T-CLAP: TEMPORAL-ENHANCED CONTRASTIVE LANGUAGE-AUDIO PRETRAINING

Yi Yuan[1]*,Zhuo Chen*[2]*,Xubo Liu*[1]*,Haohe Liu*[1]*,Xuenan Xu*[3]*,Dongya Jia*[2]*,Yuanzhe Chen*[2]
Mark D. Plumbley[1]*,Wenwu Wang*[1]

[1] University of Surrey, United Kingdom [2] ByteDance, China
[3] Shanghai Jiao Tong University, China

ABSTRACT

Contrastive language-audio pretraining (CLAP) has been developed to align the representations of audio and language, achieving remarkable performance in retrieval and classification tasks. However, current CLAP struggles to capture temporal information within audio and text features, presenting substantial limitations for tasks such as audio retrieval and generation. To address this gap, we introduce T-CLAP, a temporal-enhanced CLAP model. We use Large Language Models (LLMs) and mixed-up strategies to generate temporal-contrastive captions for audio clips from extensive audio-text datasets. Subsequently, a new temporal-focused contrastive loss is designed to fine-tune the CLAP model by incorporating these synthetic data. We conduct comprehensive experiments and analysis in multiple downstream tasks. T-CLAP shows improved capability in capturing the temporal relationship of sound events and outperforms state-of-the-art models by a significant margin. Our code and models will be released soon.

Index Terms— Contrastive language-audio pretraining, audio retrieval, zero-shot classification, multi-modal learning

1. INTRODUCTION

Contrastive pretraining has received wide attention in the field of multi-modal learning, such as contrastive language-image pretraining (CLIP) [1]. By training the models with a large number of paired samples, such as text-image pairs, these models align the latent representations across the modalities, showing excellent performance in various language-related downstream tasks, including classification [2, 3], retrieval [4, 5], captioning [6, 7], and generation [8, 9, 10]. Similar to CLIP, the recently proposed CLAP model [11] applies an audio encoder and a text encoder to learn the correspondence between text description and audio signal in a shared latent space, setting state-of-the-art performance for audio retrieval tasks. Recently, the ability of CLAP to capture rich auditory features through both audio and text inputs has become a key component in text-to-audio models [12, 13].

However, experiments indicate that current contrastive models exhibit weak modelling capabilities of temporal information [14]. For example, when processing an audio clip that features a sequence of events, such as "*A dog barking followed by a man speaking*", CLAP often generates embeddings that do not distinguish between this sequence and the sequence that occurs in reverse order, i.e. "*Man speaks followed by dog barks*". We hypothesize that the existing deficiency is caused by several factors. Firstly, previous research [15] indicates that transformer-based encoders, such as those applied in CLAP models, are not sensitive to temporal information. Secondly, the existing datasets are deficient in samples that are similar in auditory features but different in sequential order [16]. This scarcity affects the ability of CLAP to recognize the temporal features effectively. In addition, part of the training data used to develop the CLAP models are captions augmented from keywords [11], which lack sequential features that could potentially drive the model towards exploiting temporal information. These issues limit the model's ability to parse and represent the correct order of audio events in multi-event scenarios. As a result, its performance in downstream tasks that involve temporal ordering may be poor. For example, audio retrieval systems built on the CLAP models may fail to correctly retrieve an audio clip in which the order of the audio events is specified by the input text. Similarly, audio generation models built on CLAP models may generate audio clips with events in the wrong order. These examples demonstrate the strong demand for improved CLAP models that can capture temporal information (e.g., temporal order of audio events) within the audio and text data.

In this work, we propose a straightforward but novel approach, T-CLAP, to enhance the temporal understanding of CLAP models. In detail, several approaches are designed to generate samples with temporal contrastive information (i.e. audio events with different sequential order). We first create samples by mixing up labels and audio. Secondly, we take original texts as positive captions and leverage Large Language Models (LLMs) to generate negative captions for audio clips in the large-scale audio-text paired datasets. Then a temporal-focused loss is proposed to fine-tune the CLAP model with these additional text data. We further present a

new task called T-Classify to assess the temporal information captured within CLAP embeddings. T-Classify is designed to examine the ability of the CLAP models to identify and retrieve audio and text data with accurate temporal information.

