Zero-Shot Audio Captioning Using Soft and Hard Prompts

Yiming Zhang, Xuenan Xu, Ruoyi Du, Haohe Liu, Yuan Dong, Zheng-Hua Tan, *Senior Member, IEEE*, Wenwu Wang, *Senior Member, IEEE*, Zhanyu Ma, *Senior Member, IEEE*

***Abstract*—In traditional audio captioning methods, a model is usually trained in a fully supervised manner using a human-annotated dataset containing audio-text pairs and then evaluated on the test set from the same dataset. Such methods have two limitations. First, these methods are often data-hungry and require time-consuming and expensive human annotations to obtain audio-text pairs. Second, these models often suffer from performance degradation in cross-domain scenarios, i.e., when the input audio comes from a different domain than the training set, and this issue has received little attention. To address these issues, we propose a new zero-shot method for audio captioning. Our method is built on the contrastive language-audio pre-training (CLAP) model. During training, the model reconstructs the ground-truth caption using the CLAP text encoder. In the inference stage, the model generates text descriptions from the CLAP audio embeddings of given audio inputs. To enhance the ability of the model in transitioning from text-to-text generation to audio-to-text generation, we propose to use the mixed-augmentations-based soft prompt to learn more robust latent representations, leveraging instance replacement and embedding augmentation. Additionally, we introduce the retrieval-based acoustic-aware hard prompt to improve the cross-domain performance of the model by employing the domain-agnostic label information of sound events. Extensive experiments on AudioCaps and Clotho benchmarks show the effectiveness of our proposed method, which outperforms other zero-shot audio captioning approaches for in-domain scenarios and outperforms the compared methods for cross-domain scenarios, underscoring the generalization ability of our method. The code is publicly available at https://github.com/XinMing0411/zero-shot-AAC.**

***Index Terms*—Audio captioning, zero-shot, contrastive language-audio pre-training, prompt engineering**

I. Introduction

AUDIO captioning is a sophisticated audio-to-text cross-modal translation task where a model is built to analyse the content of an audio clip and articulate it using natural language [1]–[5]. The generated captions encompass not only basic descriptions of sound events and scenes but also high-level semantic information, such as the relationships among events and physical properties of sounds. Recent advancements in audio captioning have significantly elevated the state-of-the-art. Most existing methods rely on fully supervised training, employing an encoder-decoder framework containing an audio encoder and a language decoder. As a result, these approaches are data-hungry and rely on a large amount of training data with human-annotated text descriptions.

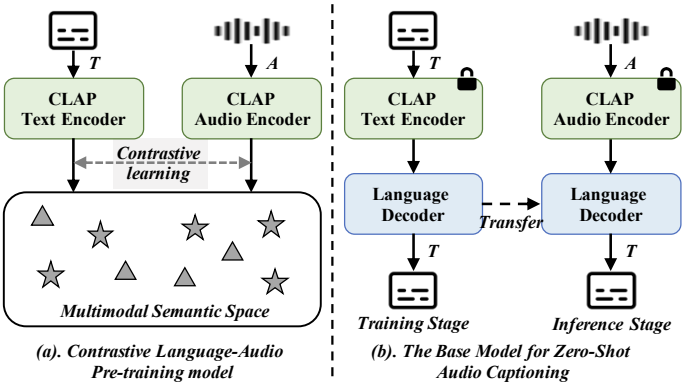


Fig. 1. (a) The structure of the CLAP model. Through contrast learning, CLAP maps the audio and text into the same semantic space. Grey triangles and pentagons represent audio and text embeddings, respectively. (b) The structure of the base zero-shot audio captioning model, where a language decoder is trained for text reconstruction using text data based on the CLAP text encoder. The CLAP audio encoder is combined with the language decoder to generate captions during inference.

However, data scarcity is a substantial challenge for audio captioning. The predominant audio captioning benchmark datasets, Clotho [2] and AudioCaps [3] contain only 19k and 49k audio-caption pairs in their training sets, respectively. These numbers pale in comparison to the vast datasets available for visual captioning (*e.g.*, about 414K paired data in the COCO Caption dataset [6]). To address this challenge, several researchers have proposed zero-shot audio captioning methods [7]–[10] that aim to generate audio captions without relying on human-annotated audio-text pairs. For example, Shaharabany *et al.* [7] introduced a training-free approach by optimizing the context cache during text generation using classifier guidance. Salewski *et al.* [8] introduced another training-free approach by reweighting the output probabilities based on audio data. These training-free methods, however, tend to exhibit limited performance in describing audio content.

Different from the training-free methods, the zero-shot methods in [9], [10] are developed with text-only training, by leveraging the Contrastive Language-Audio Pre-training (CLAP) model [11] (as shown in Fig 1 (a)), which aligns audio and text data in the embedding space. In these methods,

Y. Zhang, R. Du, Y. Dong, and Z. Ma are with the Pattern Recognition and Intelligent System Laboratory, School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing 100876, China. E-mail: {zhangyiming, duruoyi, mazhanyu, yuandong}@bupt.edu.cn.

X. Xu is with the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, 200240, China. Email: wsntxxn@sjtu.edu.cn.

Z.-H. Tan is with the Department of Electronic Systems, Aalborg University, Aalborg 9220, Denmark. E-mail: zt@es.aau.dk.

H. Liu and W. Wang are with the Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford, GU2 7XH, United Kingdom. E-mail: {haohe.liu, w.wang}@surrey.ac.uk.

(Corresponding author: Zhanyu Ma)

text-only data is utilized during training for learning the text reconstruction models. During inference, the CLAP text encoder is replaced by the CLAP audio encoder to generate the descriptions of the input audio. However, the two modalities are not well aligned, as the audio and text embeddings may form separate clusters, leading to a gap between the text embedding and its corresponding audio embedding [10], [12]. This gap impacts on the method's generalisation performance, particularly when the text-to-text generation model obtained during training is transitioned to audio-to-text generation in inference. To mitigate this, Deshmukh *et al.* [9] applied Gaussian noise to perturb the CLAP embeddings during training. However, with Gaussian noise, limited variations are added to semantic representations due to its simple pattern, leading to limited improvement of the model's robustness to the gap. Kouzelis and Katsouros [10] proposed a projection-based decoding strategy that projects audio embeddings into the text space through a weighted combination of all the text embeddings from training. However, this method requires the entire training set to be stored, leading to high demands in memory and computational load during inference.

In addition, existing fully supervised and zero-shot methods typically consider the model performance solely in in-domain scenarios, where the training and test sets come from the same source. In contrast, cross-domain scenarios where the training and test sets come from different sources have received little attention, despite being more common in real-world applications. These methods learn audio-to-text (fully supervised) or text-to-text (zero-shot) mapping with limited in-domain data, which may cause the model to overfit to dataset-specific patterns, rather than learning generalizable audio event features. As a result, they may suffer from significant performance degradation in cross-domain scenarios and fail to accurately describe out-of-domain audio clips.

To address the limitations of the existing zero-shot methods and improve their cross-domain performance, we propose a new zero-shot audio captioning method, involving a mixed-augmentations-based soft prompt and a retrieval-based acoustic-aware hard prompt. Here, "soft prompt" is a technical term borrowed from Natural Language Processing (NLP) [13], [14], referring to the continuous embeddings used as inputs to a language model, while the "hard prompt" is a term used to refer to static discrete textual prompts. In our method, the soft prompt leverages a mixed-augmentations strategy, incorporating instance replacement and embedding augmentation to enhance the model's generalization ability and robustness. With instance replacement, the original input is substituted with semantically similar but distinct text, introducing semantic perturbations that expose the model to a greater variety of semantic representations. This helps the model learn more robustly from inherent distributional variations in feature representations and reduces the risk of overfitting to specific patterns. With embedding augmentation, random perturbations are added to the text embeddings, inspired by the Gaussian noise injection concept presented in [9], [10], thus further enhancing the model's ability to transition smoothly from text-to-text generation to audio-to-text generation.

Additionally, to further enhance the model's cross-domain performance, we present a hard prompting method, leveraging an acoustic-aware retrieval strategy. More specifically, this strategy guides the model in capturing and utilizing relevant acoustic information across diverse domains by employing domain-agnostic acoustic event labels, thereby leading to more accurate captions. Through extensive experiments, we demonstrate the superior performance of the proposed method, as compared with the baseline zero-shot audio captioning methods for in-domain scenarios, and fully supervised and zero-shot audio captioning methods for cross-domain scenarios.

II. Related Work

In this section, we first give a brief overview of CLAP, whose multimodal semantic space provides the foundation of our proposed method. Then, we introduce traditional fully-supervised audio captioning methods and recent zero-shot audio captioning methods.

A. Contrastive Language-Audio Pre-training (CLAP)

CLAP [11], [15]–[17] utilizes contrastive learning to pre-train language-audio models, which map both audio and text into the same semantic space on large-scale audio-text pairs. CLAP contains two encoders: an audio encoder and a text encoder. The audio encoder $f_{\text{clap}}^{\text{Audio}}(\cdot)$ often uses well-performed audio classification models, which can be convolutional neural networks [18] or Transformers [19], as the backbone. The text encoder $f_{\text{clap}}^{\text{Text}}(\cdot)$ is usually a pre-trained masked language model (*e.g.*, BERT [20], RoBERTa [21]). CLAP utilizes noisy pairwise data for training based on the InfoNCE loss [22], learning the alignment between text and audio embeddings in a multimodal semantic space.

In this work, we use CLAP text encoder $f_{\text{clap}}^{\text{Text}}(\cdot)$ for text reconstruction in the training stage. In the inference stage, $f_{\text{clap}}^{\text{Text}}(\cdot)$ is replaced with the audio encoder $f_{\text{clap}}^{\text{Audio}}(\cdot)$ to generate descriptive text for a given audio.

