AGAPE: Training-Free Audio-Guided Adaptive Prompt Ensembling for Few-Shot Audio Classification

Kai Guo*, Xiang Xie*†⋆, Shangkai Zhao*, Shu Wu*

*Beijing Institute of Technology, Beijing, China
†Beijing Institute of Technology, Zhuhai, Guangdong, China
{guokai, xiexiang, zhaoshangkai, wushu}@bit.edu.cn

***Abstract*—Although large-scale pre-trained models have revolutionized audio classification, adapting them to downstream few-shot tasks remains challenging. Existing methods face a dilemma: conventional few-shot adaptation requires resource-intensive gradient updates, while zero-shot inference often suffers from semantic misalignment in text prompts. To address these issues in deployment-constrained scenarios, we propose Audio-Guided Adaptive Prompt Ensembling (AGAPE), a training-free framework for few-shot audio classification. Specifically, AGAPE first utilizes limited support shots to construct optimal class-specific text prototypes, thereby enhancing text-prompt-based classification. Subsequently, the final classification is achieved by fusing the similarity scores of the test sample against both these optimized text prototypes and the audio prototypes. Furthermore, we introduce the Covariance Matrix Adaptation Evolution Strategy (CMA-ES) for Test-Time Adaptation (TTA), enabling efficient model adaptation without gradient updates. Experiments on ESC-50 and FSD50K demonstrate that AGAPE significantly outperforms the baseline. Ablation studies further validate the effectiveness of our methods.**

***Index Terms*—audio classification, few-shot learning, prompt ensembling, test-time adaptation**

I. Introduction

The objective of audio classification is to categorize diverse sound classes. It serves as the fundamental basis for specific tasks such as sound event detection [1], [2], acoustic scene classification [3], [4], and anomalous sound detection [5]. The most widely adopted paradigms typically involve either training neural networks from scratch or leveraging foundation models pre-trained on large-scale audio data, followed by fine-tuning for specific downstream tasks [6]–[8]. However, fine-tuning is not always feasible for audio classification tasks due to constraints on the type, quality, and quantity of audio data in real-world scenarios [9]–[12].

CLIP [13] stands as a milestone in the field of computer vision. By pre-training on large-scale datasets using contrastive learning, it aligns visual and textual modalities, achieving state-of-the-art performance in zero-shot visual classification. Extending the CLIP paradigm to the audio domain, CLAP [14] aligns audio and text representations in a shared space, which has become a standard approach for zero-shot audio classification. However, standard CLAP models typically rely on a single generic prompt (e.g., this is a sound of `<label>`) to handle all detection tasks. While this approach suffices for simple zero-shot scenarios, it is often not the optimal choice for specific tasks or distinct sound categories.

Recently, several studies have focused on prompt optimization. For instance, TSPE [15] designs customized prompt templates tailored for specific tasks, such as acoustic scene classification and instrument classification. Similarly, PAT [16] constructs a prompt data store and leverages a cross-modal interaction approach to enhance performance. A primary limitation of these methods is their exclusive focus on purely zero-shot scenarios, making them difficult to extend to few-shot settings. Another line of research aims to improve performance by introducing learnable prompts or prompt embeddings, as demonstrated by methods like COOP [17] and PALM [18]; however, the requirement for training remains a significant limitation of these approaches.

Alternatively, Taylor et al. [19] introduced a straightforward strategy that substitutes text embeddings with the average embedding of a few audio samples, thereby shifting the optimization focus of CLAP to the audio modality. According to their findings, constructing class prototypes from audio samples yields superior robustness compared to text-based prototypes, primarily due to the noise inherent in text prompts. However, a limitation of their study is the reliance on a relatively large number of samples (e.g., 20 per class) for prototype construction. In contrast, other studies focusing on few-shot learning typically utilize fewer samples (e.g., 5 per class) to better align with real-world scenarios [12], [20].

On the other hand, in real-world audio classification scenarios, models are frequently deployed on edge devices (e.g., smartphones) to ensure real-time recognition. Under such conditions, constrained by the device's software architecture and computational resources, the parameters of the deployed models often remain frozen. Consequently, Test-Time Adaptation (TTA) techniques, particularly backpropagation-free (BP-free) approaches, have become critically important. Previously, BP-free TTA approaches were primarily applied in the field of computer vision [21]. Recently, E-BATS [22] successfully extended these techniques to the speech domain, utilizing CMA-ES [23] to enable test-time adaptation for speech models

⋆Corresponding author.

such as HuBERT [24]. However, to date, no existing work has applied similar techniques to the field of CLAP-based audio classification.

