WavCaps: A ChatGPT-Assisted Weakly-Labelled Audio Captioning Dataset for Audio-Language Multimodal Research

Xinhao Mei, Chutong Meng, Haohe Liu, Qiuqiang Kong,
Tom Ko, Chengqi Zhao, Mark D. Plumbley, Yuexian Zou, Wenwu Wang

Abstract—The advancement of audio-language (AL) multimodal learning tasks has been significant in recent years, yet the limited size of existing audio-language datasets poses challenges for researchers due to the costly and time-consuming collection process. To address this data scarcity issue, we introduce *WavCaps*, the first large-scale weakly-labelled audio captioning dataset, comprising approximately 400k audio clips with paired captions. We sourced audio clips and their raw descriptions from web sources and a sound event detection dataset. However, the online-harvested raw descriptions are highly noisy and unsuitable for direct use in tasks such as automated audio captioning. To overcome this issue, we propose a three-stage processing pipeline for filtering noisy data and generating high-quality captions, where ChatGPT, a large language model, is leveraged to filter and transform raw descriptions automatically. We conduct a comprehensive analysis of the characteristics of WavCaps dataset and evaluate it on multiple downstream audio-language multimodal learning tasks. The systems trained on WavCaps outperform previous state-of-the-art (SOTA) models by a significant margin. Our aspiration is for the WavCaps dataset we have proposed to facilitate research in audio-language multimodal learning and demonstrate the potential of utilizing large language models (LLMs) to enhance academic research. Our dataset and codes are available at https://github.com/XinhaoMei/WavCaps.

Index Terms—Audio-language dataset, multimodal learning, ChatGPT, deep learning.

I. INTRODUCTION

OVER the past decade, the field of machine listening has achieved notable progress, with the aid of deep learning techniques and the availability of audio event datasets [1]–[3]. This has resulted in the development of algorithms that can detect and identify sound events and acoustic scenes [4]–[7]. More recently, there has been a surge of interest in establishing a more profound comprehension of audio content by connecting audio and language. A number of audio-language (AL) multimodal learning tasks have been introduced, such as text-to-audio retrieval [8]–[10], automated audio captioning [11]–[13], audio question answering [14], [15], text-based sound

X. Mei, H. Liu, M. D. Plumbley, and W. Wang are with the Centre for Vision, Speech, and Signal Processing, University of Surrey, Guildford, GU2 7XH, U.K. (E-mail: [x.mei, haohe.liu, m.plumbley, w.wang]@surrey.ac.uk)

C. Meng is with Johns Hopkins University, U.S.A. (E-mail: cmeng9@jhu.edu)

Q. Kong is with The Chinese University of Hong Kong, Hong Kong, China. (E-mail: qqkong@ee.cuhk.edu.hk)

T. Ko, and C. Zhao are with ByteDance, China. (E-mail: [tom.ko, zhaochengqi.d]@bytedance.com)

Y. Zou is with the School of Electronic and Computer Engineering, Peking University, Shenzhen Graduate School, Shenzhen, 518055, China. (E-mail: zouyx@pku.edu.cn)

generation [16]–[19], and text-based sound separation [20]. Research on AL takes a stride in the direction of empowering machines to comprehend audio signals at a human-like level.

While AL research is relatively young, vision-language (VL) multimodal learning [21], the counterpart of AL multimodal learning, has been studied for decades and has contributed to the success of many vision-language applications, such as cross-modal search [22], image and video captioning [23], [24], text-to-image generation [25], [26], and visual question answering [27], [28]. Two main factors have contributed to the significant progress in VL tasks. First, advances in model architectures, especially self-attention-based Transformer models [29], have shown superior performance compared to convolutional neural networks (CNNs) [30] and recurrent neural networks (RNNs) [31], both in computer vision and natural language processing tasks. Second, large-scale VL multimodal pretraining enables models to learn transferable and robust multimodal representations that benefit VL downstream tasks [24], [32], [33]. Pre-training on large-scale (even weakly-labelled) datasets and fine-tuning on specific datasets for downstream tasks has been a prevalent methodology in VL multimodal learning tasks [34], [35].

Inspired by the progress made in the domains of vision and language, Transformer-based models and training strategies have been successfully adapted in modeling audio signals [36]. This has led to substantial improvements in the recognition and detection of sound events. However, the scale of audio-related datasets is still limited, which severely hinders the research in AL multimodal learning. Till now, the largest audio event dataset, AudioSet [1], has about 2M audio clips, and the largest audio captioning dataset, AudioCaps [37], contains only about 50k audio clips. Both of them are orders of magnitude smaller than their counterparts in the vision domain such as ImageNet [38], and MS COCO [39]. The reasons for this are twofold: First, the research within the audio community has received less attention, as compared to the vision community. Second, the collection of audio datasets is a more laborious, costly, and time-consuming process compared to that of visual datasets.

To alleviate the data scarcity problem and advance AL research, we gathered audio clips and their corresponding raw descriptions from multiple sources, drawing inspiration from the collection process of the Conceptual Captions datasets [40], [41]. Our collection encompasses data from three web

platforms—FreeSound[1], BBC Sound Effects[2], and Sound-Bible[3]—as well as from an audio tagging dataset, AudioSet. These comprehensive and varied sources are instrumental to enriching the depth and breadth of our dataset. Nevertheless, the harvested raw descriptions may range from complete sentences to fragmented phrases, lists of keywords or tags, and exhibit a high degree of noise. For instance, some descriptions do not represent the audio content at all, and certain descriptions contain extraneous information unrelated to the audio, such as recording devices, time, and locations. As a result, these unrefined raw descriptions hinder the learning of robust audio-language representations and are ill-suited for direct application in tasks like automated audio captioning. To process harvested data, Conceptual Captions 3M (CC3M) [40] applied a series of complex rules to filter and transform candidate image-caption pairs and it only keeps around 0.2% of originally harvested data. When compared to image-description pairs, the limited number of audio-description pairs available on the web makes it impractical to accept such a high rate of discarding.

To overcome this problem, we propose a three-stage processing pipeline for filtering noisy data and generating high-quality captions. In a pioneering move, we utilize ChatGPT[4], a robust large language model to automate the process, ensuring both efficiency and effectiveness. Initially, a pre-filtering stage is implemented to exclude irrelevant data based on text frequency. Subsequently, ChatGPT is employed to further process the obtained raw descriptions. This involves content-based filtering and transforming the raw descriptions into sentences resembling captions. Lastly, we refine undesirable outputs from the second stage in a post-processing stage. Ultimately, we present the first large-scale, weakly-labelled audio captioning dataset, WavCaps, which contains about 400k audio clips with paired captions.

