softmax}(f_{\text{zebra}})),$$

where $\mathcal{L}_{\text{ce}}(\cdot)$ is the cross-entropy loss. By combining zero-shot logit fusion with entropy regularization, ZEBRA reduces overfitting to base classes while preserving the semantic alignment necessary for robust novel-class generalization. Formally, we optimize following objective during few-shot training on base classes:

$$\underset{\varphi}{\text{minimize}} \sum_{(\mathbf{x},y)\in\mathcal{D}} \mathcal{L}_{\text{zebra}}\big(\mathbf{x}, y; \varphi\big). \tag{5}$$

During optimization, we minimize the cross-entropy loss while maximizing the self-entropy term, which constructively counteracts overconfident predictions, thereby discouraging overconfidence and reducing overfitting to the base classes. At inference time, predictions are computed from the fused zero-shot and prompt-learning logits, without applying entropy regularization. The predicted label $\hat{y}$ is obtained as follows:

$$\hat{y} = \underset{i\in\{1,2,\ldots,c\}}{\text{argmax}} \quad f^i_{\text{zebra}}(\mathbf{x};\varphi), \tag{6}$$

where $f_{\text{zebra}}(\mathbf{x};\varphi) \in \mathbb{R}^c$ denotes the fused logit vector over all c classes, and $f^i_{\text{zebra}}(\mathbf{x};\varphi)$ represents the logit corresponding to i_{th} class. It should be noted that few-shot training is performed using classes in $\mathcal{C}_B$, while evaluation is conducted separately on base classes $\mathcal{C}_B$ and novel classes $\mathcal{C}_N$.

It is worth noting that ZEBRA introduces no additional learnable parameters to existing prompt-learning baselines. The zero-shot logits are computed only once and can be reused during both few-shot training and inference, eliminating the need for additional forward passes through the text encoder. As a result, ZEBRA is a lightweight approach that incurs negligible computational overhead. For an overview of the proposed approach, see Figure 2.

4. Experiments and Results

Models and Datasets. For the CLIP-style audio-language backbone, we adopt Pengi [1], a generative audio-language model consisting of audio and text encoders followed by an LLM decoder. Following the setup of PALM [12], we discard the decoder and utilize only the pretrained audio and text encoders, effectively employing PENGI in a CLIP-like contrastive framework to leverage its strong pretrained representations and zero-shot generalization capability. We evaluate our method on a diverse set of speech and audio processing tasks, including instrument classification, sound event classification, emotion recognition, vocal sound classification, surveillance sound event classification, acoustic scene classification, and music analysis. Instrument classification is assessed using the Beijing Opera [15] and NS-Instruments [16] datasets. For sound event classification, we use ESC-50 and its ESC50-Actions subset [17], as well as UrbanSound8K [18]. Emotion recognition is evaluated on CREMA-D [19] and RAVDESS [20]. Vocal sound classification is conducted using VocalSound [21], while surveillance sound event classification relies on SESA [22]. Acoustic scene classification is performed on TUT2017 [23], and music analysis is evaluated using GT-Music-Genre [24].

Baseline Methods. We consider ZERO-SHOT, COOP [5], and COCOOP [6] as baselines. COOP and COCOOP are prompt learning methods originally proposed for vision-language models that replace handcrafted prompts with learnable context tokens optimized in the text encoder's input embedding space; COCOOP further introduces instance-conditioned prompts via feedback from the audio encoder, whereas COOP learns context tokens shared across instances. We adopt their audio-language adaptations as implemented in PALM [12], where the vision encoder is replaced with an audio encoder. We exclude PALM itself as a baseline because its class-specific learnable vectors limit base-to-novel generalization, while COOP and COCOOP are class-agnostic and can generalize to unseen classes. Our method builds upon COOP and COCOOP in a plug-and-play

Table 1: ***Comparison of ZEBRA with Baseline Methods.*** *While baseline prompt learning methods improve performance on base classes, they often lead to degraded accuracy on novel classes when compared with zero-shot. By incorporating ZEBRA, performance on novel classes is consistently improved while maintaining strong base-class accuracy. Values marked with ▲/▼ denote the increase/decrease in accuracy with respect to the zero-shot performance for the corresponding dataset in each row.*