B. Fully Supervised Audio Captioning

With the success of the Detection and Classification of Acoustic Scenes and Events (DCASE) Challenges[1], the flagship international data challenge in acoustic scene and event understanding [4], fully supervised audio captioning has seen significant advancements. Most research on audio captioning utilizes an audio encoder-language decoder framework trained on human-annotated audio-text paired data. These studies employed the audio encoder to extract embeddings of the input audio clip A, which are then fed into the language decoder to generate corresponding descriptive caption T. Mei *et al.* [23] proposed a full Transformer-based audio captioning method to improve the capability of modelling global and fine-grained temporal information. Ye *et al.* [24] proposed a fully supervised audio captioning model based on the multi-modal attention module, which utilizes acoustic and semantic information to generate captions. Xu *et al.* [25] pre-trained the audio encoder on audio-text retrieval tasks, enhancing the representation capability of the audio encoder for audio

[1] https://dcase.community/

captioning. Kim *et al.* [26] used a pre-trained language model (GPT-2) as the decoder to ensure text generation capability, with global and temporal information from the input audio as the prefix to guide the output of the decoder. Koh *et al.* [27] introduced the reconstruction latent space similarity regularisation to regulate model training in audio captioning. Zhang *et al.* [28] proposed a two-stage audio captioning approach to mitigate the effects of semantic disparity among the audio captions by incorporating feature space regularisation and improving the accuracy of the model-generated description text. Ghosh *et al.* [29] proposed a retrieval-augmented audio captioning method that uses the CLAP encoder to retrieve captions similar to the input audio from the external database and then the retrieved captions are used as extra guidance for the decoder to generate descriptive text.

However, the high cost of collecting audio-text paired data has limited the applicability of these methods. Therefore, reducing the dependency of audio captioning models on paired data has emerged as a prominent research focus in audio captioning.

C. Zero-Shot Audio Captioning

To further reduce the cost of paired data collection, zero-shot audio captioning aims to generate audio captions without using well-paired audio-text data [7]. Audio Flamingo [30] used a large-scale weakly aligned audio-text pair dataset to train the audio language model and evaluated the model on the AudioCaps benchmark without fine-tuning. Inspired by recent advancements in the image-text field [31]–[35], where noise injection and alignment strategies have been shown to enhance zero-shot capabilities, some works have extended these ideas to zero-shot audio captioning by utilizing pre-trained CLAP models or Large Language Models (LLMs) [7]–[10]. We categorize these studies into training-free methods and text-only training based methods. The training-free methods achieve zero-shot audio-to-text generation based on the pre-trained models without performing additional training of the language models. Shaharabany *et al.* [7] designed a classifier-guided zero-shot method, which employed the ImageBind model [36] and the pre-trained binary audibility model as classifiers to guide the LLMs towards captions. Salewski *et al.* [8] proposed a similar approach, where the similarities between the CLAP embeddings of the input audio data and the previously generated tokens are used to derive the probability of the current tokens being selected from the vocabulary for caption generation.

However, training-free methods usually achieve poor performance and fall short in zero-shot captioning capability. Compared to training-free methods, the text-only training based methods [9], [10] rely more on the multimodal modelling capabilities provided by the pre-trained CLAP model. These methods typically train the language decoder using only textual data, but not any paired audio data. However, the paired audio-text data are not well aligned within the CLAP's semantic space, as the embeddings from different modalities may form distinct clusters, resulting in a gap between text embeddings and their corresponding audio embeddings [12]. This misalignment can adversely impact on the performance of the text-only training methods, particularly when the text-to-text generation model obtained in training is transitioned to audio-to-text generation in inference.

Inspired by zero-shot image captioning methods [31], [32], Kouzelis and Katsouros [10] proposed a noise injection and an embedding shift strategy to reduce the modality gap during training, and a nearest-neighbour strategy or a projection-based decoding strategy to map the audio embeddings to text embeddings in inference by utilizing the textual data stored from the training set. Among these strategies, the projection-based decoding strategy demonstrated superior performance compared to the others. Deshmukh *et al.* [9] also proposed a noise injection method which injects a random variable into the text embeddings. Although these text-only training-based methods excel in in-domain situations, they often overlook cross-domain situations.

Inspired by the text-only training-based methods discussed above and the noise injection idea presented in [9], [10], we devise a mixed-augmentations-based soft prompt to enhance the model's generalization ability in transitioning from text-to-text generation to audio-to-text generation. In addition, we employ a retrieval-based acoustic-aware hard prompt to improve cross-domain performance by incorporating label information of acoustic events.

III. Proposed Method

In this work, we propose an alternative zero-shot audio captioning method to alleviate the reliance of the model on audio-text paired data in traditional fully supervised audio captioning methods. The overall architecture of our proposed method is illustrated in Fig. 2. In the text-only training stage, we use the CLAP text encoder to extract the embedding of the input text, and then the mixed-augmentations-based soft prompt (described in Sec. III-A) and the acoustic-aware hard prompt (described in Sec. III-B) are fed to the language decoder to reconstruct the given text. In the zero-shot inference stage (described in Sec. III-C), we replace the CLAP text encoder with the CLAP audio encoder to generate the description of the input audio.

A. The Soft Prompt based on Mixed-augmentations

An alternative approach to the zero-shot audio captioning task is to exploit the CLAP model, as shown in Fig 1 (b), where audio and text embeddings are aligned in the same semantic space through contrastive learning. During training, for a given input text T, the language decoder is trained to reconstruct the input text from the CLAP text embedding. The paired audio-text data, however, may not be well aligned within the multimodal semantic space. This can result in a gap between text embeddings and audio embeddings, which may adversely impact the ability of the model (as shown in Fig 1 (b)) to generalize effectively from training (text-to-text generation) to inference (audio-to-text generation). To address this issue, we employ a mixed-augmentations strategy, which includes instance replacement and embedding augmentation, to enable the model to learn more robust latent representations.

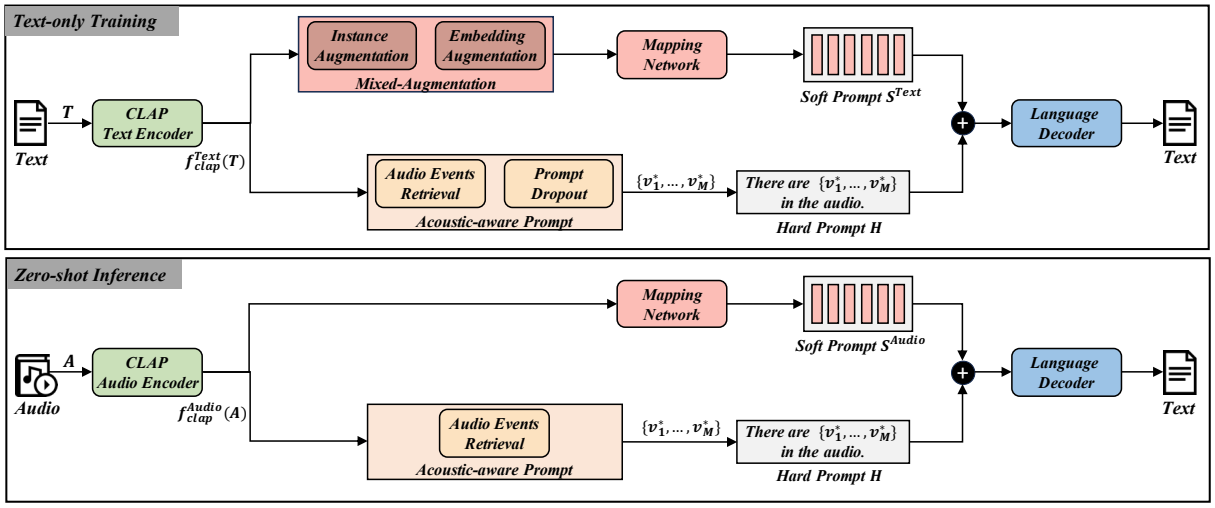


Fig. 2. The overall architecture of our proposed method. Specifically, in the training stage, we reconstruct the input text based on the acoustic-aware hard prompt and soft prompt with only textual data, so training does not require any paired data. During inference, we replace the CLAP text encoder $f_{\text{clap}}^{\text{Text}}(\cdot)$ with the CLAP audio encoder $f_{\text{clap}}^{\text{Audio}}(\cdot)$ to generate the descriptive text of the input audio.

Instance Replacement: First, we retrieve N captions in the text corpus $\mathcal{T}$ that are semantically similar to the input text T as a candidate set $\mathcal{C}_N$:

$$\mathcal{C}_N = \left\{ \underset{T_n^* \in \mathcal{T}}{\text{argmax}_N} \frac{f_{\text{clap}}^{\text{Text}}(T) \cdot f_{\text{clap}}^{\text{Text}}(T_n^*)}{\|f_{clap}^{\text{Text}}(T)\| \cdot \|f_{clap}^{\text{Text}}(T_n^*)\|} \right\}, \tag{1}$$

where argmax_N select text embeddings with top-N highest similarities, $f_{clap}^{\text{Text}}(T) \in \mathbb{R}^F$ is the CLAP text embedding of the input text T, $\|\cdot\|$ represents the norm of the embedding vector, T_n^* is the n-th candidate text, and F is the dimension of the CLAP embedding and $n \leq N$.

Then, $f_{\text{clap}}^{\text{Text}}(T_n^*)$ is randomly selected from the candidate set of text embeddings $\mathcal{C}_N$ to replace the original text embedding $f_{\text{clap}}^{\text{Text}}(T)$.

Embedding Augmentation: Inspired by the noise injection method used in [9], [10], we add a Gaussian noise $\epsilon \sim \mathcal{N}(0, \sigma)$ into the candidate text embedding $f_{\text{clap}}^{\text{Text}}(T_n^*)$ to obtain the noisy text embedding $f_{\text{clap}}^{\text{Text}}(T_n^*) + \epsilon$, where σ is the standard deviation and $\epsilon \in \mathbb{R}^F$.

Then, the noisy text embedding is fed into the mapping network $\mathcal{M}(\cdot)$ to get the text soft prompt $S^{\text{Text}} \in \mathbb{R}^{K \times F_g}$ for the language decoder,

$$S^{\text{Text}} = \mathcal{M}\left(f_{\text{clap}}^{\text{Text}}(T_n^*) + \epsilon\right), \tag{2}$$

where the soft prompt S^{Text}, a term from NLP [13], [14], refers to the inclusion of the continuous embeddings as the input to the language decoder, F_g is the dimension of the embedding for the language decoder, and K is the total length of the soft prompt S^{Text}.

B. Acoustic-aware Hard Prompt based on Retrieval

Acoustic labels signify the acoustic events or scenes within an audio clip, together with their characteristics. For example, the audio label ("*gunshots*") suggests that the audio clip likely contains sharp and loud pops, while labels like "animal" can encompass diverse sound patterns across datasets. Acoustic labels are typically obtained through pre-trained models that identify specific events in a given audio clip. Compared to the soft prompt S derived from the encoder and mapping network, acoustic labels provide domain-agnostic context about the potential audio content. This information helps the model adapt to variations across datasets and reduces the risk of overfitting to dataset-specific details. To further enhance model performance, we exploit additional acoustic-aware prompts in the text decoding process, enabling the model to focus on relevant acoustic events and improving its contextual understanding.