Our perspective aligns with Taylor et al. [19], as we aim to optimize CLAP-based audio classification using limited support samples without conducting any additional model training. Inspired by the recent success of CMA-ES-based test-time optimization in speech foundation models [22], we hypothesize that this approach can be transferred to CLAP models to realize optimization without backpropagation. Specifically, we leverage a minimal support set (3 shots) to build audio prototypes while filtering the optimal prompts for each category to construct text prototypes. By jointly leveraging these multi-modal prototypes, we perform audio classification and employ CMA-ES [23] to further boost performance. The main contributions of this paper are summarized as follows:

- We introduce Audio-Guided Adaptive Prompt Ensembling (AGAPE), a training-free framework designed for few-shot audio classification. A comprehensive prompt library is designed to facilitate prompt ensembling.
- We introduce the CMA-ES technique to achieve test-time optimization for fine-tuning-free few-shot audio classification based on CLAP.
- Comprehensive ablation studies are performed to validate the effectiveness of different prompts and analyze the roles of audio and text modalities in CLAP.

II. Methods

A. Audio-Guided Adaptive Prompt Ensembling Framework

The proposed AGAPE framework is illustrated in Fig. 1. For a few-shot audio classification task, let $\mathcal{S}_c = \{x_1, x_2, \ldots, x_N\}$ denote the support set containing N audio samples for a specific class c. Similar to the approach by Taylor et al. [19], we employ the pre-trained CLAP audio encoder, denoted as $E_a(\cdot)$, to extract the audio embeddings $\mathbf{v}_i$ for each sample in the support set:

$$\mathbf{v}_i = E_a(x_i), \quad x_i \in \mathcal{S}_c \tag{1}$$

We then construct the audio prototype $\mathbf{p}_c^{(a)}$ by computing the mean of these support embeddings:

$$\mathbf{p}_c^{(a)} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{v}_i \tag{2}$$

In parallel, we construct a prompt library consisting of L diverse templates: $\mathcal{P} = \{t_1, t_2, \ldots, t_L\}$. We use the CLAP text encoder $E_t(\cdot)$ to extract the text embedding $\mathbf{u}_j = E_t(t_j)$ for each prompt template. To select the most effective prompts for the class c, we calculate the average cosine similarity between each text embedding $\mathbf{u}_j$ and all audio embeddings in the support set:

$$s_j = \frac{1}{N} \sum_{i=1}^{N} \cos(\mathbf{u}_j, \mathbf{v}_i) \tag{3}$$

We then select the top-K prompts with the highest similarity scores s_j. The text prototype $\mathbf{p}_c^{(t)}$ is obtained by averaging the embeddings of these selected prompts:

$$\mathbf{p}_c^{(t)} = \frac{1}{K} \sum_{j \in \text{Top-}K} \mathbf{u}_j \tag{4}$$

Audio and text prototypes are cached for efficient access during the inference phase. Notably, as shown in Fig. 1, prompts with the highest similarity provide rich descriptive details about the audio class, offering more information than the category label alone.

Finally, given a query sample x_q with an audio embedding $\mathbf{z} = E_a(x_q)$, we calculate its cosine similarity with both the audio prototype and the text prototype. The final classification score $S(\mathbf{z}, c)$ is derived by fusing these similarities using a weighting parameter β:

$$S(\mathbf{z}, c) = \beta \cdot \cos(\mathbf{z}, \mathbf{p}_c^{(t)}) + (1 - \beta) \cdot \cos(\mathbf{z}, \mathbf{p}_c^{(a)}) \tag{5}$$