We classify WavCaps as a weakly-labeled dataset for the following reasons. The term "weakly-labeled dataset" is derived from the domain of weak supervision or weakly supervised learning [42]. This area of study addresses situations where the supervision signal provided for model training is imperfect [42], [43], covering the scenarios such as incomplete supervision, where only a portion of the training data has labels; inexact supervision, where the labels used are not as granular as the task demands; and inaccurate supervision, where the labels are tainted with noise or errors. In WavCaps, captions are derived by filtering and transforming harvested raw descriptions using ChatGPT. They may be incomplete if specific sound events are absent from the raw descriptions or incorrect if the descriptions themselves are erroneous, thus exemplifying the characteristics of inexact and inaccurate supervision in weakly supervised learning.

In comparison to existing audio captioning datasets [37], [44], [45], WavCaps is not only an order of magnitude larger, but also encompasses a wider range of content. We conducted experiments on multiple audio-language multimodal learning tasks to evaluate the impact of the proposed WavCaps dataset and achieved new state-of-the-art results on most tasks, surpassing previous benchmarks by significant margins.

In summary, our work offers three main contributions: (1) the introduction of a large-scale, weakly-labeled audio captioning dataset, WavCaps, for audio-language multimodal learning tasks; (2) the use of ChatGPT to automatically filter and rewrite harvested raw descriptions into caption-like sentences, showcasing its powerful data augmentation capabilities; and (3) extensive experiments conducted on downstream tasks, demonstrating the effectiveness of our proposed WavCaps dataset. Our expectation is that WavCaps will aid in advancing research in audio-language multimodal learning and also serve as a demonstration of how ChatGPT can be utilized to enrich academic research.

The remainder of this paper is organized as follows. Section II introduces related works in vision-language and audio-language areas. Details of the dataset collection and processing steps are described in Section III. Section IV introduces experiments on audio-language multimodal tasks and present the results and analysis. Finally, we conclude this work in SectionV.

II. RELATED WORKS

A. Vision-Language Datasets

In recent years, significant progress has been made in VL learning thanks to the release of large-scale VL datasets. The images or videos in these datasets are typically sourced from online platforms. Based on the annotation methods employed, VL datasets can be classified into two categories: automatically-annotated and human-annotated. Automatically-annotated datasets, such as CC3M [40], Conceptual Captions 12M (CC12M) [41], and ALIGN [34], have millions of image-caption pairs, where image-description pairs are first harvested from the web and processed automatically according to a series of predefined rules, without human intervention. These large-scale datasets are usually used as pre-training datasets to learn multimodal representations for downstream tasks. In contrast, human-annotated datasets, such as COCO Captions [39], NO-Caps [46], and Flickr [47], employ humans to annotate images or videos. The annotation process is expensive and time-consuming. Therefore, human-annotated datasets are limited in size and are generally used as fine-tuning datasets for performance evaluation. In summary, automatically-annotated datasets are larger but often noisy, while human-annotated datasets are smaller but of better quality. Pre-training on large-scale (even noisy) datasets and fine-tuning on small task-specific datasets has been a prevalent methodology in VL multimodal learning [32]–[34].

Similar to automatically-annotated VL datasets [34], [40], [41], we also harvest audio-description pairs from the web, and design a pipeline to process data automatically. However, taking into account the differences in quantity and data characteristics, we did not use complex pre-defined rules as in CC3M [40] to process the harvested data, but employed ChatGPT to filter and rewrite raw descriptions into sentence-like captions.

---

[1]https://freesound.org/

[2]https://sound-effects.bbcrewind.co.uk/

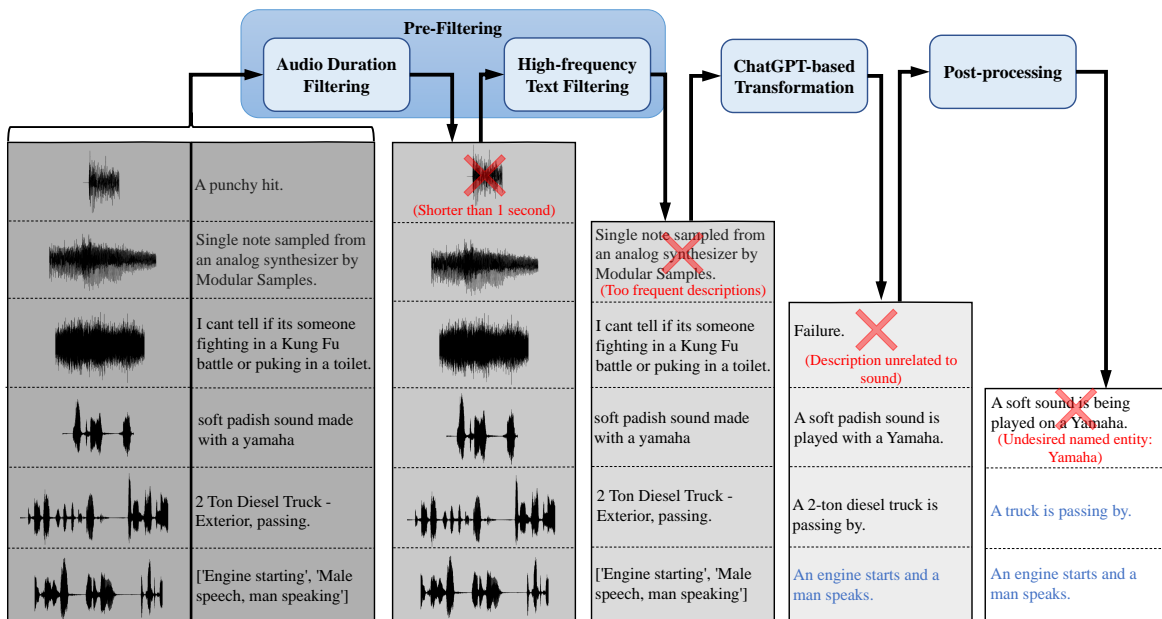[3]https://soundbible.com/

[4]https://openai.com/blog/chatgpt/

Fig. 1: Overview of the three-stage data processing pipeline.