Acoustic-aware Hard Prompt: Firstly, we need to build the vocabulary of audio events $\mathcal{V}$. We use the labels of AudioSet [37], a prevalent benchmark dataset for the audio tagging task. AudioSet contains 527 audio categories and covers various human and animal sounds, musical instruments and genres, and environmental sounds. Therefore, the vocabulary of audio events $\mathcal{V}$ is a set of 527 audio event labels $\{v_1, \ldots, v_{527}\}$, where v represents the audio event classes.

Given the text embedding $f_{\text{clap}}^{\text{Text}}(T)$, we retrieve M audio events that are most similar to $f_{\text{clap}}^{\text{Text}}(T)$ from the vocabulary $\mathcal{V}$ based on the cosine similarity of the CLAP embeddings:

$$\{v_1^*, \ldots, v_M^*\} = \left\{ \underset{v_m^* \in \mathcal{V}}{\text{argmax}_M} \frac{f_{\text{clap}}^{\text{Text}}(T) \cdot f_{\text{clap}}^{\text{Text}}(v_m^*)}{\|f_{\text{clap}}^{\text{Text}}(T)\| \cdot \|f_{\text{clap}}^{\text{Text}}(v_m^*)\|} \right\}, \tag{3}$$

where v_m^* is the m-th audio event. Therefore, the retrieved audio events are used to construct the acoustic-aware hard prompt $H =$ "There are $\{v_1^*, \ldots, v_M^*\}$ in the audio.".

We concatenate the embeddings of the acoustic-aware prompt H and the soft prompt S^{Text} along the sequence and feed them into the language decoder to reconstruct the input text T in an auto-regressive manner. The embeddings of the acoustic-aware prompts H are extracted using the embedding layer of the language decoder, resulting in embeddings with the shape $R^{L \times F_g}$, where L is the length of the tokenized sequence of the entire acoustic-aware prompt H. The model

is trained using the cross-entropy loss:

$$\mathcal{L} = -\frac{1}{|T|}\sum_{i=1}^{|T|}\log p_\theta(t_i|T_{<i}, H, S^{\text{Text}}) \tag{4}$$

where $|T|$ is the length of the input T, t_i is the i-th word token of T, $T_{<i}$ includes all the tokens from the start of T up to just before the i-th token, and $p_\theta(\cdot)$ represents the probability distribution of the output token given the previous tokens, modelled by the language decoder, with θ representing all the parameters of the model.

Event Label Dropout: To make the model robust to retrieval errors and avoid over-reliance on the acoustic-aware prompt H, we propose an event label dropout technique, where each retrieved audio event v_m^* in $\{v_1^*, \ldots, v_M^*\}$ is either retained or dropped during training. Specifically, each audio event v_m^* is retained with a probability of $1-\beta$ or dropped with a probability of β, as follows:

$$v_m^* = \begin{cases} v_m^*, & \text{with probability } 1-\beta \\ \emptyset, & \text{with probability } \beta \end{cases} \tag{5}$$

where v_m^* is either retained or replaced by $\emptyset$ to indicate that the event has been dropped. In this way, the model is trained to avoid simply concatenating audio events from acoustic-aware prompt H to generate the caption while ignoring the information in soft prompt S^{Text}.

C. Zero-shot Inference

In the inference stage, the model needs to generate the text description T for the given audio clip A. To this end, we use a CLAP audio encoder to extract the CLAP embedding $f_{\text{clap}}^{\text{Audio}}(A)$ of this audio clip, which replaces the text encoder used during model training. We process the embedding $f_{\text{clap}}^{\text{Audio}}(A)$ in a similar way to obtain its audio soft prompt S^{Audio} and retrieved audio events $\{v_1^*, \ldots, v_M^*\}$, excluding the mixed-augmentations and the event label dropout, by reformulating Eq. (2) and Eq. (3) as follows:

$$S^{\text{Audio}} = \mathcal{M}\left(f_{\text{clap}}^{\text{Audio}}(A)\right), \tag{6}$$

$$\{v_1^*, \ldots, v_M^*\} = \left\{ \underset{v_m^* \in \mathcal{V}}{\text{argmax}}_M \frac{f_{\text{clap}}^{\text{Audio}}(A) \cdot f_{\text{clap}}^{\text{Text}}(v_m^*)}{\|f_{\text{clap}}^{\text{Audio}}(A)\| \cdot \|f_{\text{clap}}^{\text{Text}}(v_m^*)\|} \right\}, \tag{7}$$

Next, the retrieved audio events are used to construct the acoustic-aware hard prompt H, formatted as "There are $\{v_1^*, \ldots, v_M^*\}$ in the audio.". We then concatenate the word embeddings of the hard prompt H with the soft prompt S^{Audio}, and feed them into the language decoder to predict the caption T in an auto-regressive manner.

IV. Experimental Settings

This section introduces the experimental settings, including model architectures, datasets, baselines and metrics, and implementation details.

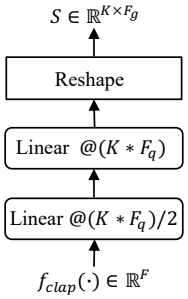


Fig. 3. The model architecture of the mapping network. It is a simple MLP containing two linear layers that maps the CLAP embedding $f_{\text{clap}}(\cdot)$ into the soft prompt S. The number after the "@" symbol indicates the feature dimension of the linear layer output.

A. Model Architectures

CLAP Encoder: In this work, we use the CLAP model[2] as our encoder which is only trained on WavCaps [11], which does not contain any human-annotated data. The CLAP audio encoder is an `HTSAT` [19] and the text encoder is a `RoBERTa` [21]. All audio clips are randomly cropped or padded to 10 seconds and sampled at 32 kHz. We use a 64-dimensional log-Mel spectrogram extracted from a Hanning window of $1,024$ points with a hop size of 320 as the input audio feature. The dimension F of the CLAP embedding is $1,024$, and all the parameters in the CLAP encoder are frozen.

Mapping Network and Language Decoder: The mapping network transforms the CLAP embedding $f_{\text{clap}}(\cdot)$ into soft prompt S. As shown in Fig. 3, the mapping network is a simple Multi-Layer Perceptron (MLP) which contains two linear layers. For the language decoder, we use the pre-trained `GPT2-base`[3] to generate text. The embedding dimension F_g is 768, and all model parameters except the CLAP encoder are trainable.

B. Datasets

We conducted our experiments on audio captioning benchmark datasets, AudioCaps [3] and Clotho [2]. AudioCaps is the largest human-annotated audio captioning dataset and contains 51K audio clips with one caption per audio clip in the training set and five captions per audio clip in the evaluation set. The audio clips are a subset of AudioSet and annotated with the aid of visual information. Clotho is the official benchmark in the DCASE challenge, and contains about 3.8K audio clips, where each audio clip has five captions. The audio clips in Clotho are collected from the Freesound platform [39], which are annotated with only audio signals, without using any visual signals. Hence, AudioCaps and Clotho are clearly different in terms of audio sources and textual style [5].

C. Baselines

In this work, we compare our method with two types of baselines: fully supervised audio captioning methods and zero-shot audio captioning methods. They have been introduced in detail in Sec. II.

[2]The weights file of the CLAP model: https://drive.google.com/drive/folders/1MeTBren6LaLWiZI8_phZvHvzz4r9QeCD

[3]The weights file of the pre-trained `GPT2-base`: https://huggingface.co/openai-community/gpt2 [38]

TABLE I
EXPERIMENTAL RESULTS FOR IN-DOMAIN SCENARIOS ON AUDIOCAPS.

Method	$BLEU_1$	$BLEU_4$	$ROUGE_L$	CIDEr	METEOR	SPICE	SPIDEr
Fully Supervised Audio Captioning							
Prefix AAC [26]	71.3 †	30.9 †	50.3 †	73.3 †	24.0 †	17.7 †	45.5 †
RECAP [29]	**72.8** †	**31.7** †	**52.1** †	**75.0** †	**25.2** †	18.3 †	**47.2** †
ACT [23]	68.4 ± 0.44 ⋆	25.2 ± 0.99 ⋆	48.0 ± 0.35 ⋆	67.5 ± 1.90 ⋆	22.8 ± 0.27 ⋆	16.9 ± 0.51 ⋆	42.2 ± 1.09 ⋆
	64.7 †	25.2 †	46.8 †	67.9 †	22.2 †	16.0 †	42.0 †
MAAC [24]	64.0 ± 0.60 ⋆	24.3 ± 0.55 ⋆	44.7 ± 0.25 ⋆	59.3 ± 1.05 ⋆	21.0 ± 0.15 ⋆	14.4 ± 0.38 ⋆	36.9 ± 0.54 ⋆
Xu *et al.* [25]	67.6 ± 0.21 ⋆	27.2 ± 0.33 ⋆	49.7 ± 0.17 ⋆	73.8 ± 1.21 ⋆	24.7 ± 0.06 ⋆	**18.4** ± 0.06 ⋆	46.1 ± 0.62 ⋆
Ours-FS	70.3 ± 0.17	26.8 ± 0.48	45.7 ± 0.18	74.5 ± 0.80	24.4 ± 0.15	18.2 ± 0.19	46.2 ± 0.39
Zero-Shot Audio Captioning							
Audio Flamingo [30] †	–	–	–	50.2	–	–	–
Shaharabany *et al.* [7] †	–	9.8	8.2	9.2	8.6	–	–
ZerAuCap [8] †	–	6.8	33.1	28.1	12.3	8.6	18.3
NoAudioCaptioning [9] ⋆	59.2 ± 1.43	15.0 ± 0.66	40.4 ± 0.37	42.4 ± 1.58	19.6 ± 0.69	13.6 ± 0.51	28.0 ± 0.96
WSAC [10] ⋆	61.1 ± 0.48	17.1 ± 0.28	43.5 ± 0.36	56.4 ± 0.44	**23.2** ± 0.09	**16.3** ± 0.29	36.3 ± 0.31
Ours	**66.0** ± 0.15	**21.3** ± 0.48	**45.7** ± 0.18	**64.4** ± 0.61	22.0 ± 0.23	15.6 ± 0.23	**40.0** ± 0.33

† The original results are listed in the paper.
⋆ The results are re-implemented by us.