METHODS →	ZERO SHOT		COOP		COCOOP		COOP + ZEBRA		COCOOP + ZEBRA	
DATASETS ↓	BASE	NOVEL	BASE	NOVEL	BASE	NOVEL	BASE	NOVEL	BASE	NOVEL
Beijing-Opera	52.00	43.48	96.10 (▲44.1)	60.87 (▲17.3)	96.00 (▲44.0)	60.88 (▲17.3)	96.03 (▲44.0)	82.61 (▲39.1)	96.20 (▲44.2)	78.26 (▲34.7)
CREMA-D	66.13	25.99	59.16 (▼6.97)	32.61 (▲6.62)	63.11 (▼3.02)	14.84 (▼11.1)	61.72 (▼4.41)	18.24 (▼7.75)	54.29 (▼11.8)	19.94 (▼6.05)
ESC50-Actions	67.50	75.00	100.0 (▲32.5)	72.50 (▼2.50)	100.0 (▲32.5)	62.50 (▼12.5)	95.00 (▲27.5)	77.50 (▲2.50)	97.50 (▲30.0)	77.50 (▲2.50)
ESC50	58.50	67.00	94.50 (▲36.0)	54.50 (▼12.5)	95.00 (▲36.5)	63.00 (▼4.00)	95.50 (▲37.0)	65.00 (▼2.00)	94.50 (▲36.0)	61.50 (▼5.50)
GT-Music-Genre	56.86	36.73	76.47 (▲19.6)	53.06 (▲16.3)	83.33 (▲26.4)	45.92 (▲9.19)	74.51 (▲17.6)	52.04 (▲15.3)	83.33 (▲26.4)	37.76 (▲1.03)
NS-Instruments	53.61	53.87	65.79 (▲12.1)	39.22 (▼14.6)	66.67 (▲13.0)	68.57 (▲14.7)	70.78 (▲17.1)	54.93 (▲1.06)	68.34 (▲14.7)	63.73 (▲9.86)
RAVDESS	23.25	38.78	59.21 (▲35.9)	32.70 (▼6.08)	60.53 (▲37.2)	40.68 (▲1.90)	59.65 (▲36.4)	42.97 (▲4.19)	60.09 (▲36.8)	43.73 (▲4.95)
SESA	60.00	93.33	95.56 (▲35.5)	76.67 (▼16.6)	91.11 (▲31.1)	93.33 (▲0.00)	91.11 (▲31.1)	98.33 (▲5.00)	93.33 (▲33.3)	95.00 (▲1.67)
TUT2017	33.33	30.72	67.81 (▲34.4)	15.86 (▼14.8)	80.14 (▲46.8)	18.88 (▼11.8)	71.69 (▲38.3)	31.53 (▲0.81)	80.37 (▲47.0)	37.35 (▲6.63)
UrbanSound8K	63.77	67.55	88.11 (▲24.3)	36.12 (▼31.4)	88.80 (▲25.0)	47.48 (▼20.0)	85.26 (▲21.4)	65.94 (▼1.61)	87.89 (▲24.1)	62.96 (▼4.59)
VocalSound	53.90	74.54	75.50 (▲21.6)	54.43 (▼20.1)	77.95 (▲24.0)	38.77 (▼35.7)	80.73 (▲26.8)	64.07 (▼10.4)	83.63 (▲29.7)	76.77 (▲2.23)
AVERAGE	53.53	55.18	79.82 (▲26.2)	48.04 (▼7.13)	82.05 (▲28.5)	50.44 (▼4.74)	80.17 (▲26.6)	59.37 (▲4.19)	81.75 (▲28.2)	59.50 (▲4.31)

Table 2: *Ablation on zero-shot logits and entropy loss.*

ZERO−SHOT	ENTROPY	BASE	NOVEL
✗	✗	79.82	48.04
✓	✗	80.31	58.83
✗	✓	81.21	48.10
✓	✓	80.17	59.37

Table 3: *Runtime and ECE of baselines with and without* ZEBRA. (†) *indicates results obtained using* ZEBRA.