Additionally, for purely zero-shot audio classification scenarios where the support set is unavailable, we utilize all templates from the prompt library to ensure the system operates effectively. In this case, the text prototype is constructed by averaging the embeddings of all generated prompts:

$$\mathbf{p}_c^{(t)} = \frac{1}{L} \sum_{j=1}^{L} \mathbf{u}_j \tag{6}$$

and the final classification score relies exclusively on the cosine similarity between the query audio embedding and the text prototype:

$$S(\mathbf{z}, c) = \cos(\mathbf{z}, \mathbf{p}_c^{(t)}) \tag{7}$$

B. Prompt Library

Following related works such as TSPE [15] and PAT [16], we first establish a dedicated prompt library for CLAP. This library is constructed with the assistance of GPT-4o [25], refined through careful manual curation. As illustrated in Table I, we broadly categorize the prompt library into five groups. Specifically, the "Basic Definitions" category consists of generic templates applicable in most scenarios. "Environmental & Spatial" prompts focus on the source location and the acoustic environment. "Acoustic & Physical" covers physical attributes such as the sounding material or pitch characteristics. "Signal Quality" describes audio quality aspects, such as the presence of background noise. The remaining prompts are assigned to the "Others" category. Unlike TSPE [15], our Prompt Library is not meticulously tailored to specific datasets; instead, it prioritizes maximum generality.

C. Test-Time Adaptation based on CMA-ES

Recently, CMA-ES [23] has been successfully applied to test-time optimization of speech foundation models [22]. Building on this, we propose a straightforward test-time adaptation method for CLAP, designed to mitigate potential noise within the AGAPE framework caused by limited support sets or suboptimal text prompt selection.

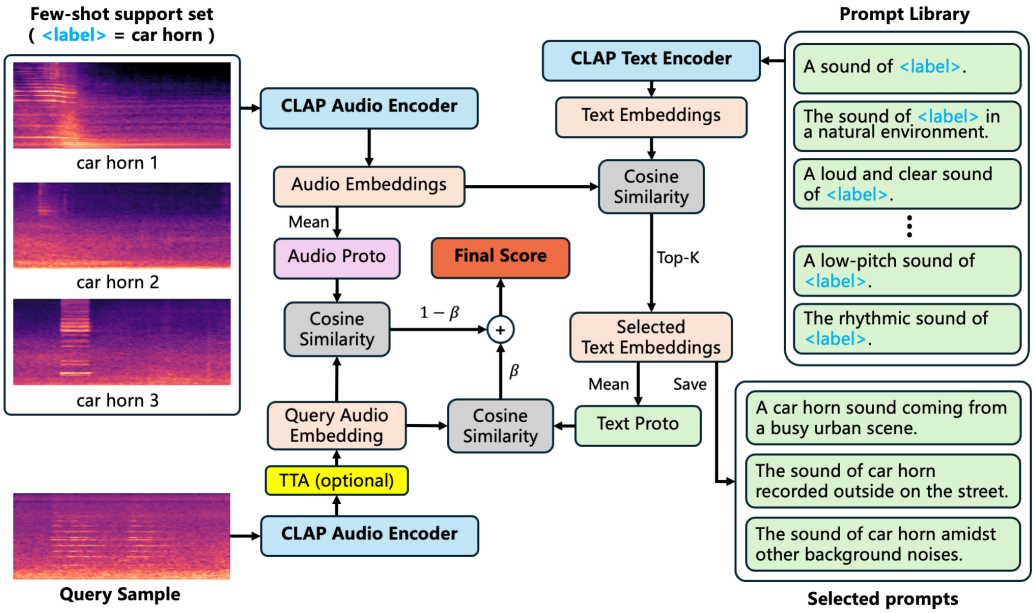

Fig. 1. Overview of Audio-Guided Adaptive Prompt Ensembling framework.

TABLE I
CATEGORIZATION OF THE PROMPT LIBRARY.

Category	Examples
Basic Definitions	"A sound of `<label>`." "An audio recording of `<label>`."
Environmental & Spatial	"The sound of `<label>` in a natural environment." "An echoing sound of `<label>` in a large space."
Acoustic & Physical	"A metallic sound of `<label>`." "A high-pitch sound of `<label>`."
Signal Quality	"A clear recording of `<label>` with minimal background noise." "A noisy recording of `<label>` in a crowded place."
Others	"The sound of `<label>` interacting with water." "A continuous and smooth sound of `<label>` over time."