TABLE II
THE EXPERIMENTAL RESULTS FOR IN-DOMAIN SCENARIOS ON THE CLOTHO DATASET

Method	$BLEU_1$	$BLEU_4$	$ROUGE_L$	CIDEr	METEOR	SPICE	SPIDEr
Fully Supervised Audio Captioning							
Prefix AAC [26]	56.0 †	16.0 †	37.8 †	39.2 †	17.0 †	11.8 †	25.5 †
RECAP [29]	56.3 †	16.5 †	38.3 †	39.8 †	17.9 †	12.2 †	21.4 †
ACTUAL [28]	56.6 †	16.1 †	37.5 †	40.9 †	17.6 †	12.1 †	26.5 †
RLSSR [27]	55.1 †	16.8 †	37.3 †	38.0 †	16.5 †	11.1 †	24.6 †
ACT [23]	**58.4** ± 0.21 ⋆	**16.9** ± 0.30 ⋆	38.5 ± 0.30 ⋆	41.6 ± 0.46 ⋆	17.8 ± 0.08 ⋆	12.1 ± 0.14 ⋆	26.9 ± 0.26 ⋆
MAAC [24]	57.0 ± 0.53 ⋆	16.0 ± 0.40 ⋆	37.7 ± 0.37 ⋆	41.3 ± 0.56 ⋆	17.7 ± 0.22 ⋆	12.3 ± 0.13 ⋆	26.8 ± 0.32 ⋆
	57.7 †	17.4 †	37.7 †	41.9 †	17.4 †	11.9 †	26.9 †
Xu *et al.* [25]	56.9 ± 0.15 ⋆	16.0 ± 0.39 ⋆	37.9 ± 0.33 ⋆	41.8 ± 0.69 ⋆	17.9 ± 0.15 ⋆	12.7 ± 0.07 ⋆	27.3 ± 0.34 ⋆
	-	16.4 †	38.6 †	42.1 †	-	12.6 †	27.4 †
Ours-FS	58.0 ± 0.54	16.7 ± 0.19	**38.7** ± 0.10	**42.6** ± 0.27	**18.0** ± 0.05	**12.8** ± 0.07	**27.7** ± 0.15
Zero-Shot Audio Captioning							
ZerAuCap [8] †	–	2.9	25.4	14.0	9.4	5.3	9.7
NoAudioCaptioning [9] ⋆	51.8 ± 1.02	11.3 ± 0.80	34.7 ± 0.87	29.2 ± 1.25	15.6 ± 0.38	10.3 ± 0.24	19.7 ± 0.66
WSAC [10] ⋆	54.5 ± 0.05	12.6 ± 0.14	35.9 ± 0.04	35.7 ± 0.33	16.9 ± 0.02	11.8 ± 0.01	23.8 ± 0.17
Ours	**56.4** ± 0.24	**15.6** ± 0.22	**37.5** ± 0.17	**40.3** ± 0.47	**17.3** ± 0.17	**11.9** ± 0.19	**26.1** ± 0.27

† The original results are listed in the paper.
⋆ The results are re-implemented by us.

Fully Supervised Audio Captioning: The fully supervised methods we compared include: *ACT* [23], *MAAC* [24], *Xu et al.* [25], *Prefix AAC* [26], *RLSSR* [27], *RECAP* [29], and *ACTUAL* [28]. All of which are open sourced and not trained with any additional data.

Zero-shot Audio Captioning: The zero-shot methods we compared include: *Audio Flamingo* [30], *Shaharabany et al.* [7], *ZerAuCap* [8], *NoAudioCaptioning* [9], and *WSAC* [10]. Audio Flamingo [30] is a large audio language model and achieves SOTA in several audio understanding tasks. Shaharabany et al. [7] and ZerAuCap [8] are training-free zero-shot methods. NoAudioCaptioning [9] and WSAC [10] are text-only training based methods. It should be noted that we re-implemented NoAudioCaptioning [9] and WSAC [10] using the same pre-trained CLAP model to ensure the fairness of the comparison.

D. Metrics

Similar to other audio captioning works, we use common captioning metrics, including $BLEU_n$ [40], $ROUGE_L$ [41], *METEOR* [42], *CIDEr* [43], *SPICE* [44], and *SPIDEr* [45] for evaluation. For all metrics, higher scores indicate better performance.

E. Implementation Details

In our work, we train the network using the AdamW optimizer with a weight decay of 0.02, an initial learning rate of 1×10^{-5}, a batch size of 32, a warm-up iteration of $3,000$ and a total training iteration of $15,000$. The model is trained on a 2080Ti GPU. We construct the hyperparameter tuning experiments on the Clotho dataset (described in the Sec. V-D). We set the number of candidates N in instance replacement as 5, the number of audio events M in the acoustic-aware prompt as 4, the noise variance σ in embedding augmentation as 0.1, the length K of the soft prompt as 10, and the dropout rate β of event labels as 0.6. We use beam search with a beam size of 3 to generate captions during inference.

V. RESULTS AND DISCUSSION

This section shows results followed by discussions of comparative experiments. In all tables, the **bold** font represents the **best** result for each metric in the same setting. Some works do not provide cross-domain results so we re-train these models using five different random seeds and report the mean and standard deviation of metrics.

A. In-domain Audio Captioning

Tables I and II compare our proposed method and baselines for in-domain scenarios, where the training and test sets come from the same benchmark dataset. It should be specially noted that we re-implemented the text-only training-based baseline methods using the same CLAP model to ensure a fair comparison. The fully supervised methods use audio-text paired data. The *Ours-FS* model operates in a fully supervised setting, where we use audio data and the CLAP audio encoder instead of the CLAP text encoder and mixed-augmentations strategy during training.

We have the following observations from the results for in-domain scenarios on the Clotho and AudioCaps datasets: 1) The fully supervised audio captioning methods tend to achieve better experimental performance than the zero-shot audio captioning methods. This is expected as the fully-supervised methods are trained using audio-text pairs, and the models learn the "audio-to-text" conversion ability well. The zero-shot methods suffer from the need to migrate from "text-to-text" in training to "audio-to-text" in inference, thus the discrepancy between training and inference results in decreased in-domain performance. 2) Our proposed zero-shot method outperforms other zero-shot audio captioning methods in most metrics. We attribute this improvement to the use of mixed-augmentations-based soft prompt and retrieval-based acoustic-aware hard prompt in model training. The mixed-augmentations strategy introduces partial perturbations to the CLAP embeddings, allowing the model to learn robust representations [46] and enhancing its generalization ability from text-to-text generation to audio-to-text generation. In addition, the acoustic-aware hard prompt provides acoustic event information, which can help improve the model's performance by making it more context-aware. Furthermore, we found that both the mixed-augmentations strategy and the acoustic-aware hard prompt continue to be effective in the fully-supervised setting. The *Ours-FS* method demonstrates superior performance on the Clotho dataset and comparable performance on the AudioCaps dataset, further validating the effectiveness of these strategies. 3) Our proposed zero-shot method, which does not utilize any paired data, achieves about 85.9% of the performance of the fully-supervised state-of-the-art method RECAP [29], which obtains a *CIDEr* score of 75.0 on the AudioCaps dataset, and about 96.4% of the performance of Xu *et al.* [25], attaining a *CIDEr* score 41.8 on the Clotho dataset. This demonstrates the effectiveness of our method.

B. Cross-domain Audio Captioning

Cross-domain scenarios are where the training and test sets come from different benchmark datasets. The model is trained using only data from the *Source* benchmark, and any data from the training set of the *Target* benchmark is prohibited. In the real world, the audio in *Target* domain is often agnostic, so the cross-domain performance can better represent the generalizability of the model in real-world applications.

Table III shows the experimental results of our method and baseline methods in cross-domain scenarios, where the "*Source* $\Longrightarrow$ *Target*" refers to the scenario where the model is trained on the training set of the *Source* dataset and evaluated on the test set of the *Target* dataset. It is important to note that neither the training nor the validation set of the *Target* dataset is used in model training and selection. From the experimental results, we find the following: 1) Both fully-supervised and zero-shot methods exhibit performance degradation in cross-domain scenarios compared to the in-domain scenarios, particularly on the AudioCaps dataset. 2) The fully-supervised methods often struggle to generalize beyond the training domain. In contrast, zero-shot methods benefit from augmentation strategies, such as noise injection, which help the model learn more robust feature representations. As a result, the performance degradation for the fully supervised methods tends to be more severe, as compared to the zero-shot methods. 3) Our proposed model outperforms all the baselines, including both fully-supervised and zero-shot methods, across most evaluation metrics. The mixed-augmentation-based soft prompt introduces perturbations that contribute to learning robust feature representations. Moreover, the acoustic-aware hard prompts play a critical role by providing domain-agnostic guidance based on labels of the acoustic events.

The success of zero-shot audio captioning approach enables the scaling up of the training of audio captioning models using text-only data. This is promising since human-annotated audio captioning data is scarce. However, there is limited availability of text data for describing audio, compared to the textual data available from other fields, such as visual or music captioning. This leads to two options for scaling up the training data: 1) using data from other fields, or 2) generating synthetic audio captioning data from LLMs. The former is human-annotated real data, presenting better diversity and larger scale, while the latter's domain is better aligned with the target. Therefore, we explore whether the dataset size or the field alignment plays an important role by comparing training data from different fields.