B. Audio-Language Datasets

Compared with flourishing research on VL multimodal learning, research on audio-language multimodal learning is limited due to the lack of AL datasets. Almost all AL tasks, such as automated audio captioning [11], [48], [49], language-based audio retrieval [9], [10], text-to-audio generation [16], [17] and language-queried sound separation [20], rely on human-annotated audio captioning datasets, AudioCaps [37] and Clotho [44], [50]. Collecting human-annotated datasets is expensive and time-consuming, necessitating meticulously planned collection processes, such as determining the information to be made available to human annotators (e.g., visual aids), and careful post-processing to ensure their quality and accuracy. AudioCaps is the largest human-annotated audio captioning dataset, containing about 50k audio clips sourced from AudioSet [1], the largest audio event dataset. The training set of AudioCaps has one human-annotated caption per audio clip, while the validation and test sets have five human-annotated captions per audio clip. Clotho contains around 6k audio clips sourced from the FreeSound platform, each audio clip has five human-annotated captions. Although other human-annotated audio captioning datasets have been proposed, such as MACS [45] and AudioCaption [51], they are still limited in size and not widely used due to their inferior quality compared to AudioCaps and Clotho. The methodologies and post-processing procedures of these human-annotated datasets vary significantly. For more detailed information about these datasets, we recommend readers to refer to the respective papers. The insufficient size of these human-annotated audio-language datasets poses a significant obstacle to AL research.

Other audio researchers have also tried to harvest audio clips and their descriptions from the web. Koepke et al. [9] introduced SoundDescs dataset, which is sourced from BBC Sound Effects archive. Soham et al. [52] crawled 5k audio clips and their descriptions from two sources on the web and released WavText5K for language-based audio retrieval task. Wu et al. [53] introduced LAION-Audio-630K, the largest audio-language dataset to date, by crawling audio and descriptions from multiple sources, where most of the audio clips come from Freesound. In addition, Wu et al. investigated to employ T5 [54], a pre-trained large language model, to generate captions based on ground truth tags of audio clips in AudioSet. Much like LAION-Audio-630K, the majority of the data included in our WavCaps dataset also originates from Freesound. There are two main differences between our work and other online-harvested AL datasets. First, raw descriptions harvested from the web are very noisy. Existing harvested AL datasets did not filter or process these raw descriptions. Instead, we utilized ChatGPT to sift through and rephrase the unprocessed descriptions into sentences that resemble captions, making our WavCaps dataset suitable for all kinds of audio-language tasks including automated audio captioning. Second, classes in AudioSet are very unbalanced and most of the audio clips only have a single label [55], therefore, using T5 to transform tags to captions makes the training data unbalanced and noisy. We propose to create captions for the subset of AudioSet [56] that contains precise timing information for sound events, as labeled by humans.

Since the release of our WavCaps dataset in March 2023, our methodology has inspired a variety of subsequent research works and WavCaps has been used to train diverse audio-language models. These include the creation of new datasets [57]–[59], the development of text-to-audio generation models [60]–[62], and the advancement of large audio language models [63]–[65]. This breadth of research demonstrates the significant impact and ongoing relevance of our work in the evolving field of audio-language multimodal learning.

TABLE I: Example prompts to ChatGPT for FreeSound and AudioSet strongly-labelled subset. Transformation examples are are ignored. 'SL' refers to 'strongly-labelled'.

Data sources	Prompts
FreeSound BBC Sound Effects SoundBible	I will give you a list of descriptions of sounds. Process each individually. Extract the type of the sound and generate an audio caption describing the sound events. The audio caption should be less than 20 words. Delete the author of the sound. Delete locations, city names, country names. Delete the time. Delete device names. Delete the proper noun modifiers, number modifiers, and unit modifiers. Summarize each output into one sentence. Replace all named entities with their hypernyms. Replace people names with "someone". Do not write introductions or explanations. Only describe the sound events and do not use "heard", "recorded". Start each output sentence with its index. Make sure you are using grammatical subject-verb-object sentences. Output "Failure." if the description is not related to sound.
AudioSet SL	I will give you a number of lists containing sound events occurred sequentially in time. Process each individually. Write an one-sentence audio caption to describe these sounds. Make sure you are using grammatical subject-verb-object sentences. Directly describe the sounds and avoid using the word "heard". The caption should be less than 20 words.

III. WAVCAPS DATASET

In this section, we introduce the collection and processing steps of the WavCaps dataset. Firstly, we provide an introduction to the data sources and their respective characteristics. We then describe our proposed three-stage processing pipeline, including pre-filtering, ChatGPT-based transformation, and post-processing. Finally, we present a detailed analysis of the WavCaps dataset. Fig. 1 shows the overview of the three-stage data processing pipeline.

A. Data Sources

FreeSound [66] is an online collaborative sound sharing site started in 2005. The initial goal of FreeSound is to give support to sound researchers and sound artists who usually have trouble in finding royalty-free sound samples. After more than 10 years of development, there are more than $560\,000$ audio clips[5] uploaded by registered users, and these audio clips cover diverse contents such as music, environmental sounds, synthesized sound effects and even noises. When uploading an audio clip, each user is asked to give a short description as well as annotations of other attributes such as tags of the sound events about the uploaded audio clip. Ideally, we would like to use user-uploaded descriptions directly as audio captions. However, these raw descriptions are extremely noisy. For example, some are not related to the audio content (such as the recording device or personal feeling), some are not sentences but only nouns or verbs, and some are too detailed and include many specific information such as place names and people names. In addition, there are a large number of repetitive descriptions uploaded by the same user when they upload many audio clips at the same time. Due to these distinct attributes, processing descriptions sourced from FreeSound can prove to be extremely challenging.

BBC Sound Effects contains over 33k audio clips recorded around the world over the past 100 years. These audio clips contain extensive contents from the BBC Radiophonic workshop, the Blitz in London, BBC TV and Radio productions, and the BBC Natural History Unit archive. The raw descriptions in BBC Sound Effects begin with nouns describing the sounding objects or acoustic scenes of the sound content, followed by detailed descriptions. However, most of the detailed descriptions are also not complete sentences, and some include specific information such as recording equipment, time and places. Although these raw descriptions are noisy and cannot be directly used as audio captions, these raw descriptions all describe the content of the audio clips.