The pseudocode of our TTA algorithm is presented in Algorithm 1. Specifically, for a query sample with an initial audio embedding $\mathbf{z}_0$, we seek an optimal latent prompt $\boldsymbol{\delta}^*$ to construct an adapted embedding $\hat{\mathbf{z}}$:

$$\hat{\mathbf{z}} = \mathbf{z}_0 + \boldsymbol{\delta}^* \tag{8}$$

Following [22], the optimization objective of CMA-ES is to minimize the entropy of the predicted class distribution, thereby encouraging the model to output high-confidence predictions. According to the original CLAP protocol [14], converting logits (i.e., cosine similarities) into probabilities requires the softmax function for single-label classification and the sigmoid function for multi-label classification.

For single-label classification, the optimal latent prompt $\boldsymbol{\delta}^*$ is defined as:

$$\boldsymbol{\delta}^* = \arg\min_{\boldsymbol{\delta}} \left(-\sum_{c \in \mathcal{C}} P(c|\hat{\mathbf{z}}) \log P(c|\hat{\mathbf{z}}) \right) \tag{9}$$

The probability $P(c|\hat{\mathbf{z}})$ is computed using a temperature-scaled softmax function:

$$P(c|\hat{\mathbf{z}}) = \frac{\exp(\tau \cdot S(\hat{\mathbf{z}}, c))}{\sum_{k \in \mathcal{C}} \exp(\tau \cdot S(\hat{\mathbf{z}}, k))} \tag{10}$$

where $\mathcal{C}$ is the set of all target classes, and τ is the temperature hyperparameter.

For multi-label classification scenarios, we employ the sigmoid function and compute the sum of binary entropies. The optimization objective is formulated as minimizing the total uncertainty across all independent classes:

$$\boldsymbol{\delta}^* = \arg\min_{\boldsymbol{\delta}} \sum_{c \in \mathcal{C}} -\Big(P(c|\hat{\mathbf{z}}) \log P(c|\hat{\mathbf{z}}) + (1 - P(c|\hat{\mathbf{z}})) \log(1 - P(c|\hat{\mathbf{z}})) \Big) \tag{11}$$

Here, each class probability $P(c|\hat{\mathbf{z}})$ is independently estimated using the sigmoid activation function:

$$P(c|\hat{\mathbf{z}}) = \frac{1}{1 + \exp(-\tau \cdot S(\hat{\mathbf{z}}, c))} \tag{12}$$

In each iteration, CMA-ES generates a population of candidates, evaluates their entropy loss, and updates its search distribution. Since a relatively comprehensive set of classes $\mathcal{C}$ is a prerequisite for this optimization, we treat the proposed TTA as an optional refinement step.

TABLE II
COMPARISON WITH STATE-OF-THE-ART METHODS ON ESC-50 AND FSD50K. (%)

Setting	Shot	Method	ESC-50 (Acc)	FSD50K (mAP)
Zero-Shot	0	CLAP [14]	93.70	43.43
		TSPE [15]	94.55	**49.75**
		PAT [16]	94.80	45.52
		Our Prompt Library	**95.40**	48.81
Few-Shot	20	Taylor et al. [19]	97.0	56.1
	3	Taylor et al. [19]	95.79±1.43	51.75±0.91
		Our Text proto only (β = 1)	95.47±1.45	53.82±0.28
		Our Ensemble (β = 0.5)	**96.72±1.33**	**58.29±0.70**
		Our Text proto only (β = 1) with TTA	95.80±1.49	55.11±0.24
		Our Ensemble (β = 0.5) with TTA	**96.80±0.69**	**58.72±0.58**

Algorithm 1 Test-Time Adaptation via CMA-ES

Require: Query audio embedding $\mathbf{z}_0$, CMA-ES state $\boldsymbol{\theta}$, max iterations T, population size λ
Ensure: Predicted label(s) $\hat{y}$
Initialize CMA-ES state $\boldsymbol{\theta}$ and best prompt $\boldsymbol{\delta}^*$
for iteration $t = 1$ to T **do**
  Generate candidates: $\mathcal{D} \leftarrow \{\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_\lambda\} \sim \text{Sample}(\boldsymbol{\theta})$
  Evaluate losses $\mathcal{L}(\mathbf{z}_0 + \boldsymbol{\delta})$ for all $\boldsymbol{\delta} \in \mathcal{D}$
  Update best solution:
    $\boldsymbol{\delta}^* \leftarrow \arg\min_{\boldsymbol{\delta} \in \mathcal{D} \cup \{\boldsymbol{\delta}^*\}} \mathcal{L}(\mathbf{z}_0 + \boldsymbol{\delta})$
  Update state $\boldsymbol{\theta}$ based on losses
end for
Inference:
Compute adapted embedding $\hat{\mathbf{z}} \leftarrow \mathbf{z}_0 + \boldsymbol{\delta}^*$
Predict $\hat{y}$ using scores from $\hat{\mathbf{z}}$
return $\hat{y}$