Table IV shows the cross-domain performance of our proposed method trained on textual data from different fields and evaluated on Clotho and AudioCaps. We use the textual data from three fields for training: LLM-generated audio captioning corpus (*ChatGPT*[4], *FreeSound*[5], *WavCaps* [11]), visual captioning corpus (*COCO Captions* [6]), and music captioning corpus (*MusicCaps* [47], *LP-MusicCaps MSD* [48]). For the text from *ChatGPT*, we used GPT-3.5 to generate 31K text based on in-context learning. Specifically, we provide example captions from Clotho or AudioCaps and ask ChatGPT to generate similarly styled audio descriptions based on the examples[6]. This enables the model to generate a large number of audio captions automatically, with a simple and fully automated process. The text data in *FreeSound* comes from the subset of *WavCaps*, collected through an online collaborative sound-sharing site [39]. *WavCaps* [11] is a large-scale weakly-labeled audio captioning dataset that collects audio clips and their raw descriptions from web sources and uses ChatGPT to filter and clean noisy descriptions. *COCO Captions* [6] is a

[4]https://chat.openai.com/

[5]https://freesound.org/

[6]The prompt template is shown in Appendix A

TABLE III
THE EXPERIMENTAL RESULTS FOR CROSS-DOMAIN SCENARIOS ON THE AUDIOCAPS AND CLOTHO DATASET

Method	AudioCaps ⟹ Clotho				Clotho ⟹ AudioCaps			
	ROUGE_L	CIDEr	METEOR	SPICE	ROUGE_L	CIDEr	METEOR	SPICE
Fully Supervised Audio Captioning								
Prefix AAC [26] †	27.6	19.2	11.2	7.4	33.0	21.1	14.4	8.3
RECAP [29] †	27.6	19.5	11.0	8.4	28.1	19.1	11.2	13.6
ACT [23] ⋆	26.1 ± 0.44	13.4 ± 0.68	10.2 ± 0.25	5.5 ± 0.39	35.2 ± 0.22	23.7 ± 0.87	16.4 ± 0.17	10.7 ± 0.31
MAAC [24] ⋆	24.8 ± 0.83	16.4 ± 1.28	10.3 ± 0.35	5.8 ± 0.10	35.9 ± 0.20	25.4 ± 0.45	17.1 ± 0.23	10.9 ± 0.18
Xu *et al.* [25] ⋆	**29.2** ± 0.04	**22.8** ± 0.51	**12.8** ± 0.07	**8.5** ± 0.22	35.8 ± 0.29	25.6 ± 0.85	16.7 ± 0.30	11.1 ± 0.20
Ours-FS	29.1 ± 0.24	22.4 ± 0.65	**12.8** ± 0.12	**8.5** ± 0.09	**36.1** ± 0.23	**30.9** ± 0.72	**18.0** ± 0.22	**12.6** ± 0.13
Zero-shot Audio Captioning								
NoAudioCaptioning [9] ⋆	26.6 ± 0.45	17.5 ± 2.00	11.1 ± 0.59	7.4 ± 0.60	34.1 ± 1.18	23.3 ± 1.68	16.7 ± 0.36	10.6 ± 0.34
WSAC [10] ⋆	26.6 ± 0.34	20.6 ± 0.31	12.0 ± 0.11	8.2 ± 0.08	35.5 ± 0.15	25.6 ± 0.22	17.3 ± 0.10	12.0 ± 0.08
Ours	**29.8** ± 0.55	**24.8** ± 0.55	**13.2** ± 0.46	**9.3** ± 0.44	**36.1** ± 0.51	**33.8** ± 0.93	**18.0** ± 0.28	**12.3** ± 0.18

† The original results are listed in the paper.
⋆ The results are re-implemented by us.

TABLE IV
THE EXPERIMENTAL RESULTS UNDER TEXTUAL DATA FROM DIFFERENT FIELDS

Data Field	Dataset	Size	Source Dataset ⟹ Clotho				Source Dataset ⟹ AudioCaps			
			ROUGE_L	CIDEr	METEOR	SPICE	ROUGE_L	CIDEr	METEOR	SPICE
Audio Captioning	ChatGPT[4]	31K	25.5 ± 0.31	16.3 ± 0.62	10.6 ± 0.19	6.3 ± 0.11	27.3 ± 0.27	15.5 ± 0.39	11.7 ± 0.17	7.1 ± 0.22
	Freesound[5]	84K	30.4 ± 0.20	22.0 ± 0.84	**12.6** ± 0.20	7.8 ± 0.15	28.6 ± 0.42	22.3 ± 0.94	12.3 ± 0.34	6.7 ± 0.21
	WavCaps [11]	190K	**30.6** ± 0.36	**22.1** ± 0.86	**12.6** ± 0.22	**7.9** ± 0.20	**33.4** ± 1.21	**31.6** ± 1.96	**15.5** ± 0.61	**9.1** ± 0.59
Visual Captioning	COCO Captions [6]	414K	25.9 ± 0.24	10.0 ± 0.55	8.9 ± 0.28	5.1 ± 0.44	27.8 ± 0.53	10.6 ± 1.11	10.6 ± 0.55	6.2 ± 0.69
Music Captioning	MusicCaps [47]	13K	21.1 ± 1.34	6.6 ± 0.90	8.8 ± 0.19	4.5 ± 0.42	20.4 ± 1.84	9.6 ± 0.42	9.8 ± 0.13	6.3 ± 0.89
	LP-MusicCaps MSD [48]	526K	15.9 ± 0.72	0.9 ± 0.10	6.1 ± 0.11	1.0 ± 0.16	15.0 ± 0.56	0.8 ± 0.08	6.2 ± 0.24	0.9 ± 0.13

TABLE V
THE ABLATION EXPERIMENT RESULTS OF DIFFERENT COMPONENTS.

Setting		Components			In-Domain Scenarios				Cross-Domain Scenarios			
		IA	EA	AP	ROUGE_L	CIDEr	METEOR	SPICE	ROUGE_L	CIDEr	METEOR	SPICE
ZS-Base Model	*a).*				29.9 ± 0.82	15.7 ± 0.37	13.1 ± 0.52	7.5 ± 0.62	29.8 ± 1.00	13.8 ± 0.71	14.1 ± 0.62	7.8 ± 0.67
	b).	✓			33.0 ± 0.66	25.9 ± 0.97	15.0 ± 0.30	9.6 ± 0.34	31.9 ± 0.28	18.2 ± 0.79	14.8 ± 0.30	7.8 ± 0.49
	c).		✓		34.7 ± 0.30	30.4 ± 0.89	15.7 ± 0.14	10.5 ± 0.32	33.4 ± 0.21	20.6 ± 0.48	15.3 ± 0.05	9.2 ± 0.15
	d).			✓	35.0 ± 0.44	31.9 ± 0.93	16.1 ± 0.17	10.2 ± 0.18	34.1 ± 0.52	25.9 ± 0.88	16.6 ± 0.37	10.6 ± 0.45
	e).	✓	✓		36.0 ± 0.27	32.6 ± 0.46	16.1 ± 0.19	11.0 ± 0.29	32.9 ± 0.35	19.6 ± 0.76	15.3 ± 0.19	9.2 ± 0.17
ZS-Full Model	*f).*	✓	✓	✓	**37.5** ± 0.17	**40.3** ± 0.47	**17.3** ± 0.17	**11.9** ± 0.19	**36.1** ± 0.51	**33.8** ± 0.93	**18.0** ± 0.28	**12.3** ± 0.18

human-annotated benchmark dataset in visual captioning. For the music captioning corpus, *MusicCaps* [47] is annotated by ten professional musicians and *LP-MusicCaps MSD* [48] is a large language model based pseudo music caption dataset.

From the results shown in Table IV, we have the following findings. 1) The choice of textual data domain significantly influences the model's cross-domain performance. The method trained on text data from *WavCaps* in the audio captioning field achieve acceptable cross-domain performance compared to fully supervised and other zero-shot methods, as shown in Table III. This discrepancy likely arises from the field-specific focus of captions: Visual captions emphasize visual objects, while music captions focus on elements like genre and rhythm. Audio captions, however, prioritize audio events and environmental context, making field alignment a critical factor for effective model generalization. 2) For textual data within the audio captioning field, we observe that increasing the dataset size (from *ChatGPT* to *WavCaps*) leads to consistent cross-domain performance improvements on both Clotho and AudioCaps. This trend suggests that expanding the quantity of field-aligned textual data can significantly enhance the model's ability to generalize across diverse audio datasets. Compared to paired audio-text data and manually annotated audio captions, GPT-generated text data is significantly easier to obtain, as it bypasses the complex and time-consuming process of collecting and annotating paired datasets. Our method utilizes in-context learning to automatically generate large volumes of audio captions, enabling a simpler and more efficient approach to captioning. This shift from relying on scarce paired data to utilizing abundant, scalable textual data could pave the way for a new paradigm in audio captioning, broadening its applicability and advancing research in the field.

TABLE VI
THE NUMBER OF CANDIDATES N IN INSTANCE REPLACEMENT

N	ROUGE_L	CIDEr	METEOR	SPICE
1	36.5 ± 0.17	35.8 ± 0.51	16.6 ± 0.11	11.2 ± 0.15
3	37.0 ± 0.15	36.8 ± 0.36	**16.9** ± 0.10	**11.7** ± 0.18
5	**37.1** ± 0.10	**36.9** ± 0.36	**16.9** ± 0.10	**11.7** ± 0.10
7	37.0 ± 0.18	36.6 ± 0.49	16.8 ± 0.09	11.5 ± 0.14
10	37.2 ± 0.22	36.1 ± 0.70	16.8 ± 0.11	11.4 ± 0.13

C. Ablation Studies

In this section, we conduct ablation experiments for in-domain and cross-domain scenarios by training the models on Clotho. The results are shown in Table V, where 'IA',

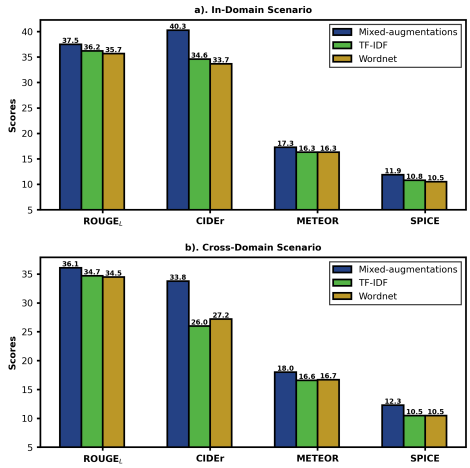


Fig. 4. The experimental results for different text augmentation strategies, where a) shows the results on the Clotho dataset under the in-domain scenario, and b) illustrates the cross-domain scenario. In the cross-domain scenario, where the text corpus from the Clotho training set is used for model training, while the model evaluation is performed on the AudioCaps dataset.