SoundBible is a website for sharing free and royalty free sound effects and audio clips. We harvested about 1500 royalty free sound effects with their raw descriptions from SoundBible, and these audio clips also cover a wide range of contents. Most of the raw descriptions in SoundBible are well-written sentences describing the audio content, but some of them still contain redundant information unrelated to the audio content. As a result, these raw descriptions remain unsuitable for direct use as audio captions.

AudioSet Strongly-Labelled Subset [56] is a sound event detection dataset that is included to increase the size of our proposed WavCaps dataset. The original AudioSet dataset [1] contains about 2M audio clips with unbalanced, weakly-labelled clip-level tags, and most of these audio clips only has one single tag. To investigate whether accuracy of the classifier trained on AudioSet is impaired by the weak labels, Shawn et al. [56] have also made available a strongly-labelled subset of approximately 100k audio clips from the AudioSet dataset. This subset has been annotated by humans and includes precise timing information for the sound events that occur within each clip. With these strongly-labelled temporal information, template-based methods or large language models can be employed to generate captions for audio clips in AudioSet strongly-labelled subset.

B. Data Processing

Online-harvested raw descriptions are very noisy and thus cannot be directly used as captions. To address this issue, image captioning datasets, such as CC3M, and CC12M, have applied a series of complex filtering and transformation steps to process raw image-description pairs, including image-based filtering, text-based filtering, image and text-based filtering and text transformation. The processing steps are stringent, resulting in a significant discard rate that we cannot afford. In order to retain as much data as possible, we propose a three-stage processing pipeline by simplifying the filtering and transformation steps according to the characteristics of the harvested audio-description pairs, as shown in Figure 1.

---

[5]Including data up to the November 30th 2022.

TABLE II: Examples of raw descriptions and ChatGPT's outputs.

Data sources	Raw descriptions	ChatGPT's output
FreeSound	This sound is of a book falling down the staircase in the Stanford University library west stacks.	A book is falling down a staircase.
	After 2h and 15min. Excerpt.	Failure.
BBC Sound Effects	Timber & Wood - Rip saw, carpenters' workshop.	Someone is using a rip saw in a carpenter's workshop.
	Motor Car: Ford Ecsort 1300 (Automatic) - Ford Escort 1300, exterior, passing with horn. (1300cc engine, automatic transmission.)	A car is passing with its horn.
SoundBible	Tasmanian Devil growling screaming hissing. Warning sounds from a Tasmanian Devil in Zoo.	An animal is growling, screaming, and hissing.
	Large Tibetan Bells ringing in a temple. Could also use for Monastery or Monks.	Bells are ringing.
AudioSet SL	['Accelerating, revving, vroom', 'Race car, auto racing']	A race car is accelerating and revving.
	['Female speech, woman speaking', 'Whoosh, swoosh, swish']	A woman is speaking while something whooshes.

TABLE III: Statistics of harvested raw data and WavCaps dataset.

Data sources	Before processing			After processing		
	num. of audio	avg. audio duration (s)	avg. text length	num. of audio	avg. audio duration (s)	avg. text length
FreeSound	567078	56.87	17.74	262300	85.98	6.77
BBC Sound Effects	33064	115.75	15.91	31201	115.04	9.67
SoundBible	1576	11.20	17.90	1232	13.12	5.87
AudioSet SL subset	108317	10.00	-	108317	10.00	9.79
WavCaps	-	-	-	403050	67.59	7.80

Pre-Filtering. Initially, we apply minimal pre-filtering to remove undesirable data, such as descriptions that do not pertain to audio content. The pre-filtering consists of audio-duration filtering and high-frequency text filtering. For audio-duration filtering, audio clips with a duration of less than one second are removed. This is mainly because short audio clips might not contain enough meaningful content and require extremely long padding during training. We limit our high-frequency text filtering solely to data obtained from FreeSound. This is because the raw descriptions from BBC Sound Effects and SoundBible are generally considered to be consistent with their corresponding audio content, and the sound event labels in the AudioSet strongly-labelled subset are manually annotated. Regarding the FreeSound data, we have noticed that high-frequency descriptions that are shared across multiple audio clips are typically uploaded by the same user who uploads multiple clips simultaneously. These descriptions are more likely to be unrelated to the audio content. Therefore, we apply high-frequency text filtering to exclude descriptions that are shared by more than 5 audio clips. These two filtering steps have removed about 265k data samples from FreeSound.

ChatGPT-based Transformation. In light of the previously discussed characteristics of the raw data, we propose three fundamental principles that we believe are essential for effectively converting raw metadata into audio captions:

- Use a single-sentence, accurate description of the audio's content using concise syntax that follows the grammatical subject-verb-object structure;
- Avoid the use of named entities such as individual names, locations, and recording devices, which are not related to the audio content;
- Omit any irrelevant subjective information, like personal emotions or opinions, that is unrelated to the sound.

Given the varied nature of raw descriptions, it is challenging to apply conventional rule-based natural language processing methods as in CC3M [40] or CC12M [41]. Attempting to do so may result in a high percentage of metadata and audio clips being discarded, comparable to that observed in CC3M.

To tackle the challenge of converting raw descriptions into captions, we propose using ChatGPT, a powerful conversational large language model trained by OpenAI[6] to perform this task automatically. Unlike traditional large language models such as T5 [54] used in LAION-Audio-630K [53], ChatGPT has been shown to excel at generating human-like responses to natural language prompts, and has garnered widespread attention for its powerful understanding, reasoning, and dialogue abilities. By designing prompts that account for the characteristics of different data sources, ChatGPT can effectively filter out sound-unrelated information and rewrite raw descriptions in to audio caption-like sentences that meet the requirements we proposed in prompts. This approach has the potential to significantly reduce the discard rate of raw descriptions and improve the quality of converted captions. Prompts we used are shown in Table I. In order to make use of ChatGPT's in-context learning ability, several transformation examples are also included in the prompts and they are different for each data source (ignored in Table I). These examples can significantly improve the caption quality. It is worth noting that other advanced conversational LLMs emerged recently, such as GPT-4 and LLaMA3 [7], could also be used. In terms of our empirical tests, however, the captions generated by these models retain similar semantic information, despite their variations in the choice of the words.