III. EXPERIMENTS

A. Datasets

Following the protocol of previous audio-prototype-based few-shot audio classification baseline [19], we conduct experiments on two datasets: ESC-50 [26] and FSD50K [27]. Specifically, ESC-50 consists of 2,000 audio samples distributed across 50 categories, with each 5-second clip annotated with a single label. The FSD50K dataset is larger, comprising over 50,000 audio clips spanning 200 sound classes. Compared to ESC-50, FSD50K features greater diversity and supports multi-label annotations, where a single sample may contain multiple sound events. We adhere to the standard evaluation protocols for both datasets: we perform 5-fold cross-validation on ESC-50, and utilize the official development and evaluation splits for FSD50K. We adopt accuracy and mean Average Precision (mAP) as evaluation metrics for ESC-50 and FSD50K, respectively.

B. Experimental Setup

Consistent with prior works [15], [19], we utilize the 2023 version of CLAP as our backbone model. By default, we set the fusion weight β to 0.5 for combining audio and text prototype similarities, so that both prototypes contribute equally. For the ESC-50 dataset, we set K to 3, while for FSD50K, we set K to 10, as the latter is more complex and potentially contains more noise. In few-shot settings, the support set size N is set to 3, and the experiments are repeated 10 times. For the CMA-ES-based test-time adaptation, we configure the number of iterations to 10, with a population size of 16 and an initial standard deviation $\sigma = 0.05$. We set τ to 10.

IV. RESULTS

A. Main Results

The experimental results are shown in Table II. We first evaluate the performance in a fully zero-shot scenario. The experimental results indicate that our designed prompt library achieves effective zero-shot audio classification performance, thereby demonstrating its generalizability and robustness. However, on FSD50K, utilizing the full prompt library yielded inferior results compared to our reproduction of TSPE, which employs 20 meticulously designed templates. We attribute this to the introduction of unnecessary noise when incorporating all 50 templates. Consequently, we recommend pruning the prompt library according to specific scenarios to enhance performance in zero-shot audio classification tasks.

Next, we evaluate the performance under the few-shot audio classification scenario. Initially, we observe that employing solely text prototypes yields performance comparable to using audio prototypes. Specifically, audio prototypes outperform text prototypes on the ESC-50 dataset, whereas the opposite trend is observed on FSD50K. We attribute this discrepancy to the higher complexity and noise levels inherent in the FSD50K samples. Consequently, in the low-shot regime (e.g., 3-shot), audio prototypes constructed from a limited support set tend to be less effective. On the FSD50K dataset, prompt selection significantly enhances text-based audio classification performance; however, this benefit is less pronounced on ESC-50. Consequently, in scenarios analogous to ESC-50, directly employing a text-prototype-based zero-shot approach may offer greater cost-effectiveness.

On the other hand, the joint utilization of audio and text prototypes (with $\beta = 0.5$) yields superior performance com-