METHODS	COOP	COOP†	COCOOP	COCOOP†
TIME (Train)	19m 36s	19m 45s	45m 37s	45m 41s
TIME (Test)	2m 40s	2m 42s	3m 37s	3m 38s
ECE (BASE)	0.0677	0.0755	0.0821	0.0777
ECE (NOVEL)	0.2033	0.1997	0.2738	0.2253

manner, seamlessly integrating with their learned prompt representations without introducing additional trainable parameters or requiring further hyperparameter tuning. It operates on top of the existing framework, preserving their training protocols and generalization capabilities, while providing consistent improvements with minimal computational overhead and no modifications to the underlying model architecture.

Implementation Details. Except for ZERO-SHOT, all methods are trained for 50 epochs under the standard few-shot setting, using 16 randomly sampled training examples per *base* class. Inference is conducted on the full test set using the predefined *base–novel* class split. Few-shot training is done with stochastic gradient descent (SGD) at a learning rate of 0.05, and performance is measured in terms of accuracy. Each method is evaluated with three different random seeds, and we report the average results. For ZERO-SHOT, we adopt the default prompt template: `This is a recording of {CLASS NAME}`, for a fair comparison. We run all experiments using `NVIDIA RTX A6000` GPU. For the contribution weights in Equation 3, we empirically set $\lambda_{\text{zs}} = 0.5$ and $\lambda_{\text{pr}} = 0.5$. We also scale the $\mathcal{L}_{\text{ent}}$ by a factor of 0.05 during few-shot training.

Results and Discussion. We report a comparison of ZERO-SHOT, prompt-learning methods, and their ZEBRA-enhanced variants in Table 1. As discussed earlier, baseline methods (COOP and COCOOP) substantially improve performance on base classes compared to ZERO-SHOT (e.g., $+26.2\%$ and $+28.5\%$, respectively). However, they generalize poorly to novel classes, exhibiting notable drops in accuracy (e.g., -7.13% and -4.74%, respectively) and, in both cases, performing below the ZERO-SHOT novel-class average of 55.18%. On the other hand, ZEBRA enhances their novel-class performance by $+4.19\%$ and $+4.31\%$, respectively, bringing the accuracy above the ZERO-SHOT novel-class average of 55.18% while maintaining competitive performance on base classes. Results demonstrate that, on average, ZEBRA improves novel-class generalization while maintaining base performance, achieving these gains with no additional parameters and negligible computational overhead (see Table 3).

Ablative Analysis. We analyze the impact of the fusion of zero-shot logits and self-entropy loss term on ZEBRA's performance in Table 2 (using the COOP baseline). The results indicate that most of the improvement comes from incorporating zero-shot logits, while the self-entropy term provides a further, albeit marginal, gain. We also report runtime (training and testing, averaged across all datasets) and Expected Calibration Error (ECE) in Table 3. ZEBRA introduces negligible computational overhead while consistently reducing ECE on average across both base and novel classes, demonstrating improved calibration without sacrificing efficiency.

5. Conclusion

We introduced ZEBRA, a lightweight, plug-and-play framework for improving base-to-novel generalization in audio–language models. While existing prompt-learning methods substantially boost base-class performance, they often suffer from degraded generalization to novel classes, frequently underperforming the zero-shot baseline. ZEBRA effectively mitigates this trade-off by leveraging zero-shot knowledge and entropy-regularized logits, consistently improving novel-class accuracy while preserving strong base-class performance. ZEBRA adds no learnable parameters and incurs negligible computational overhead, making it a simple yet effective enhancement to existing prompt-learning methods.

6. Generative AI Use Disclosure

We confirm that an LLM was used solely for writing refinement (grammar, wording, and clarity). All ideas, analyses, and conclusions are the authors' own.