Table II presents examples of the raw descriptions and final processed captions. It can be observed that ChatGPT can transform fragmented descriptions (e.g., nouns and phrases) into sentences, remove redundant information that is either too specific or not related to sound, and condense lengthy sentences into more concise captions. This capability enhances the usability of the data and can be considered a form of data augmentation, enriching the dataset with more coherent audio captions. However, ChatGPT also serves as a critical quality control measure, ensuring the relevance and accuracy of the dataset, due to its ability to output a "Failure" message for

---

[6]https://openai.com/

[7]https://llama.meta.com/

Fig. 2: Comparative visualization of raw descriptions between ChatGPT-augmented captions, Figure 2a shows the Jaccard similarity between raw descriptions and ChatGPT-augmented captions, Figure 2b presents word clouds of top 100 words in the WavCaps dataset (top) and the entire harvested raw descriptions (bottom), stop words are ignored, and Figure 2c displays the readability of raw descriptions and ChatGPT-augmented captions as indicated by Flesch-Kincaid grade levels.

the descriptions that are not related to audio content. This ability, distinct from data augmentation, focuses on selectively refining the dataset by removing irrelevant or misleading entries, thereby enhancing the overall data quality and utility.

While ChatGPT has shown promising results in converting raw descriptions into captions, in some cases, it may still fail to follow the prompts to produce captions that meet our requirements. For instance, ChatGPT may encounter difficulty in removing numbers, names of individuals, and geographic locations from raw descriptions. Moreover, a small percentage of descriptions that have no relation with the audio content may not be filtered out.

Post-Processing. To address the aforementioned cases of incorrect processing, we implement post-processing steps to refine the captions generated by the system. To identify erroneous outputs that still include numbers, geographic locations, and individuals' names, we utilize a pre-trained named entity recognition model from spaCy [8]. Any captions with named entity information are then processed again by ChatGPT to remove these details using the same prompts but different examples. In most cases, this second round of processing successfully removes the named entity information. However, if post-processing still results in captions with named entity information, we discard these captions to ensure the quality of the final dataset. Finally, we exclude captions that are too brief by setting a minimum length threshold of three words. This ensures that the captions are informative and descriptive enough to convey a meaningful description of the audio content.

C. Dataset Analysis

Table III provides statistics for the raw data collected from four different sources before and after processing. It can be observed that more than half of the data samples from FreeSound were filtered out after processing, with the majority

TABLE IV: Comparative overview of main audio-language datasets between our proposed WavCaps dataset.

Dataset	Num. audios	Duration (h)	Text source
AudioCaps [37]	52904	144.94	Human
Clotho [44]	5929	37.00	Human
MACS [45]	3537	9.83	Human
WavText5K [52]	4072	23.20	Online raw-data
SoundDescs [9]	32979	1060.4	Online raw-data
LAION-Audio-630K [53]	633526	4325.39	Online raw-data
WavCaps	403050	7567.92	ChatGPT

Fig. 3: Distribution of audio duration from each data source in the WavCaps dataset. The audio clips from AudioSet are excluded since they are all 10 seconds long.

being removed by the high-frequency text filter. Since audio clips with a duration less than one second were excluded, the average duration of samples from FreeSound significantly increased. Conversely, only a small number of samples from BBC Sound Effects and SoundBible were removed. The dataset includes $28\,678$ unique words and $19\,974$ stemmed words. Notably, there has been a significant decrease in the average caption length.

Following [50], we employed Jaccard similarity to quantify the overlap between the original raw descriptions and the captions augmented by ChatGPT. The Jaccard similarity is
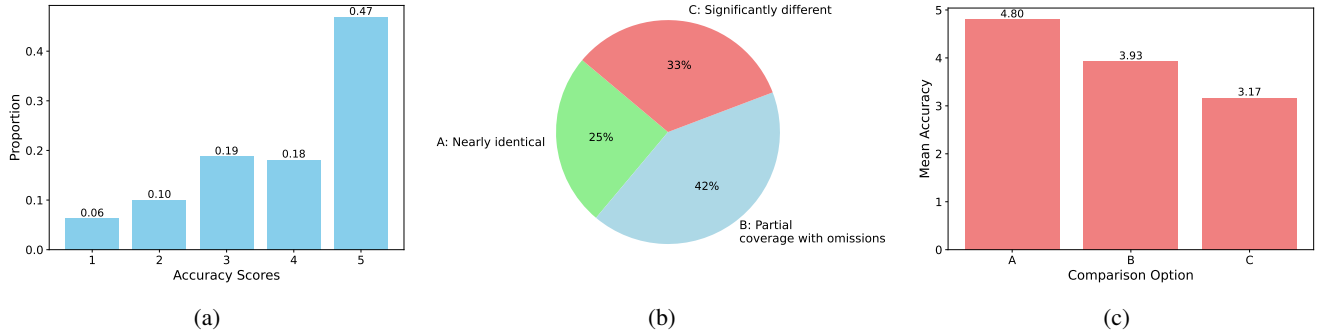
Fig. 4: Visualization of human validation results. (a) The distribution of Mean Opinion Score (MOS) ratings for ChatGPT-augmented captions with 1 indicating no correspondence and 5 indicating an exact match to the audio content. (b) The evaluators' categorization of the similarity between ChatGPT-augmented captions and human-annotated captions: (A) nearly identical, (B) partial coverage with omissions, and (C) significantly different. (c) The mean accuracy scores of the captions corresponding to each of the three comparative categories between ChatGPT-augmented captions and human-annotated captions.

a metric used to gauge the similarity and diversity of sample sets. In our analysis, the Jaccard similarity is calculated as the proportion of the shared words to the total unique words in both the raw description and the ChatGPT-augmented caption. This can be formulated as:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{1}$$

where A is the set of words in the raw description and B is the set of words in the ChatGPT-augmented caption. The normalized distribution of the Jaccard similarity scores, as presented in Figure 2a, demonstrates a generally low level of lexical overlap across various sources. Such Jaccard similarity scores, coupled with the observed reduction in caption length, suggest that the ChatGPT-augmented captions have undergone considerable transformation from the original descriptions, often involving substantial word deletion.

Fig. 2b displays the top 100 word clouds in the WavCaps dataset and the entire harvested raw descriptions, respectively. The top word clouds in the WavCaps dataset contain meaningful sound-related words such as sounding objects and sound events, whereas the top word clouds in the raw descriptions are largely devoid of meaning. This also demonstrates that ChatGPT successfully extracted sound-related information and removed information unrelated to sound. Moreover, we utilized the Flesch-Kincaid grade level [67] to evaluate the readability of both raw descriptions and ChatGPT-enhanced captions, where a lower Flesch-Kincaid grade level indicates the text is easier to read and comprehend. As depicted in Figure 2c, the lower grade levels of the augmented captions across all the data sources reflect that ChatGPT has simplified the original raw descriptions. This simplification is aligned with our prompt design shown in Section III-B, which aims to remove audio-irrelevant information and employ concise syntax.