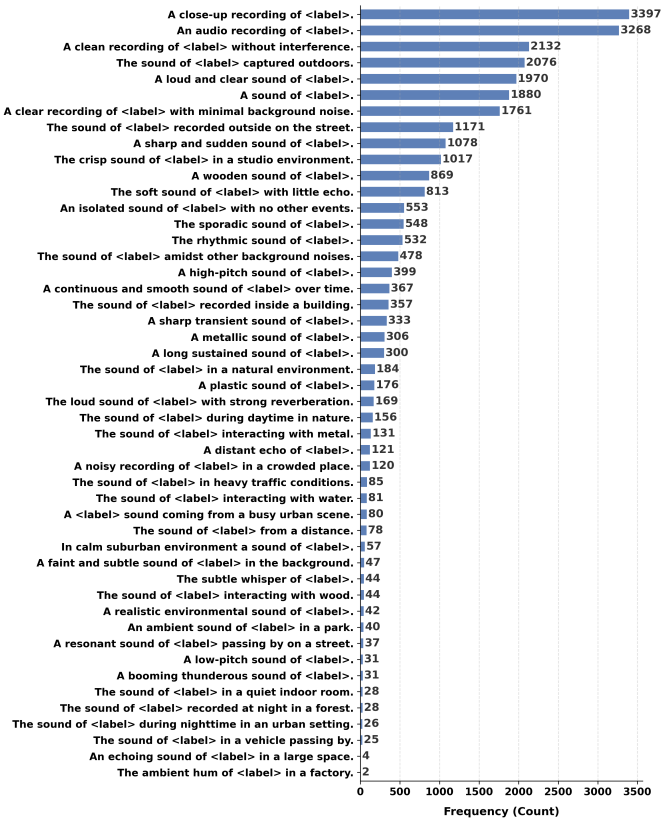

Fig. 2. Frequency of different templates.

pared to using either audio or text prototypes in isolation. This demonstrates that although CLAP maps audio and text into a common latent space, a residual modality gap persists. Crucially, this gap can be effectively bridged through a simple weighted summation of similarity scores.

B. Effectiveness of Test-Time Adaptation

We evaluated the impact of incorporating Test-Time Adaptation (TTA). As shown in Table II, the results indicate that CMA-ES-based TTA can, to a certain extent, enhance CLAP-based audio classification performance. With the introduction of TTA, our method achieves performance in the 3-shot setting comparable to that of the baseline in the 20-shot setting, further demonstrating the effectiveness of our approach.

Notably, this improvement is significantly more pronounced on the FSD50K dataset. We attribute this phenomenon to the complexity of the datasets. FSD50K features a larger number of classes and multi-label samples; in such scenarios, TTA effectively resolves prediction ambiguity. In contrast, ESC-50 is a relatively simple dataset where the CLAP model is already highly confident. Consequently, forcing adaptation in such high-confidence cases can be counterproductive. Therefore, we conclude that the deployment of TTA for audio classification should be context-dependent: it may be unnecessary for simple tasks with fewer categories, but it serves as a valuable booster for complex tasks characterized by high cardinality and noise.

C. Effectiveness of Prompt Templates

To explore the effects of different prompt templates in our designed prompt library, we counted the frequency of each template being selected as a top-K (K = 3) prompt across multiple experiments, as illustrated in Fig. 2.

Overall, the frequency of prompt templates exhibited a distribution similar to the normal distribution. Consistent with our expectations, general-purpose prompt templates appeared with the highest frequency. Notably, those containing the specific term "recording" were selected significantly more often than others. This phenomenon may be attributed to the pre-training bias of the CLAP model, where captions containing metadata-like descriptions (e.g., 'an audio recording of...') are prevalent.

In the intermediate frequency range, we also observed several interesting phenomena. For instance, "high-pitch" templates were selected significantly more frequently than "low-pitch" ones, and the template "A wooden sound of `<label>`" proved far more effective than "The sound of `<label>` interacting with wood". These findings provide valuable insights for manually constructing more effective prompts for CLAP.

Furthermore, some prompt templates were rarely or never selected. For instance, we designed the prompt template "A sound of `<label>` captured in an industrial environment" within our library. Despite its semantic plausibility, it never ranked among the top-3 candidates based on similarity scores. We therefore suggest that these templates could be optimized or omitted in future work.

D. Effectiveness of Text and Audio Prototypes

As illustrated in Fig. 3, we also investigated the impact of selecting different values of β on the two datasets without TTA. The experimental results demonstrate that as β increases from 0 to 1, the performance first increases and then decreases. Optimal performance is achieved on both datasets when β is set to approximately 0.6 and 0.7, respectively. This is consistent with our hypothesis that the synergistic effect of the two modalities can reduce noise introduced by support samples to a certain extent, thereby achieving better audio classification performance. On the other hand, the optimal value of β leans towards 1 rather than being exactly 0.5, suggesting that text prototypes contribute more significantly to the classification.