Our experimental results show that T-CLAP outperforms baseline CLAP models and achieves the state-of-the-art in T-Classify, audio-text retrieval and zero-shot classification tasks. To further evaluate the representation of T-CLAP, we conduct experiment generation tasks (e,g. text-to-audio generation). We use different CLAP models as the text encoder for AudioLDM [12], a state-of-the-art latent diffusion-based generation model. Both objective and subjective evaluation results demonstrate that T-CLAP improves the performance of AudioLDM in generating sound events aligned with the temporal relationships described in the text captions.

The paper is organized as follows. Section 2 introduces the T-CLAP model, including the structure of the CLAP model, the proposed data processing pipelines and the loss functions. Section 3 presents the experimental settings and provides the evaluation metrics of the model. Section 4 discusses the results and the conclusion is drawn in Section 5.

2. TEMPORAL-ENHANCED CLAP

2.1. Problem Formulation

Let $G = \{(a_i, t_i)\}_{i=1}^{N}$ be a dataset with N audio-caption samples, where the audio clip a_i and its associated audio caption t_i form an audio-text pair indexed by i. The CLAP utilizes an audio encoder $f_{\text{audio}}(\cdot)$ and a text encoder $f_{\text{text}}(\cdot)$ to extract the feature of audio clips and their captions, and then projects them into a shared dimension D, i.e., $f_{\text{text}}(a_i) \in R^D$ and $f_{\text{audio}}(t_i) \in R^D$. Using cosine similarity, the similarity of each audio and text embedding can be measured as:

$$s_{ij} = \frac{f_{\text{text}}(a_i) \cdot f_{\text{audio}}(t_j)}{\left\| f_{\text{text}}(a_i) \right\|_2 \cdot \left\| f_{\text{audio}}(t_j) \right\|_2} \quad (1)$$

where the CLAP is trained to increase the similarity score of each audio with its matched text s_{ii}, while decreasing similarity with other text s_{ij} $(i \neq j)$.As discussed in Section 1, original CLAP models suffer from data limitations and encoder insensitivity in temporal features, resulting in insufficient performance in audio event order understanding and generation.

2.2. Model

Our goal is to train an enhanced CLAP that performs better on capturing temporal features, i.e., presenting the text embedding with correct sequential information. We leverage transfer learning and utilize a pretrained CLAP [11] to learn the feature with temporal information. In detail, we employ HTSAT [17] as the audio encoder and RoBERTa [18] as the text encoder. The multi-layer perceptron (MLP) layers and activation functions are then applied to project audio and text embedding into the same dimension for contrastive learning.

2.3. Training Set

To enhance the model in capturing temporal information, we introduce two novel approaches to generate temporally contrastive captions (e.g., captions describing audio events in a wrong order). We first generate audio clips with different audio events by mixing and concatenating audio clips and creating their corresponding text descriptions with matching temporal ordering phrases.

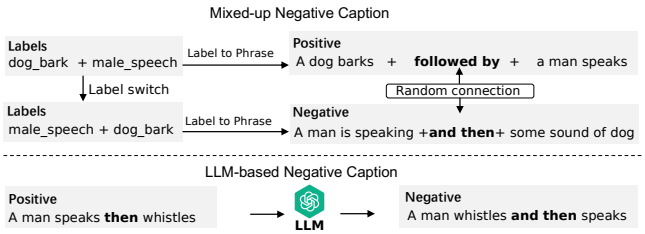

Fig. 1. Pipelines of generating the negative captions.

As shown in Figure 1, labels of each audio are converted into short captions connected with phrases, such as *"followed by", "and then"*, to represent the temporal ordering of audio events. A negative caption is then generated by switching the order of the description for each audio event within each sentence.