Additionally, the WavCaps dataset includes $330\,609$ unique captions. Among these, $311\,242$ captions appear only once, showcasing a wide variety of expressions. However, there are $19\,367$ captions that recur in the dataset, indicating instances where the exactly same caption (i.e., the same sequence of words) is used more than once. It is noteworthy that captions recurring with a frequency of more than 5 times tend to be shorter, having an average of 4.9 words, compared to the average length of all the captions, which have an average of 7.8 words. An informal analysis suggests that these shorter, frequently appearing captions typically depict common and single sound events. This trend suggests that, while there is considerable variation in the dataset, certain sound events tend to be described with a consistent, standardized vocabulary and captions.

Table IV provides a comparison of key statistics between WavCaps and other audio-language datasets. Human-labeled datasets are generally limited in size, but the captions are in high quality. SoundDescs [9] is sourced from the BBC Sound Effects with no processing applied to the raw descriptions. LAION-Audio-630K has the highest number of audio clips. A majority of the audio clips in LAION-Audio-630K (around 420k) are sourced from FreeSound, resulting in a substantial overlap between their dataset and ours. To preserve the diversity of clip durations, we did not exclude any long audio clips, leading to a longer total duration in our dataset compared to others. Fig. 3 illustrates the distribution of audio durations in the WavCaps dataset. The AudioSet Strongly-labeled subset is not considered in Fig. 3, as all audio clips in AudioSet have a consistent duration of 10 seconds. Overall, in comparison to human-labelled audio captioning datasets, WavCaps is an order of magnitude larger and encompasses a greater diversity of content. In contrast to online-harvested datasets such as SoundDescs and LAION-Audio-630k, our method entails converting raw descriptions into captions and removing noisy or sound-unrelated descriptions. This approach ultimately leads to the development of the most extensive weakly-labeled audio captioning dataset that exists.

We categorize WavCaps as a large-scale pre-training resource and have not divided it into subsets, in a similar way to the practice adopted in CC12M [41], or VideoCC 3M [68], for the reason that the dataset is collected from diverse sources and the captions are weakly-labelled. It is important

to note that if the original meta descriptions do not capture a particular aspect of the audio events or even do not accurately describe the audio content, the resulting captions tend to reflect these limitations, as we only processed the text metadata, but not the audio data. In addition, high-quality audio captions may include spatial-temporal information about sound events, as observed in some human-annotated captions in datasets like Clotho and AudioCaps [37], [44]. Such information can enhance the contextual understanding of the soundscapes. However, spatial-temporal details are generally absent in our harvested descriptions, particularly those sourced from three web sources. Given that spatial-temporal information is not a perquisite for effective audio captions, its omission aligns with our design principles. Consequently, our captions primarily lack spatial-temporal details, focusing instead on the coarse-grained descriptions of audio clips. Notably, the captions from AudioSet strongly-labelled subset might include temporal relationships, as these audio tags are provided to ChatGPT in sequence, based on their occurrence in time. In a future study, it would be interesting to further improve the granularity of the captions with the spatiotemporal information of the acoustic events.

D. Human Validation

To ensure the robustness of our methodology, we incorporated a human validation process for a subset of the WavCaps dataset. We randomly selected 160 samples from FreeSound, which are also included in the test set of Clotho. We recruited eight evaluators and engaged them in a two-stage process: initially, they were asked to listen to the audio samples and assessed the accuracy of the ChatGPT-augmented captions. The evaluators rated each caption on a Mean Opinion Score (MOS) scale from 1 to 5, where 1 means that the caption fails to correctly describe the audio clip at all, and 5 indicates a perfect description of the audio clip without errors. Subsequently, for the comparison phase, the evaluators were instructed to conduct a comparison of the ChatGPT-augmented captions against the human-annotated captions from the Clotho dataset. They were instructed to categorize the relationship between the two sets of captions from three options, (A) nearly identical in content, (B) ChatGPT-augmented captions partially covering the sound events as the Clotho annotations but omitting certain details or sound events, and (C) significantly different in content and detail. This comparative analysis was critical in identifying the relative accuracy and descriptive quality of the ChatGPT-augmented captions compared to human annotation standards. In the entire evaluation process, captions were labeled as "Caption A" and "Caption B" to ensure that evaluators remained unaware of each caption's origin, whether ChatGPT-augmented or human-annotated.

The human validation results are shown in Figure 4. The mean accuracy score obtained for ChatGPT augmented captions was 3.89, indicating a high level of precision in the ChatGPT-augmented captions. Notably, more than 40% of the evaluated captions received a perfect score of 5, indicating that the majority of the ChatGPT-augmented captions accurately describe the audio events. However, there were instances of lower scores, which underscores the rationale for labeling this dataset as weakly-labeled. As the captions are augmented from the raw descriptions, and if the raw descriptions do not cover all of the sound events, the augmented captions tend to miss such sound events too. In terms of the comparison options, the majority of evaluators selected option (B), suggesting that the ChatGPT-augmented captions usually missed some sound events or lacked the specificity found in human-annotated captions. This outcome is aligned with the initial intent of our prompt design, which intentionally discards specific details and generates more general captions. For option (C), which indicated a significant difference in content and detail, the mean accuracy score was 3.17. If we regard the human-annotated captions as ground-truth, the mean accuracy score for option (C) is expected to be as low as possible. This discrepancy can partially be attributed to the annotation process used for Clotho dataset [44], [50], where annotators had access solely to audio without contextual information, leading to possible interpretations of the same sound as different events. In conclusion, the human validation results affirm the overall quality of our dataset. The results highlight the efficacy of the ChatGPT-augmented captions that are generally accurate but could miss some details or sound events.

IV. EXPERIMENTS

To evaluate the impact of the proposed WavCaps dataset, we conducted experiments on several audio-language multimodal learning tasks, including audio-language retrieval, automated audio captioning, zero-shot audio classification, and text-to-sound generation. In this section, we provide a description of the AL tasks we considered, along with the corresponding experimental settings, results and analysis. For all the experiments excluding text-to-sound generation, audio clips are sampled with a 32k Hz sampling rate, and we use 64-dimensional log mel-spectrograms extracted by a 1024-point Hanning window with a hop size of 320 samples as input audio features.