'EA', and 'AP' are abbreviations for the instance augmentation, embedding augmentation, and acoustic-aware prompt, respectively. The *ZS-Base model* does not use any components and its model structure only contains the CLAP encoder, the mapping network, and the language decoder. The audio features are extracted using the CLAP audio encoder and fed into the trained mapping network and language decoder to generate the caption of the given audio during the inference stage. The model structure is shown in Fig. 1 (b). The settings (*b*, *c*, *d*) show that the components we proposed can improve the model performance in all metrics compared to the base model in setting *a*. In particular, the settings (*b*, *c*) show that both instance replacement and embedding augmentation can significantly improve the in-domain performance of the model. These strategies enable the model to learn robust representations by introducing partial perturbations to the CLAP embedding, thus improving the model's generalisation from text-to-text generation to audio-to-text generation and enhancing the performance of zero-shot audio captioning. Acoustic-aware prompts (setting *d*) guides the language decoder through acoustic-aware prompts for audio events, thus enabling the model to achieve better cross-domain generalization performance compared to setting *e*, while maintaining comparable in-domain performance. Interestingly, this suggests that the acoustic-aware strategy is more robust than manipulating the text sentence. Our proposed method, *ZS-Full model*, in the setting *f* achieves significant improvements in all metrics (especially in the CIDEr metric) in both in-domain and cross-domain scenarios, indicating the effectiveness of our proposed model.

In addition, we conduct the ablation experiments to compare our method with other text augmentation strategies used for audio captioning [49]–[51]. The results are shown in Fig. 4. The *Mixed-augmentations* strategy is our proposed method. The *TF-IDF* strategy [52] enhances the text by replacing low-information words with low TF-IDF values while retaining high-information words with high TF-IDF values. The *WordNet* strategy [53], on the other hand, leverages WordNet [54] to replace words in the sentence with synonyms or words of similar meaning to achieve augmentation. It can be observed that our proposed *Mixed-augmentations* strategy consistently outperforms the *TF-IDF* and *WordNet* strategies across various metrics in both in-domain and cross-domain scenarios. We attribute this improvement to the limitations of the *TF-IDF* and *WordNet* strategies, which perform text augmentation at the word level. Such word-level replacements provide limited enhancement to the semantic richness of the text. In contrast, our *Mixed-augmentations* strategy operates directly in the semantic space, thereby leading to more robust augmentation.

TABLE VII
THE VARIANCE σ^2 OF NOISE IN EMBEDDING AUGMENTATION

σ^2	ROUGE$_L$	CIDEr	METEOR	SPICE
1×10^{-4}	34.8 ± 0.72	31.0 ± 1.83	16.0 ± 0.43	10.1 ± 0.43
1×10^{-3}	35.4 ± 0.35	33.7 ± 0.58	16.3 ± 0.13	10.6 ± 0.18
1×10^{-2}	**36.1** ± 0.31	**36.8** ± 0.45	**16.5** ± 0.09	**11.0** ± 0.12
1×10^{-1}	34.2 ± 0.17	32.1 ± 0.53	15.2 ± 0.11	9.7 ± 0.14
1	34.4 ± 0.23	32.5 ± 0.75	15.3 ± 0.15	10.0 ± 0.21

TABLE VIII
THE LENGTH K OF SOFT PROMPT

K	ROUGE$_L$	CIDEr	METEOR	SPICE
1	35.7 ± 0.29	35.5 ± 0.42	16.2 ± 0.11	10.5 ± 0.12
5	36.8 ± 0.04	39.2 ± 0.09	16.9 ± 0.04	11.4 ± 0.00
10	**37.5** ± 0.17	**40.3** ± 0.47	**17.3** ± 0.17	**11.9** ± 0.19
15	37.2 ± 0.11	40.2 ± 0.78	**17.3** ± 0.19	11.6 ± 0.21
20	37.1 ± 0.40	39.3 ± 2.40	17.2 ± 0.35	11.4 ± 0.12

D. Analysis on Hyper-parameters

In the following, we conduct hyper-parameter tuning experiments to investigate and discuss the effects of different hyper-parameters on the model performance. We fix the other hyper-parameters in the full model in each tuning experiment.

1) The number of candidates N in instance replacement: We first show the effect of the number of candidates N in the instance replacement. We select the number of candidates N from values $\{1, 3, 5, 7, 10\}$. The results are shown in Table VI. When N is 5, the model performs better in most metrics. As N continues to increase, the model performance starts to deteriorate since augmented text samples contain texts that are far away from the original text so that the model can learn an accurate "text-to-text" conversion.

2) The variance σ^2 of noise in embedding augmentation: In Table VII, we present the results under different variances. We find that the model performance is sensitive to the variance scale. As the variance increases, the model performance improves progressively, suggesting that appropriate noise applied to the text embedding can significantly enhance the generalization ability of the model and weaken the gap between training stage and inference stage. However, when the variance exceeds 1×10^{-2}, the model performance decreases rapidly due to excessive noise.

3) The length K of soft prompt: We select the number of length K from values $\{1, 5, 10, 15, 20\}$. Table VIII shows the experimental results under different lengths K. We can find

TABLE IX
THE NUMBER OF AUDIO EVENTS M IN ACOUSTIC-AWARE PROMPT

M	ROUGE$_L$	CIDEr	METEOR	SPICE
1	33.8 ± 0.37	27.6 ± 0.98	15.0 ± 0.16	8.4 ± 1.70
2	33.6 ± 0.64	28.3 ± 0.47	15.3 ± 0.22	9.9 ± 0.15
3	33.9 ± 0.69	30.5 ± 1.56	15.7 ± 0.33	10.1 ± 0.38
4	**35.3** ± 0.27	**34.3** ± 0.71	16.1 ± 0.13	10.4 ± 0.21
5	35.2 ± 0.29	34.1 ± 0.67	**16.3** ± 0.18	**10.5** ± 0.21
7	34.8 ± 0.76	32.3 ± 1.22	15.8 ± 0.30	10.3 ± 0.33
10	34.1 ± 0.20	31.5 ± 0.59	15.3 ± 0.27	10.1 ± 0.14

TABLE X
THE RATE β OF EVENT LABEL DROPOUT

β	ROUGE$_L$	CIDEr	METEOR	SPICE
0	36.3 ± 0.64	34.7 ± 0.21	16.9 ± 0.38	11.4 ± 0.20
0.2	36.7 ± 0.24	36.0 ± 0.68	17.1 ± 0.11	11.6 ± 0.25
0.4	37.3 ± 0.12	39.9 ± 0.21	17.2 ± 0.08	11.7 ± 0.18
0.6	**37.5** ± 0.17	**40.3** ± 0.47	**17.3** ± 0.17	**11.7** ± 0.19
0.8	36.7 ± 0.56	37.5 ± 0.74	17.2 ± 0.29	11.6 ± 0.42
1	36.1 ± 0.41	35.4 ± 0.56	16.3 ± 0.19	11.2 ± 0.16

that the best performance is achieved in almost all metrics when K is 10. When K is 1, the inferior results are achieved because of the limited expressiveness of the model.

4) The number of audio events M in acoustic-aware prompt: Table IX presents experimental results using different audio event numbers M. The model performance is the best when we set M to 4 or 5. When M is less than 4, the model performance improves with increasing M due to more acoustic information guidance. However, when M is greater than 5, the performance of the model decreases due to the increase in the irrelevance of the retrieved sound events.

5) The rate β of event label dropout: Table X demonstrates the effect of different dropout rate β on the performance. We can see that the CIDEr score gradually increases as β increases, indicating that dropout can prevent the model from relying heavily on the audio events. When β exceeds 0.6, the model performance decreases as useful audio events information is discarded so the model cannot leverage the guidance.

E. Multilingual Audio Captioning

In addition, since only text is involved in the training stage, we can more easily use advanced language-based tools to investigate the potential applications of our proposed method, such as multilingual audio captioning, multi-styled audio captioning (literary style, children's style, etc.)

For example, when it comes to multilingual captioning systems, we use the Mistral [55] large language model, which is a multilingual pre-trained text generation model with 7 billion parameters[7], to replace the GPT-2 as a language decoder for multilingual audio captioning. We use the *DeepL*[8] to translate the Clotho English text data into different languages (Chinese, French). The additional language token L (e.g., <en>, <fr>) is fed into the language decoder with acoustic-aware prompt H and soft prompt S to generate language-specific audio captions.

We conduct experiments as shown in Table XI, where the model configurations for the *ZS-Base Model* and *ZS-Full Model* are similar to those in the ablation studies (Table V), with the difference being the use of Mistral as the language decoder to generate multilingual audio captioning. Our proposed method, the *ZS-Full Model*, achieves comparable results with the fully supervised method in most metrics and even achieves better results in English compared to the experimental results in Table II. We believe that Mistral has more powerful text generation capabilities compared to GPT-2, and therefore can exploit multimodal semantic information and generate descriptive text more accurately. While the multilingual training data is three times the size of the monolingual data due to translations, the performance improvement is not solely due to the larger dataset. Given the inherent challenges of multilingual audio captioning, we attribute the better results to Mistral's stronger modeling capabilities rather than just the increased data size. In addition, the *ZS-Base Model* still achieves inferior performance in all the metrics compared to our proposed method, the *ZS-Full Model*, which demonstrates that our proposed mixed-augmentation strategy and the auditory-aware retrieval strategy can also improve the generalization performance of zero-shot audio captioning in the multi-lingual scenario.

F. Qualitative Analysis

1) In-domain Audio Captioning: Table XII shows the visualization results for the AudioCaps and Clotho datasets in the in-domain setting, where *red* and *blue* are the sound events objects and their actions behavior, respectively. The last row is the retrieved audio events in the acoustic-aware prompts. We can find that benefiting from the guidance provided by the acoustic-aware prompt and the improvement of model's generalization ability provided by the mixed-augmentations strategy, our proposed zero-shot method does not use any paired audio-text data for training, but can still accurately recognize the audio events and describe the contents of the audio clip during inference. In addition, the event label dropout can mitigate the over-reliance of the model on prompts: in the fourth sample, the retrieved sound events provide irrelevant information ('country', 'field recording', and 'noise'), but the model manages to generate accurate descriptions, overcoming the interference of noisy guidance.

2) Cross-domain Audio Captioning: We also present the ground truth captions and the generated captions of our proposed method in the cross-domain setting, shown in Table XIII. We can observe that the training corpus has a tremendous impact on the style of the generated text. For instance, in the second sample, the training set of AudioCaps contains lots of short, generalized text, which results in concise captions. In the third sample, the text generated by ChatGPT results in speculative descriptions "*demanding attention from its passengers*".