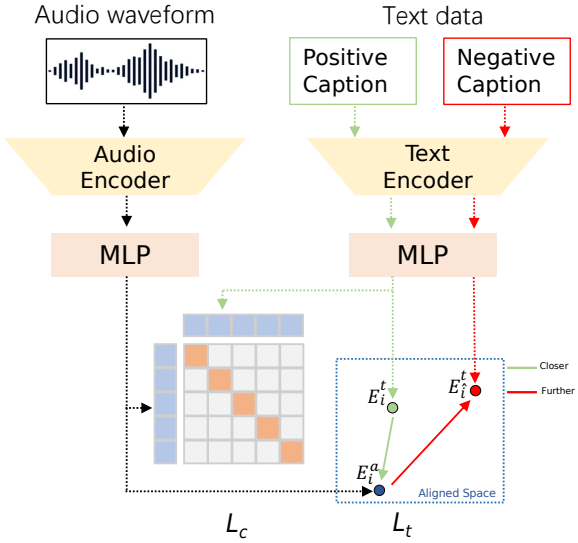

Fig. 2. Pipeline for training T-CLAP, with original contrastive loss L_c on the left and proposed temporal-focused loss L_t on the right.

Secondly, we apply LLMs, e.g. ChatGPT, to generate negative captions of existing large-scale audio-caption datasets. The LLMs are guided to identify the descriptions for audio events in each sentence and generate the text with the same audio events in a different order. We have observed that this approach provides more diversity of temporal-contrastive

samples than artificially generated data (e.g., mixed-up audio and caption), which cannot cover various conditions and sentence combinations.

Nevertheless, we found that audio captions generated by LLMs tend to exhibit specific characteristics. For example, a substantial number of negative captions consist of audio events connected with words like "first" or structures like " , " to represent temporal information, leading the model to overfit on these specific conditions. Hence, we combine both mixed-up datasets and LLM-generated datasets for training, where mixed-up datasets provide straightforward contrastive samples and LLM-generated datasets provide comprehensive and diverse samples. Statistic details of the datasets are discussed in Section 3.

2.4. Loss Functions

The key idea is to train the model to embed audio and text features with temporal information. This means that an audio caption describing the correct order of audio events results in a higher similarity score than a caption describing audio events in a wrong (or different) order. Inspired by [24], we introduce a temporal-focused loss that guides the model to minimize the distance between target audio and positive caption embedding within the aligned latent space, while pushing it away from the embedding presented in the negative order. To effectively enhance our model and maintain the retrieval capabilities, we use a combination of two loss functions for fine-tuning CLAP:

$$L_{\text{train}} = L_c + \lambda_l L_t \tag{2}$$

where L_c is the contrastive loss that maintains the capability of the CLAP model, while L_t is the newly introduced temporal-focused loss for guiding the model with temporal information, and λ_l is a weight to balance the two losses. As shown in Figure 2, for each audio-text sample i, the model first computes the aligned-presentation of audio embedding E_i^a, the positive text embedding E_i^t and negative text embedding E_i^t. The term L_t then works as a single-contrastive-loss to reinforce the similarity between audio and positive text, while minimizing it with negative text, shown as:

$$L_t = -\sum_{i=1}^{N} \log \frac{\exp(E_i^a \cdot E_i^t)}{\exp(E_i^a \cdot E_i^t) + \exp(E_i^a \cdot E_i^t)} \tag{3}$$

3. EXPERIMENTS

3.1. Dataset

Training set. In this paper, we first utilize ESC-50 [25], which comprises 50 distinct categories of sounds with each audio sample lasting 5 seconds, as the supplementary database for mix-up negative captions (ESC-mixed-up). Applying the dataset generation pipeline introduced in Section 2.3, the ESC-mixed-up dataset consists of $50,000$ 10-second audio clips. Each audio clip is composed of distinct audio events and matched with one positive and one negative caption. In addition, we employ LLMs to generate the negative texts corresponding to captions in AudioCaps [26] and Clotho [27]. Both AudioCaps and Clotho are audio datasets with human-annotated captions. Specifically, AudioCaps includes about $50,000$ audio samples, each lasting 10 seconds, and Clotho has about $5,000$ audio clips of varying duration (ranging from 15 to 30 seconds). In addition, Auto-ACD [28] is applied as the large training set to maintain the retrieval capability of T-CLAP. Auto-ACD contains 1.9M audio samples with paired captions generated through LLMs on both visual and audio information. Together we collected 2.04M audio-caption pairs for training.