A. Audio-Language Retrieval

Audio-language retrieval involves searching for an audio clip or a caption in a database based on a query from another modality. To perform this task, the model learns Acoustic Semantic Embeddings (ASE) [10] that map paired audio clips and captions closer in the embedding space, while keeping embeddings for non-paired audio clips and captions far apart. We evaluate audio-language retrieval under three training settings: zero-shot, pretraining, and fine-tuning.

1) Models: Following previous works [10], [53], we build an ASE model based on a two-tower architecture, where an audio encoder is employed to encode audio representations while an language encoder is used to encode captions. Two types of audio encoders, a CNN14 from pretrained audio neural networks (PANNs) [55] and a Transformer network-HTSAT [70] are considered, which are both pretrained on AudioSet with an audio tagging task. A pretrained BERT-base network [71] is employed as the language encoder. A 2-layer multilayer perceptron with a ReLU activation in between

TABLE V: Experimental results of audio-language retrieval on test sets of AudioCaps and Clotho. "AC" refers to "AudioCaps", "LA" refers to "LAION-Audio-630K", "ZS" refers to "zero-shot", "PT" refers to "pretraining", and "FT" refers to "fine-tuning". Higher score means better performance.

Model	Training Dataset	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 [9]	AC or Clotho	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-ASE [10]	AC or Clotho	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 [69]	AC or Clotho	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-HTSAT [52]	AC+Clotho+WavText5K	34.6	70.2	82.0	41.9	73.1	84.6	16.7	41.1	54.1	20.0	44.9	58.7
LAION [53]	AC+Clotho	36.7	70.9	83.2	45.3	78.0	87.7	12.0	31.6	43.9	15.7	36.9	51.3
LAION [53]	AC+Clotho+LA	32.7	68.0	81.2	43.9	77.7	87.6	15.6	38.6	52.3	23.7	48.9	59.9
LAION [53]	AC+Clotho+LA+AudioSet	36.1	71.8	83.9	46.8	82.9	90.7	16.1	38.3	51.1	22.7	48.5	60.8
LAION (fusion) [53]	AC+Clotho+LA+AudioSet	35.1	71.5	83.6	45.8	80.9	91.6	18.2	42.5	54.4	25.7	51.5	63.4
HTSAT-BERT-ZS	LA	14.4	39.4	53.9	15.3	40.1	56.2	13.8	35.5	47.4	15.8	39.7	51.8
HTSAT-BERT-PT	AC+Clotho+LA	34.2	69.6	83.3	41.9	76.2	85.5	17.4	43.0	56.0	23.9	46.4	61.3
CNN14-BERT	AC+Clotho	30.3	66.1	79.7	36.7	70.6	83.0	15.6	39.7	52.6	19.9	41.2	55.9
CNN14-BERT-ZS	WavCaps	25.4	58.2	73.2	32.4	63.6	77.1	17.5	41.7	54.9	21.7	45.7	57.6
CNN14-BERT-PT	AC+Clotho+WavCaps	34.7	69.1	82.5	44.6	76.3	86.2	21.2	46.4	59.4	25.9	52.6	65.8
CNN14-BERT-FT	AC+Clotho+WavCaps	35.1	70.0	82.1	45.7	76.1	87.7	**21.5**	**47.9**	**61.9**	**27.1**	**52.7**	**66.3**
HTSAT-BERT	AC+Clotho	39.2	74.9	86.5	49.5	81.9	91.5	15.6	38.4	52.0	21.0	43.8	55.7
HTSAT-BERT-ZS	WavCaps	28.6	61.1	75.8	40.2	69.4	80.3	16.5	38.8	50.9	20.0	43.3	56.6
HTSAT-BERT-PT	AC+Clotho+WavCaps	39.7	74.5	86.1	51.7	82.3	90.6	19.5	45.2	58.2	23.4	50.9	63.4
HTSAT-BERT-FT	AC+Clotho+WavCaps	**42.2**	**76.5**	**87.1**	**54.6**	**85.2**	**92.4**	19.7	45.7	59.4	26.9	52.6	64.9

is appended after these two encoders to project respective features into the shared embedding space. Cosine similarity is used to quantify the similarity between audio embeddings and language embeddings, and can be formulated as

$$s_{ij} = \frac{f(a_i) \cdot g(t_j)}{||f(a_i)||_2 ||g(t_j)||_2} \quad (2)$$

where $f(\cdot)$ is the audio encoder and $g(\cdot)$ is the language decoder, a_i is the audio clip indexed with i in a batch, t_j is the caption indexed with j in a batch, and s_{ij} is the similarity score. The model is trained with a normalized temperature-scaled cross entropy loss (NT-Xent) [72] in a bi-directional manner, and can be formulated as:

$$\mathcal{L} = -\frac{1}{2B} \sum_{i=1}^{B} \log \frac{\exp(s_{ii}/\tau)}{\sum_{j=1}^{B} \exp(s_{ij}/\tau)} + \\ \log \frac{\exp(s_{ii}/\tau)}{\sum_{j=1}^{B} \exp(s_{ji}/\tau)} \quad (3)$$

where B is the batch size, and τ is a temperature hyper-parameter. This training strategy is also known as contrastive language-audio pretraining (CLAP) [73].

2) Experimental Setup: We first train our two models on the merged training sets of AudioCaps and Clotho as baselines. For the zero-shot setting, we exclude all overlapping samples in AudioCaps and Clotho from the WavCaps dataset. We provide IDs of overlapping samples with the release of the dataset[9]. The baseline and zero-shot models are trained for 15 epochs with a batch size of 128 and a learning rate of 5×10^{-5} using the Adam [74] optimizer. For the supervised setting, we merge WavCaps, AudioCaps and Clotho together as a large training set (validation and test sets of AudioCaps and Clotho are not included). The models are trained for 40 epochs and other hyperparameters are the same as in zero-shot settings. For the fine-tuned setting, we aim to study the impact of pretraining on WavCaps dataset. We further fine-tune the

[9]https://huggingface.co/datasets/cvssp/WavCaps/tree/main/json_files/blacklist

models trained under pretraining setting on AudioCaps and Clotho for 20 epochs, respectively. The temperature hyperparameter τ is set to 0.07 for all settings. For the HTSAT audio encoder, all the audio clips are randomly cropped or padded to 10 seconds, because HTSAT requires fixed-sized inputs. For the CNN14 audio encoder that can receive variable length inputs, we set the maximum input duration as 30 seconds and audio clips longer than 30 seconds are randomly cropped. During training, audio clips with similar duration are grouped within a batch. Model checkpoints are selected based on their performance on validation sets after each epoch and the final model performance is evaluated on the test sets of AudioCaps and Clotho using recall at rank k (R@k). For a query, R@k is 1 if the positive item appears in the top k retrieved items, otherwise 0. The final R@k is averaged across the dataset.