E. Visualization

As illustrated in Fig. 4, we employ t-SNE [28] to visualize the feature distributions of the "Dog" and "Cat" categories from the FSD50K dataset. The visualization reveals that despite CLAP's alignment objectives, a discernible gap persists between audio and text modalities. Specifically, visual inspection indicates distinct preferences across categories: for the "Dog" class, the audio prototype appears more representative of the sample distribution, whereas for the "Cat" class, the text prototype resides closer to the cluster centroid. However, in real-world scenarios, it is difficult to determine a priori which prototype is superior for a given category. Consequently, our approach jointly utilizes both prototypes, thereby bridging

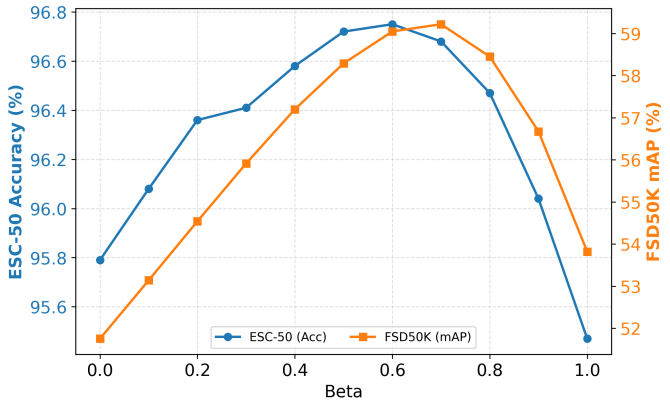

Fig. 3. Experimental results of different β.

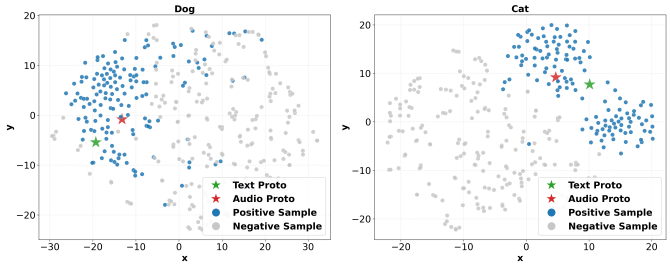

Fig. 4. t-SNE visualization of the "dog" and "cat" categories in FSD50K.

this modality gap and enhancing the robustness of audio classification against such uncertainty.

V. Conclusion

In this paper, we presented Audio-Guided Adaptive Prompt Ensembling (AGAPE), a novel training-free framework designed to adapt large-scale pre-trained models like CLAP to downstream few-shot audio classification tasks. By leveraging audio-guided prompt selection and multimodal prototypes, our method effectively mitigates the semantic misalignment often found in zero-shot inference. Furthermore, we integrated the CMA-ES algorithm to achieve efficient Test-Time Adaptation for CLAP without the computational burden of gradient updates. Extensive experiments demonstrate that AGAPE significantly outperforms the baseline. For future work, we plan to conduct deeper investigations into optimizing prompt templates and explore robust TTA strategies to address challenges in scenarios with extremely limited category diversity or class-imbalanced test streams. Additionally, we will investigate the impact of CMA-ES hyperparameters, such as the number of iterations, as they significantly affect both computational cost and overall performance.

References

[1] David Diaz-Guerra, Archontis Politis, Parthasaarathy Sudarsanam, Kazuki Shimada, Daniel A. Krause, Kengo Uchida, Yuichiro Koyama, Naoya Takahashi, Shusuke Takahashi, Takashi Shibuya, Yuki Mitsufuji, and Tuomas Virtanen, "Baseline models and evaluation of sound event localization and detection with distance estimation in DCASE2024 challenge," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2024 Workshop (DCASE2024)*, Tokyo, Japan, October 2024, pp. 41–45.

[2] Konstantinos Drossos, Stylianos I. Mimilakis, Shayan Gharib, Yanxiong Li, and Tuomas Virtanen, "Sound event detection with depthwise separable and dilated convolutions," in *2020 International Joint Conference on Neural Networks (IJCNN)*, 2020, pp. 1–7.

[3] Florian Schmid, Paul Primus, Toni Heittola, Annamaria Mesaros, Irene Martín-Morató, and Gerhard Widmer, "Low-complexity acoustic scene classification with device information in the DCASE 2025 challenge," 2025.

[4] Yanxiong Li, Jiaxin Tan, Guoqing Chen, Jialong Li, Yongjie Si, and Qianhua He, "Low-Complexity Acoustic Scene Classification Using Parallel Attention-Convolution Network," in *Interspeech 2024*, 2024, pp. 567–571.