Testing set. We evaluate the performance of the proposed method on several downstream tasks, including retrieval, zero-shot classification, and text-to-audio generation. For the retrieval tasks, we follow the baseline models and evaluate AudioCaps and Clotho. For temporal-focused retrieval (T-Classify), we apply AudioCaps, Clotho and ESC-mixed-up datasets with temporal-negative samples. For zero-shot classification tasks, we use ESC-50, Urbansound8k [29] and VGGSound [30] for evaluations. Urbansound8K has $8,000$ samples and is labelled into ten different groups, while VGGSound is a larger dataset with more than 300 classes.

3.2. Implementation

We fine-tune the CLAP [11] with our proposed loss function L_{train} introduced in Section 2.4. Specifically, the temporal loss L_t is only applied on a combination of AudioCaps, Clotho, and ESC-mixed-up datasets, where each sample is combined with a negative caption. We take the Auto-ACD with only positive captions as the primary database for the contrastive loss L_c and apply a ratio of $1:4$ to add the other three datasets within each batch. The weight between two loss functions λ_l is set to 0.5. All the models are trained over 30K steps with a batch size of 512. The learning rate is set to 1×10^{-4} with a linear warm-up for the first $10,000$ steps.

3.3. Evaluation

Four different tasks are applied to evaluate the performance of T-CLAP on different tasks. First, we follow the baseline models on retrieval tasks, including text-to-audio (T2A) and audio-to-text (A2T) retrieval tasks.

In addition, we evaluate the performance of the T-CLAP model on zero-shot audio classification across three datasets, including ESC-50 [25], Urbansound8K [29], and VGGSound [30]. For ESC-50, we follow the original test-split defined within the CLAP [11] experiments, while we examine the performance on the entire Urbansound8K dataset and the official

Table 1. The retrieval performance between different systems, CLAP_m and CLAP_l are models trained by Microsoft [19, 20] and LAION [11], using different structures and datasets respectively [11]. For the training set, "AC" and "CL" are short for AudioCaps and Clotho datasets, "WT5K", "LA", "ES" and "AU" refer to Wavtext5K, LAION-Audio-630k, ESC-mixed-up and Auto-ACD respectively. T-CLAP is marked with † as a fine-tuned model from CLAP_l

Model	Training Set	AudioCaps						Clotho					
		Text-to-Audio			Audio-to-Text			Text-to-Audio			Audio-to-Text		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
MMT [21]	AC or CL	36.1	72.0	84.5	39.6	76.8	86.7	6.7	21.6	33.2	7.0	22.7	34.6
ML-ACT [22]	AC or CL	33.9	69.7	82.6	39.4	72.0	83.9	14.4	36.6	49.9	16.2	37.6	50.2
TAP [23]	AC or CL	36.1	72.0	85.2	41.3	75.5	86.1	16.2	39.2	50.8	17.6	39.6	51.4
CLAP [19]	AC+CL+WT5K	34.6	70.2	82.0	41.9	73.1	84.6	16.7	41.1	**54.1**	20.0	44.9	58.7
CLAP_m [20]	4.6M-Audio	33.5	70.4	80.2	47.8	80.2	90.7	16.2	39.6	51.4	**23.6**	46.7	**60.3**
CLAP_l [11]	AC+CL+LA	34.2	71.1	84.1	43.1	79.5	90.1	15.3	38.4	51.2	20.8	**51.2**	60.0
T-CLAP†	AC+CL+ES+AU	**39.7**	**74.6**	**86.9**	**49.8**	**82.5**	**91.9**	**17.3**	**39.9**	53.6	21.8	44.9	57.4

testing set of VGGSound, respectively. Similar to the baseline approach, we evaluate the performance as a T2A retrieval task with a text prompt formatted as " *A sound of* ***Label*** " according to the ***Label*** of each audio sample. Notably, we have excluded all the testing samples from the ESC-50 dataset in our mixed-up dataset and VGGSound audio samples in Auto-ACD to prevent any overlap and premature access of data during the training stages.