7. References

[1] S. Deshmukh, B. Elizalde, R. Singh, and H. Wang, "Pengi: An audio language model for audio tasks," *Advances in Neural Information Processing Systems*, vol. 36, pp. 18 090–18 108, 2023.

[2] B. Elizalde, S. Deshmukh, M. Al Ismail, and H. Wang, "Clap learning audio concepts from natural language supervision," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[3] A. Guzhov, F. Raue, J. Hees, and A. Dengel, "Audioclip: Extending clip to image, text and audio," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 976–980.

[4] J. Zhang, J. Huang, S. Jin, and S. Lu, "Vision-language models for vision tasks: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[5] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *International Journal of Computer Vision (IJCV)*, 2022.

[6] K. Zhou., J. Yang, C. C. Loy, and Z. Liu, "Conditional prompt learning for vision-language models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16 816–16 825.

[7] 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 *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[8] J. Gu, Z. Han, S. Chen, A. Beirami, B. He, G. Zhang, R. Liao, Y. Qin, V. Tresp, and P. Torr, "A systematic survey of prompt engineering on vision-language foundation models," *arXiv preprint arXiv:2307.12980*, 2023.

[9] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," *ACM computing surveys*, vol. 55, no. 9, pp. 1–35, 2023.

[10] A. Hanif, F. Shamshad, M. Awais, M. Naseer, F. S. Khan, K. Nandakumar, S. Khan, and R. M. Anwer, "Baple: Backdoor attacks on medical foundational models using prompt learning," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2024, pp. 443–453.

[11] R. Imam, A. Hanif, J. Zhang, K. W. Dawoud, Y. Kementchedjhieva, and M. Yaqub, "Noise is an efficient learner for zero-shot vision-language models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 5820–5829.

[12] A. Hanif, M. T. Agro, M. A. Qazi, and H. Aldarmaki, "Palm: Few-shot prompt learning for audio language models," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024, pp. 18 527–18 536.

[13] A. Seth, R. Selvakumar, S. Kumar, S. Ghosh, and D. Manocha, "Pat: Parameter-free audio-text aligner to boost zero-shot audio classification," in *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 2025, pp. 12 376–12 394.

[14] A. Hanif, M. T. Agro, F. Shamshad, and K. Nandakumar, "Trojanwave: Exploiting prompt learning for stealthy backdoor attacks on large audio-language models," in *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2025, pp. 18 628–18 644.

[15] M. Tian, A. Srinivasamurthy, M. Sandler, and X. Serra, "A study of instrument-wise onset detection in beijing opera percussion ensembles," in *2014 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, 2014, pp. 2159–2163.

[16] J. Engel, C. Resnick, A. Roberts, S. Dieleman, D. Eck, K. Simonyan, and M. Norouzi, "Neural audio synthesis of musical notes with wavenet autoencoders," 2017.

[17] K. J. Piczak, "ESC: Dataset for Environmental Sound Classification," in *Proceedings of the 23rd Annual ACM Conference on Multimedia*. ACM Press, pp. 1015–1018. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2733373.2806390

[18] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *Proceedings of the 22nd ACM international conference on Multimedia*, 2014, pp. 1041–1044.

[19] 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 transactions on affective computing*, vol. 5, no. 4, pp. 377–390, 2014.

[20] 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.

[21] Y. Gong, Y.-A. Chung, and J. Glass, "Psla: Improving audio tagging with pretraining, sampling, labeling, and aggregation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2021.

[22] T. Spadini, "Sound events for surveillance applications," 2019.

[23] T. Heittola, A. Mesaros, and T. Virtanen, "TUT Acoustic Scenes 2017, Development dataset," Department of Signal Processing, Tampere University of Technology, Tech. Rep., 2017. [Online]. Available: https://www.cs.tut.fi/sgn/arg/dcase2017/challenge/task-acoustic-scene-classification

[24] B. L. Sturm, "An analysis of the gtzan music genre dataset," in *Proceedings of the second international ACM workshop on Music information retrieval with user-centered and multimodal strategies*, 2012, pp. 7–12.