[5] Youde Liu, Jian Guan, Qiaoxi Zhu, and Wenwu Wang, "Anomalous sound detection using spectral-temporal information fusion," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 816–820.

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

[7] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D. Plumbley, "PANNs: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 28, pp. 2880–2894, Nov. 2020.

[8] Heinrich Dinkel, Yongqing Wang, Zhiyong Yan, Junbo Zhang, and Yujun Wang, "CED: Consistent ensemble distillation for audio tagging," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 291–295.

[9] Noboru Harada, Daisuke Niizumi, Daiki Takeuchi, Yasunori Ohishi, and Masahiro Yasuda, "First-shot anomaly detection for machine condition monitoring: A domain generalization baseline," *Proceedings of 31st European Signal Processing Conference (EUSIPCO)*, pp. 191–195, 2023.

[10] Eunjeong Koh, Fatemeh Saki, Yinyi Guo, Cheng-Yu Hung, and Erik Visser, "Incremental learning algorithm for sound event detection," in *2020 IEEE International Conference on Multimedia and Expo (ICME)*, 2020, pp. 1–6.

[11] Riyansha Singh, Parinita Nema, and Vinod K Kurmi, "Towards robust few-shot class incremental learning in audio classification using contrastive representation," in *Interspeech 2024*, 2024, pp. 5023–5027.

[12] Yongjie Si, Yanxiong Li, Jialong Li, Jiaxin Tan, and Qianhua He, "Fully Few-shot Class-incremental Audio Classification Using Expandable Dual-embedding Extractor," in *Interspeech 2024*, 2024, pp. 4788–4792.

[13] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al., "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[14] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang, "CLAP: learning audio concepts from natural language supervision," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[15] Nishit Anand, Ashish Seth, Ramani Duraiswami, and Dinesh Manocha, "TSPE: Task-specific prompt ensemble for improved zero-shot audio classification," in *2025 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*. IEEE, 2025, pp. 1–5.

[16] Ashish Seth, Ramaneswaran Selvakumar, Sonal Kumar, Sreyan Ghosh, and Dinesh 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. 12376–12394.

[17] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu, "Learning to Prompt for Vision-Language Models," *International Journal of Computer Vision*, vol. 130, no. 9, pp. 2337–2348, Sept. 2022.

[18] Asif Hanif, Maha Tufail Agro, Mohammad Areeb Qazi, and Hanan Aldarmaki, "PALM: Few-shot prompt learning for audio language models," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Yaser Al-Onaizan, Mohit Bansal, and

Yun-Nung Chen, Eds., Miami, Florida, USA, Nov. 2024, pp. 18527–18536, Association for Computational Linguistics.

[19] James Taylor and Wolfgang Mack, "Improving Audio Classification by Transitioning from Zero- to Few-Shot," in *Interspeech 2025*, 2025, pp. 2585–2589.

[20] Yongjie Si, Yanxiong Li, Jiaxin Tan, Qianhua He, and Il-Youp Kwak, "Fully Few-shot Class-incremental Audio Classification Using Multi-level Embedding Extractor and Ridge Regression Classifier," in *Interspeech 2025*, 2025, pp. 1318–1322.

[21] Malik Boudiaf, Romain Mueller, Ismail Ben Ayed, and Luca Bertinetto, "Parameter-free online test-time adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 8344–8353.

[22] Jiaheng Dong, Hong Jia, Soumyajit Chatterjee, Abhirup Ghosh, James Bailey, and Ting Dang, "E-BATS: Efficient backpropagation-free test-time adaptation for speech foundation models," in *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[23] Nikolaus Hansen, "The CMA evolution strategy: A tutorial," 2023.

[24] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 29, pp. 3451–3460, Oct. 2021.

[25] OpenAI, "GPT-4 technical report," 2024.

[26] Karol J. Piczak, "ESC: Dataset for Environmental Sound Classification," in *Proceedings of the 23rd Annual ACM Conference on Multimedia*. pp. 1015–1018, ACM Press.

[27] Eduardo Fonseca, Xavier Favory, Jordi Pons, Frederic Font, and Xavier Serra, "FSD50K: An open dataset of human-labeled sound events," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 829–852, 2022.

[28] Laurens van der Maaten and Geoffrey Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.