3) Multilingual Audio Captioning: Table XIV shows the samples of English, French, and Chinese audio captions generated by our proposed model. Our method can generate

[7]https://huggingface.co/mistralai/Mistral-7B-v0.1
[8]https://www.deepl.com/

TABLE XI
THE IN-DOMAIN EXPERIMENTAL RESULTS ON MULTILINGUAL AUDIO CAPTIONING

Setting	English			French			Chinese		
	ROUGE_L	CIDEr	METEOR	ROUGE_L	CIDEr	METEOR	ROUGE_L	CIDEr	METEOR
Supervised Model	**37.9** ± 0.28	41.8 ± 1.31	**17.6** ± 0.21	**29.8** ± 2.81	**29.3** ± 2.76	13.4 ± 1.22	**28.2** ± 0.94	**20.6** ± 1.58	**14.6** ± 0.33
ZS-Base Model	32.3 ± 0.82	24.4 ± 1.18	14.3 ± 0.51	26.0 ± 0.56	18.5 ± 1.46	12.4 ± 0.35	25.5 ± 1.06	15.7 ± 2.34	13.9 ± 0.36
ZS-Full Model	37.7 ± 0.44	**42.1** ± 0.50	**17.6** ± 0.13	29.6 ± 2.84	28.6 ± 3.25	**13.5** ± 1.29	27.9 ± 0.52	20.4 ± 0.34	14.3 ± 0.19

TABLE XII
THE SAMPLE RESULTS OF THE IN-DOMAIN AUDIO CAPTIONING

Sample	AudioCaps		Clotho	
	YqeSl7YZAfs4.wav	YonBZOH88OYs.wav	t34t trafik[1].wav	Ronda - The Old Shrine - La antigua Ermita.wav
Ground Truth	*faucet running* and a *man speaks*	repeated bursts of *spray*	*car horns honk* in traffic and *people shout* in the background	*birds* are *singing* while *people talk* in the background
Prediction	a *man* is *speaking* and *water* is *running* from a faucet	*spraying* and *hissing*	*cars* are *honking* their horns and *people* are *talking* in the background	*birds* are *chirping* and *people* are *talking* in the background
Audio Events	water tap, faucet, sink (filling or washing), bathtub (filling or washing), male speech, man speaking	spray, hiss, air brake, steam	vehicle horn, car horn, honking, honk, air horn, truck horn, traffic noise, roadway noise	country, bird, field recording, noise

TABLE XIII
THE SAMPLE RESULTS OF THE CROSS-DOMAIN AUDIO CAPTIONING

Sample	Clotho ⟹ AudioCaps	AudioCaps ⟹ Clotho	ChatGPT ⟹ AudioCaps	WavCaps ⟹ Clotho
	YfBYDJWChe5c.wav	Blade Big.wav	YwoadpeAGHUQ.wav	steam train 05.wav
Ground Truth	a *person snoring*	*metal sliding* together such as *swords* or *knives*	an *emergency siren blaring* steadily	a *person talks* on board a *train* while it *rattles* along the tracks
Prediction	a *person* is *snoring*	*clanking* and *clanking*	an *ambulance siren wails* urgently, demanding attention from its passengers	a train is *moving* on a track with a *clickety-clack* sound
Audio Events	snoring, snort, babbling, groan	dishes, pots, and pans, cutlery, silverware, scrape, heavy metal	fire engine, fire truck (siren), emergency vehicle, ambulance (siren)	train, railroad car, train wagon, clickety-clack

TABLE XIV
THE SAMPLE RESULTS OF THE MULTILINGUAL AUDIO CAPTIONING

Sample	enoesque-Thunder and Rain 1.wav	Pencil Writing.wav
Ground Truth	*rain* starts *pouring down* and *thunder makes* a boom	a *person writes* several words on a chalkboard
English	*thunder* is *rumbling* and *rain* is *falling*	a *person* is *writing* on a chalkboard with chalk
French	la *pluie tombe* sur le sol à un rythme régul.	*quelqu'un écrit* sur un tableau
Chinese	大雨倾盆而下	有人在黑板上写字

descriptive text for the corresponding audio in an end-to-end process, regardless of the language, providing a solid basis for applying the multilingual audio captioning method.

VI. CONCLUSION

We have presented a novel zero-shot audio captioning method that does not employ human-labeled audio-text paired data but only uses the text corpus for model training. Our proposed method avoids the reliance on highly costly paired data. To enhance the model's generalization ability during the transition from text-to-text generation to audio-to-text generation and to improve the cross-domain performance of the model, we devise a mixed-augmentation strategy and a retrieval-based acoustic-aware prompt strategy. Extensive experiments were conducted on AudioCaps and Clotho to demonstrate the effectiveness of our proposed method. Our proposed method performs better on most metrics for the in-domain setting than other zero-shot audio captioning methods. In the cross-domain setting, our proposed method outperforms the compared methods in all metrics, including both fully supervised and zero-shot audio captioning methods. Moreover, our proposed method shows the potential of multilingual audio captioning. Experimental results show that our method can generate multilingual descriptive text for input audio in an end-to-end style.

Our proposed method relies on the multimodal modeling abilities of CLAP. Its performance on downstream tasks, including audio captioning, is influenced by CLAP's ability in multimodal modeling [56]. Therefore, improving CLAP's training strategy to enhance its modeling ability is an exciting future direction of research. Other potential directions include exploring the effectiveness of our proposed method in other audio-text multimodal tasks, such as Music Captioning and Audio Question Answering, and studying multilingual and multi-styled methods to promote the democratization of audio captioning.

ACKNOWLEDGEMENT

We sincerely thank the associate editor and the reviewers for their constructive comments and valuable suggestions, which have greatly improved the quality of this manuscript. This work was supported by the National Natural Science Foundation of China (Grant 62225601, U23B2052) , in part by the Beijing Natural Science Foundation Project No. L242025, and in part by the scholarship from China Scholarship Council No. 202306470064. The work was conducted during Y. Zhang's academic visit to Prof W. Wang's research lab at the University of Surrey.

REFERENCES

[1] K. Drossos, S. Adavanne, and T. Virtanen, "Automated audio captioning with recurrent neural networks," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2017, pp. 374–378.

[2] K. Drossos, S. Lipping, and T. Virtanen, "Clotho: An audio captioning dataset," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 736–740.

[3] C. D. Kim, B. Kim, H. Lee, and G. Kim, "AudioCaps: Generating captions for audios in the wild," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 119–132.

[4] S. Lipping, K. Drossos, and T. Virtanen, "Crowdsourcing a dataset of audio captions," in *Proceedings of Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, 2019, p. 139.

[5] X. Xu, Z. Xie, M. Wu, and K. Yu, "Beyond the status quo: A contemporary survey of advances and challenges in audio captioning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

[6] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick, "Microsoft COCO Captions: Data collection and evaluation server," *arXiv preprint arXiv:1504.00325*, 2015.

[7] T. Shaharabany, A. Shaulov, and L. Wolf, "Zero-shot audio captioning via audibility guidance," *arXiv preprint arXiv:2309.03884*, 2023.

[8] L. Salewski, S. Fauth, A. Koepke, and Z. Akata, "Zero-shot audio captioning with audio-language model guidance and audio context keywords," *arXiv preprint arXiv:2311.08396*, 2023.

[9] S. Deshmukh, B. Elizalde, D. Emmanouilidou, B. Raj, R. Singh, and H. Wang, "Training audio captioning models without audio," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 371–375.

[10] T. Kouzelis and V. Katsouros, "Weakly-supervised automated audio captioning via text only training," *Proceedings of Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, pp. 81–85, 2023.

[11] X. Mei, C. Meng, H. Liu, Q. Kong, T. Ko, C. Zhao, M. D. Plumbley, Y. Zou, and W. Wang, "WavCaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 3339–3354, 2024.

[12] V. W. Liang, Y. Zhang, Y. Kwon, S. Yeung, and J. Y. Zou, "Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning," *Advances in Neural Information Processing Systems*, vol. 35, pp. 17 612–17 625, 2022.

[13] X. Li and P. Liang, "Prefix-tuning: Optimizing continuous prompts for generation," in *Association for Computational Linguistics*, 2021.

[14] G. Qin and J. Eisner, "Learning how to ask: Querying LMs with mixtures of soft prompts," in *Association for Computational Linguistics*, 2021.

[15] 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 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

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

[17] B. Elizalde, S. Deshmukh, and H. Wang, "Natural language supervision for general-purpose audio representations," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 336–340.

[18] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "PANNs: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.

[19] 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 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 646–650.

[20] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *Proceedings of the North American Chapter of the Association for Computational Linguistics*, pp. 4171–4186, 2019.

[21] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "ROBERTA: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[22] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[23] X. Mei, X. Liu, Q. Huang, M. D. Plumbley, and W. Wang, "Audio captioning transformer," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, 2021.

[24] Z. Ye, H. Wang, D. Yang, and Y. Zou, "Improving the performance of automated audio captioning via integrating the acoustic and semantic information," *Proceedings of Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, pp. 40–44, 2021.

[25] X. Xu, Z. Xie, M. Wu, and K. Yu, "The SJTU system for DCASE2022 challenge task 6: Audio captioning with audio-text retrieval pre-training," *DCASE 2022 Challenge, Tech. Rep.*, 2022.

[26] M. Kim, K. Sung-Bin, and T.-H. Oh, "Prefix tuning for automated audio captioning," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

[27] A. Koh, X. Fuzhao, and C. E. Siong, "Automated audio captioning using transfer learning and reconstruction latent space similarity regularization," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 7722–7726.

[28] Y. Zhang, H. Yu, R. Du, Z.-H. Tan, W. Wang, Z. Ma, and Y. Dong, "ACTUAL: Audio captioning with caption feature space regularization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

[29] S. Ghosh, S. Kumar, C. K. R. Evuru, R. Duraiswami, and D. Manocha, "RECAP: retrieval-augmented audio captioning," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 1161–1165.

[30] 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," *Proceedings of the International Conference on Machine Learning*, 2024.

[31] D. Nukrai, R. Mokady, and A. Globerson, "Text-only training for image captioning using noise-injected clip," *Findings of the Association for Computational Linguistics: EMNLP*, pp. 4055–4063, 2022.

[32] W. Li, L. Zhu, L. Wen, and Y. Yang, "Decap: Decoding clip latents for zero-shot captioning via text-only training," *The International Conference on Learning Representations (ICLR)*, 2023.

[33] M. J. Mirza, L. Karlinsky, W. Lin, H. Possegger, M. Kozinski, R. Feris, and H. Bischof, "LaFTer: Label-free tuning of zero-shot classifier using language and unlabeled image collections," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[34] Y. Meng, S. Yang, X. Hu, R. Zhao, L. Li, Z. Shi, and Z. Zou, "Zero-shot text-driven physically interpretable face editing," *arXiv preprint arXiv:2308.05976*, 2023.