We also introduce a new task, T-Classify, to measure the ability of temporal feature retrieval. For each test sample i with negative captions (AudioCaps and Clotho testing set), T-Classify evaluates the performance on T2A by calculating the percentage of cases where the model can successfully provide higher cosine similarity on the positive captions and lower on negative captions ($s_{ii} > s_{i\hat{i}}$). Same as for A2T tasks, for each test sample j with negative audio clips (ESC-mixed-up testing set), T-Classify measures the percentage that the model calculates a higher score on positive audios than negative audios ($s_{jj} > s_{\hat{j}j}$). For both two tasks, we calculate the percentage of correct judgments by aggregating the results across the testing sets.

Lastly, we evaluate the effectiveness of temporal-enhanced embedding in text-to-audio generation tasks. Specifically, we train various AudioLDM [12] using different CLAP models as the input encoders. In addition to objective metrics on previous experiments for audio generation, we conduct human evaluations to assess if the AudioLDM, trained with the enhanced CLAP, can generate outputs whose temporal features align with the given text conditions.

4. RESULTS

All the results are evaluated based on an average of each experiment trained three times with random seeds.

Text and audio retrieval. Table 1 presents the performance comparison of T2A and A2T retrieval tasks on benchmark testing sets. Our proposed T-CLAP demonstrates superior performance over the baseline models in most scenarios, particularly across most tasks on the AudioCaps testing set. For tasks performed on the Clotho testing set, the proposed method does not achieve start-of-the-art. This could be attributed to the fact that: (*i*) the majority of the training data consists of 10-second audio clips, while samples with various lengths (e.g., Clotho) only make up a small proportion. This discrepancy may limit the model's ability to generalize across different audio lengths effectively. (*ii*) Clotho features longer audio samples that encompass more complex contextual concepts, which may not be captured by the existing retrieval benchmarks. This observation aligns with findings from previous studies [31], indicating a potential gap in current retrieval metrics to evaluate performance on more contextually rich audio samples.

Table 2. The results of zero-shot classification on three benchmark datasets.

Model	ESC-50	Urbansound8K	VGGSound
Wav2CLIP [32]	41.4	40.4	10.0
AudioClip [33]	69.4	65.3	-
CLAP [19]	82.6	73.2	-
BLAT [34]	80.6	77.3	14.9
CLAP_l [11]	91.0	75.8	23.8
T-CLAP	$\mathbf{96.5}_{\pm 0.25}$	$\mathbf{78.4}_{\pm 0.38}$	$\mathbf{42.8}_{\pm 0.60}$

Zero-shot classification. To examine the generalization and robustness of the CLAP models, Table 2 shows the zero-shot classification results on three benchmark datasets, where T-CLAP outperforms the top five models from previous experiments. Notably, T-CLAP presents a significant gap from the baseline models in the VGGSound [30] dataset, which consists of larger-scale data with over 300 classes of labels. This significant enhancement could be attributed to the varied and comprehensive content offered by the expansive 2M training dataset, which improves the robustness of the model.

Table 3. The retrieval performance on the temporal feature on the T-Classify task. For Text-to-Audio all three datasets are used (AudioCaps, Clotho, and ESC-mixed-up) while only ESC-mixed-up is applied for Audio-to-Text.

Model	T-Classify	
	Text-to-Audio	Audio-to-Text
MMT [21]	53.1	55.6
CLAP_m [20]	45.7	44.1
CLAP_l [11]	56.2	53.2
T-CLAP	**87.2**$_{\pm 0.47}$	**72.0**$_{\pm 1.35}$

Temproal-featured retrieval. We use the T-Classify task mentioned in Section 3 to compare the CLAP models on both Text-to-Audio and Audio-to-Text tasks with temporal information. It is noted that only the ESC-mixed-up testing set is applied for Audio-to-Text tasks as this is the only dataset with negative audio clip samples. The results of the baseline models in Table 3 demonstrate that the previous retrieval models cannot discern the temporal feature, with the overall performance at around 50% (selected randomly). Especially in CLAP_m, the T-classify score does not surpass 50%, indicating that the model prefers to assign higher similarity scores to negative targets than to correct ones. On the other hand, our proposed approach achieves significant improvements in the T-Classify task. With more than 87% of accuracy on Text-to-Audio tasks, T-CLAP shows an enhanced capability of embedding the audio feature with temporal information.

Table 4. The results of AudioLDM trained with different CLAP as text encoder, where all the models are trained under same configuration and steps. MOS_r and MOS_t presents human evaluations on overall quality and temporal information accuracy.

Model	KL ↓	FAD ↓	MOS_r ↑	MOS_t ↑
AudioLDM_M	$2.6_{\pm 0.11}$	$2.4_{\pm 0.17}$	-	-
AudioLDM_L	**2.2**$_{\pm 0.13}$	$2.5_{\pm 0.24}$	3.5	2.8
AudioLDM_T	$2.3_{\pm 0.13}$	**1.8**$_{\pm 0.08}$	**3.7**	**3.8**

Text-to-audio generation. To compare the effectiveness of temporal-enhanced embedding, we train AudioLDM models using different CLAP models as the condition encoder for $500,000$ steps. AudioLDM_M, AudioLDM_L, and AudioLDM_T presents the model trained with CLAP-Microsoft [20], CLAP-LAION [11] and T-CLAP respectively. Table 4 shows that the AudioLDM model employing T-CLAP achieve better results on text-to-audio benchmarks. Nevertheless, when integrated with T-CLAP encoders, AudioLDM achieves significantly higher scores in the Mean Opinion Score (MOS) for the accuracy of temporal information. Both the objective and subjective evaluations show that the enhanced sequential feature within T-CLAP embedding contributes to the improvement of AudioLDM in terms of both the overall quality and the temporal feature correctness.

5. CONCLUSION

We propose a novel method to enhance the CLAP on temporal information. We first design two pipelines to generate the negative captions and a temporal-focused loss to guide the model. In addition, we introduce a new task, T-Classify, to evaluate the performance of identifying the temporal feature. Compared with the baseline models, T-CLAP provides improved performance in modelling temporal features, as well as achieving state-of-the-art results on retrieval and zero-shot classification tasks. Evaluations on AudioLDM trained with different CLAP models show that the improvements of audio and text embedding in T-CLAP can contribute to performance enhancement in downstream tasks. However, this paper mainly focuses on the situation of sounds, which may result in performance reduction on tasks under other audio domains (e.g., speech, music). In addition, all the negative audio samples are generated artificially, which affects the robustness of the temporal capability in audio encoders. Future work includes the extension of this method on more audio content, such as speech and music, as well as on more downstream tasks.

6. REFERENCES

[1] 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*, 2021.

[2] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu, "Conditional prompt learning for vision-language models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[3] Kaiyang Zhou, Jingkang Yang, Chenchange Loy, and Ziwei Liu, "Learning to prompt for vision-language models," *International Journal of Computer Vision*, vol. 130, no. 9, pp. 2337–2348, 2022.

[4] Aneeshan Sain, Ayan Kumar Bhunia, Pinaki Nath Chowdhury, Subhadeep Koley, Tao Xiang, and Yi-Zhe Song, "Clip for all things zero-shot sketch-based image retrieval, fine-grained or not," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[5] Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo, "Effective conditioned and composed image retrieval combining clip-based features," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[6] Ron Mokady, Amir Hertz, and Amit H Bermano, "CLIPcap: CLIP prefix for image captioning," *arXiv:2111.09734*, 2021.

[7] Yoad Tewel, Yoav Shalev, Idan Schwartz, and Lior Wolf, "Zero-shot image-to-text generation for visual-semantic arithmetic," *arXiv:2111.14447*, 2021.

[8] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, et al., "EDIFFI: Text-to-image diffusion models with an ensemble of expert denoisers," *arXiv:2211.01324*, 2022.

[9] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen, "GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models," in *International Conference on Machine Learning*. PMLR, 2022.

[10] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al., "Photorealistic text-to-image diffusion models with deep language understanding," *Advances in Neural Information Processing Systems*, vol. 35, pp. 36479–36494, 2022.