[35] Z. Duan, Y. Ding, C. Gou, Z. Zhou, E. Smith, and L. Liu, "Ezigen: Enhancing zero-shot subject-driven image generation with precise subject encoding and decoupled guidance," *arXiv preprint arXiv:2409.08091*, 2024.

[36] R. Girdhar, A. El-Nouby, Z. Liu, M. Singh, K. V. Alwala, A. Joulin, and I. Misra, "ImageBind: One embedding space to bind them all," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 15 180–15 190.

[37] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "AudioSet: An ontology and human-labeled dataset for audio events," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 776–780.

[38] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[39] F. Font, G. Roma, and X. Serra, "Freesound technical demo," in *Proceedings of the 21st ACM International Conference on Multimedia*, 2013.

[40] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.

[41] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*, 2004, pp. 74–81.

[42] S. Banerjee and A. Lavie, "METEOR: An automatic metric for mt evaluation with improved correlation with human judgments," in *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 2005, pp. 65–72.

[43] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "CIDER: Consensus-based image description evaluation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4566–4575.

[44] P. Anderson, B. Fernando, M. Johnson, and S. Gould, "SPICE: Semantic propositional image caption evaluation," in *European Conference on Computer Vision*. Springer, 2016, pp. 382–398.

[45] S. Liu, Z. Zhu, N. Ye, S. Guadarrama, and K. Murphy, "Improved image captioning via policy gradient optimization of SPIDEr," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 873–881.

[46] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proceedings of the 25th International Conference on Machine learning*, 2008, pp. 1096–1103.

[47] A. Agostinelli, T. I. Denk, Z. Borsos, J. Engel, M. Verzetti, A. Caillon, Q. Huang, A. Jansen, A. Roberts, M. Tagliasacchi *et al.*, "MusicLM: Generating music from text," *arXiv preprint arXiv:2301.11325*, 2023.

[48] S. Doh, K. Choi, J. Lee, and J. Nam, "LP-MusicCaps: LLM-based pseudo music captioning," *The International Society for Music Information Retrieval (ISMIR)*, 2023.

[49] Y. Koizumi, D. Takeuchi, Y. Ohishi, N. Harada, and K. Kashino, "The NTT DCASE2020 challenge task 6 system: Automated audio captioning with keywords and sentence length estimation," *Proceedings of Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, 2020.

[50] D. Takeuchi, Y. Koizumi, Y. Ohishi, N. Harada, and K. Kashino, "Effects of word-frequency based pre-and post-processings for audio captioning," *Proceedings of Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, 2020.

[51] J.-H. Cho, Y.-A. Park, J. Kim, and J.-H. Chang, "Hyu submission for the dcase 2023 task 6a: Automated audio captioning model using almixgen and synonyms substitution," in *Proceedings of Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, 2023.

[52] Q. Xie, Z. Dai, E. Hovy, T. Luong, and Q. Le, "Unsupervised data augmentation for consistency training," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6256–6268, 2020.

[53] J. Wei and K. Zou, "EDA: Easy data augmentation techniques for boosting performance on text classification tasks," *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 6382–6388, 2019.

[54] G. A. Miller, "Wordnet: a lexical database for english," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[55] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. d. l. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier *et al.*, "Mistral 7b," *arXiv preprint arXiv:2310.06825*, 2023.

[56] Y. Yuan, Z. Chen, X. Liu, H. Liu, X. Xu, D. Jia, Y. Chen, M. D. Plumbley, and W. Wang, "T-CLAP: Temporal-enhanced contrastive language-audio pretraining," in *Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing*. IEEE, 2024.

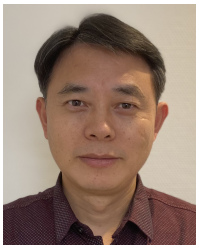
Yiming Zhang received his B.E. degree from Beijing University of Posts and Telecommunications (BUPT), China, in 2018, and the M.S. degree from Beijing University of Posts and Telecommunications (BUPT), China, in 2021. Currently, he is pursuing the Ph.D. degree. His research interests include audio captioning and audio generation.

Xuenan Xu received his B.S. degree from Shanghai Jiao Tong University in 2019. He is currently working towards his Ph.D. degree with the Department of Computer Science and Engineering, Shanghai Jiao Tong University. His main research interests include audio understanding and generation and multi-modal learning.

Ruoyi Du received his B.E. degree in telecommunication with management from Beijing University of Posts and Telecommunications (BUPT), China, in 2020, where he is currently pursuing the Ph.D. degree. His research interests include pattern recognition and computer vision.

Haohe Liu received his B.Eng. degree in Computer Science from Northwestern Polytechnical University, Xi'an, China, in 2020. He is currently a final-year PhD student at the Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey, UK. His research focuses on audio generative modeling, speech quality enhancement, source separation, and audio recognition. Haohe has authored or co-authored over 40 research publications, appearing in top-tier journals and conferences such as IEEE TPAMI, IEEE/ACM TASLP, IEEE JSTSP, ICML, AAAI, NeurIPS, INTERSPEECH, and ICASSP. Haohe is widely recognized as the primary author of AudioLDM and AudioLDM2, alongside other notable projects including VoiceFixer, AudioSR, SemantiCodec, MusicLDM, and NaturalSpeech. Haohe was honored with the 2024 Postgraduate Researcher of the Year Award (CSEE, University of Surrey).

Yuan Dong is currently a Professor at Beijing University of Posts and Telecommunications. He received the Ph.D. degree from the Shanghai Jiao Tong University, China, in 1999. His research interests include text analysis, machine translation and natural language processing.



Zheng-Hua Tan (M'00–SM'06) is a Professor and a co-head of the Centre for Acoustic Signal Processing Research (CASPR) at Aalborg University, Denmark. He is also a Co-Lead of the Pioneer Centre for AI, Denmark. He was a Visiting Scientist at MIT, USA, an Associate Professor at SJTU, China, and a postdoctoral fellow at KAIST, Korea. His research interests include machine learning, deep learning, noise-robust speech processing, and multi-modal signal processing. He has (co-)authored over 280 refereed publications. He was the Chair of the IEEE Signal Processing Society Machine Learning for Signal Processing Technical Committee. He is the Lead Editor for IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING Special Series on AI in Signal and Data Science. He was an Associate Editor for the IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH AND LANGUAGE PROCESSING. He has served as an Editorial Board Member for several journals including Computer Speech and Language. He is the General Chair for ICASSP 2029 and was a TPC Vice-Chair for ICASSP 2024, the General Chair for IEEE MLSP 2018 and a TPC Co-Chair for IEEE SLT 2016.

Wenwu Wang (M'02-SM'11) was born in Anhui, China. He received the B.Sc., M.E., and the Ph.D. degrees, all in the field of automation, from Harbin Engineering University, China, in 1997, 2000, and 2002, respectively. He then worked with King's College London, Cardiff University, Tao Group Ltd. (now Antix Labs Ltd.), and Creative Labs, before joining University of Surrey, U.K., in May 2007, where he is currently a Professor in Signal Processing and Machine Learning, and Associate Head in External Engagement, School of Computer Science and Electronic Engineering, University of Surrey, UK. He is also an AI Fellow at the Surrey Institute for People Centred Artificial Intelligence. His current research interests include signal processing, machine learning and perception, artificial intelligence, machine audition (listening), and statistical anomaly detection. He has (co)-authored over 400 papers in these areas. He has been recognized as a (co-)author or (co)-recipient of more than 15 awards, including the 2022 IEEE Signal Processing Society Young Author Best Paper Award, ICAUS 2021 Best Paper Award, DCASE 2020, 2023 and 2024 Judge's Award, DCASE 2019 and 2020 Reproducible System Award, and LVA/ICA 2018 Best Student Paper Award. He is an Associate Editor (2024-2026) for IEEE Transactions on Multimedia. He was a Senior Area Editor (2019-2023) and Associate Editor (2014-2018) for IEEE Transactions on Signal Processing, and an Associate Editor (2020-2025) for IEEE/ACM Transactions on Audio Speech and Language Processing. He was the elected Chair (2023-2024) of IEEE Signal Processing Society (SPS) Machine Learning for Signal Processing (MLSP) Technical Committee, and a Board Member (2023-2024) of IEEE SPS Technical Directions Board. He is currently the elected Chair (2025-2027) of the EURASIP Technical Area Committee on Acoustic Speech and Music Signal Processing, and an elected Member (2021-2026) of the IEEE SPS Signal Processing Theory and Methods Technical Committee. He was on the organization committee of IEEE ICASSP 2019 and 2024, INTERSPEECH 2022, IEEE MLSP 2013 and 2024, and IEEE SSP 2009. He is a Technical Program Co-Chair of IEEE MLSP 2025. He has been a keynote or plenary speaker at more than 20 international conferences and workshops.

Zhanyu Ma is currently a Professor at Beijing University of Posts and Telecommunications, Beijing, China, since 2019. He received the Ph.D. degree in electrical engineering from KTH Royal Institute of Technology, Sweden, in 2011. From 2012 to 2013, he was a Postdoctoral Research Fellow with the School of Electrical Engineering, KTH. He has been an Associate Professor with the Beijing University of Posts and Telecommunications, Beijing, China, from 2014 to 2019. His research interests include pattern recognition and machine learning fundamentals with a focus on applications in computer vision, multimedia signal processing. He is a Senior Member of IEEE.

APPENDIX A
PROMPT TEMPLATES FOR IN-CONTEXT LEARNING

In this appendix, we describe the prompt template used to generate audio captions through in-context learning. The purpose of the template is to instruct the language model to generate high-quality audio descriptions based on a few examples provided in the prompt.

TABLE XV
THE PROMPT TEMPLATE FOR AUDIO CAPTION

Prompt Template
Generate the sentences describing the content of the audio.
Each sentence should be 25 words or less and focus solely on the audio aspect.
Do not include words describing visual objects, such as size, shape, color, etc.
Each sentence should describe one or several audio events.
The sentences should be similar in style and content to the following examples.
Examples:{***Example Captions***}
Output Caption:

Table XV presents the prompt template used to generate audio captions, where ***Example Captions*** are randomly selected audio captions samples from AudioCaps or Clotho. The entire process of generating audio captions does not require paired audio-text data, nor does it involve complex data preprocessing or post-processing steps.