[11] Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo 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*, 2023.

[12] Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D. Plumbley, "AudioLDM: Text-to-Audio generation with latent diffusion models," in *International Conference on Machine Learning*, 2023.

[13] Yi Yuan, Haohe Liu, Xubo Liu, Qiushi Huang, Mark D Plumbley, and Wenwu Wang, "Retrieval-augmented text-to-audio generation," in *International Conference on Acoustics, Speech, and Signal Processing*, 2023.

[14] Zixian Ma, Jerry Hong, Mustafa Omer Gul, Mona Gandhi, Irena Gao, and Ranjay Krishna, "CREPE: Can vision-language foundation models reason compositionally?," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[15] Koustuv Sinha, Robin Jia, Dieuwke Hupkes, Joelle Pineau, Adina Williams, and Douwe Kiela, "Masked language modeling and the distributional hypothesis: Order word matters pretraining for little," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2021.

[16] Zeyu Xie, Xuenan Xu, Mengyue Wu, and Kai Yu, "Enhance Temporal Relations in Audio Captioning with Sound Event Detection," in *Proc. INTERSPEECH 2023*, 2023, pp. 4179–4183.

[17] Ke Chen, Xingjian Du, Bilei Zhu, Zejun Ma, Taylor Berg-Kirkpatrick, and Shlomo Dubnov, "HTS-AT: A hierarchical token-semantic audio transformer for sound classification and detection," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2022.

[18] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov, "Roberta: A robustly optimized bert pretraining approach," in *CoRR*, 2019.

[19] Soham Deshmukh, Benjamin Elizalde, and Huaming Wang, "Audio retrieval with Wavtext5K and CLAP training," in *CoRR*, 2022.

[20] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang, "CLAP learning audio concepts from natural language supervision," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2023.

[21] Andreea Maria Oncescu, A Koepke, João F Henriques, Zeynep Akata, and Samuel Albanie, "Audio retrieval with natural language queries," in *Interspeech*, 2021.

[22] Xinhao Mei, Xubo Liu, Jianyuan Sun, Mark D Plumbley, and Wenwu Wang, "On metric learning for audio-text cross-modal retrieval," in *Interspeech*, 2022.

[23] Yifei Xin, Dongchao Yang, and Yuexian Zou, "Improving text-audio retrieval by text-aware attention pooling and prior matrix revised loss," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2023.

[24] Roni Paiss, Ariel Ephrat, Omer Tov, Shiran Zada, Inbar Mosseri, Michal Irani, and Tali Dekel, "Teaching CLIP to count to ten," *arXiv:2302.12066*, 2023.

[25] Karol J Piczak, "ESC: dataset for environmental sound classification," in *Proceedings of the ACM International Conference on Multimedia*, 2015.

[26] Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim, "AudioCaps: Generating captions for audios in the wild," in *Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 2019.

[27] Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen, "Clotho: An audio captioning dataset," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020.

[28] Luoyi Sun, Xuenan Xu, Mengyue Wu, and Weidi Xie, "A large-scale dataset for audio-language representation learning," *arXiv preprint arXiv:2309.11500*, 2023.

[29] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *International Conference on Multimedia*, 2014.

[30] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman, "VGGSound: A large-scale audio-visual dataset," in *International Conference on Acoustics, Speech, and Signal Processing*, 2020.

[31] Ho Hsiang Wu, Oriol Nieto, Juan Pablo Bello, and Justin Salomon, "Audio-text models do not yet leverage natural language," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2023.

[32] Ho Hsiang Wu, Prem Seetharaman, Kundan Kumar, and Juan Pablo Bello, "Wav2clip: Learning robust audio representations from clip," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2022.

[33] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel, "AudioCLIP: Extending CLIP to image, text and audio," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2022.

[34] Xuenan Xu, Zhiling Zhang, Zelin Zhou, Pingyue Zhang, Zeyu Xie, Mengyue Wu, and Kenny Q Zhu, "Blat: Bootstrapping language-audio pre-training based on audioset tag-guided synthetic data," *arXiv:2303.07902*, 2023.