3) Results and Analysis: Table V presents the audio-language retrieval results on the AudioCaps and Clotho datasets, where top half of the table shows the results of existing methods and bottom half of the table shows our results. In the zero-shot setting, both of our models show strong ability for zero-shot retrieval. When compared with previous SOTA models that were trained only on respective datasets, both models outperform the previous SOTA models on the Clotho dataset and achieve comparable results on the AudioCaps dataset. This demonstrates that the models trained on WavCaps in zero-shot setting generalize well on both AudioCaps and Clotho datasets. In the pretraining setting, the performance is improved compared to baselines and zero-shot models, and also achieved SOTA results on both datasets. It can be observed that using more training data leads to greater improvement on audio-to-text metrics compared to text-to-audio metrics. With fine-tuning, both of our models further improve the performance, and outperform all existing methods by a significant margin on both audio-to-text retrieval (16.7% improvement on $R@1$ on AudioCaps and 5.4% improvement on $R@1$ on Clotho) and text-to-audio retrieval (15.0% improvement on $R@1$ on AudioCaps and 18.1% improvement on $R@1$ on Clotho).

In order to demonstrate the effectiveness of our proposed

WavCaps dataset, we primarily compare our results with models trained on the LAION-Audio-630K dataset [53]. For LAION's models, we observe a notable decline in performance on the AudioCaps dataset when the LAION-Audio-630K dataset is added to the training, while there is an improvement in results for the Clotho dataset. After incorporating AudioSet into the training, LAION's model demonstrates improved performance on AudioCaps, but remains close to their baseline. In our case, our models outperform LAION's models in most metrics for both datasets, despite utilizing less data. When incorporating WavCaps into the training, our models exhibit significant improvements on both datasets. It is important to highlight that LAION's training set, which includes AudioSet, is approximately six times larger than ours, totaling 2.63 million in size. Due to the differences in the architecture between our model and LAION's model, we trained our HTSAT-BERT model on the LAION-Audio-630K dataset under both zero-shot and pretraining settings. This approach ensures a fair comparison by maintaining consistency in all training parameters, with the only variation being the dataset used. The LAION-Audio-630K dataset we used contains $594\,404$ audio clips, which is slightly smaller than the official dataset since we do not have access to some of the audio clips. The comparison between the official LAION-Audio-630K dataset and our version is shown in Appendix A. It can be observed that the models trained on WavCaps greatly outperform those trained on LAION-Audio-630K under both training settings. Our experimental results reveal that the quality of the audio-language dataset significantly influences the model performance. By adopting our proposed three-stage processing pipeline to filter out noisy data and rewrite raw descriptions into captions, we achieve SOTA performance using a smaller quantity of higher-quality data.

Finally, our two models show varying performance levels on the AudioCaps and Clotho datasets. The CNN14-based model underperforms the HTSAT-based model on the AudioCaps dataset, yet it surpasses the HTSAT-based model on the Clotho dataset. A potential explanation for this could be the variable duration of audio clips in Clotho, which range from 15 to 30 seconds. Randomly cropping these clips into 10 seconds in the HTSAT model may result in information loss. As CNN14 is capable of handling variable duration, it outperforms HTSAT on the Clotho dataset. We believe that the performance of our models could be further improved by employing feature fusion methods, as seen in [53], to process variable-duration audio clips. However, we plan to explore this in future work.

B. Automated Audio Captioning

Automated audio captioning is the task of generating a natural language sentence to describe the content of an audio clip, which mainly concerns environmental sounds and ignores possible voice content [11]. We continue to evaluate audio captioning on AudioCaps and Clotho datasets.

1) Models: Audio captioning is generally solved by an encoder-decoder model, where an encoder is leveraged to extract audio features and a decoder is employed to generate captions based on audio features extracted from the encoder.

We build our model based on the baseline of Detection and Classification of Acoustic Scenes and Events (DCASE) 2022 challenge task 6 [12]. Similar to the audio-language retrieval approach mentioned earlier, we explore two types of audio encoders, namely, a CNN14 and an HTSAT, for our audio captioning models. The language decoder is a pretrained language model, BART based network [82]. BART is a sequence-to-sequence language model consisting of both Transformer encoder and decoder blocks, pretrained on large text corpora. The model is trained with a cross-entropy loss and can be formulated as:

$$\mathcal{L}_{\text{CE}}(\theta) = -\frac{1}{T} \sum_{t=1}^{T} \log p(y_t | y_{1:t-1}, x, \theta) \qquad (4)$$

where x is an input audio clip, y_t is the t-th ground truth token in a sentence whose length is T, and θ are the parameters of the audio captioning model.

2) Experimental Setup: We first explored a zero-shot training setting, where all overlapping samples from the Clotho and AudioCaps datasets are excluded for training. Then, we adopt a two-stage training paradigm similar to the fine-tuning setting in audio-language retrieval. The whole model is first pretrained on WavCaps, together with the training sets from Clotho and AudioCaps, using a learning rate of 5×10^{-5} and a batch size of 48 for 15 epochs. The pretrained model is further fine-tuned on AudioCaps and Clotho for 20 epochs with a learning rate of 5×10^{-6}, respectively. During the whole training process, we ensure that no data from validation or test sets of Clotho and AudioCaps are used for training. To assess the impact of pretraining using WavCaps dataset, we train the two models exclusively on AudioCaps and Clotho as baselines. The performance is evaluated using conventional metrics including BLEU_n [83], ROGUE_l [84], METEOR [85], CIDEr [86], SPICE [87], and SPIDEr [88], where SPIDEr is generally employed as the main metric in the literature.

3) Results and Analysis: Results are presented in Table VI. In the zero-shot training scenario, both models demonstrate robust zero-shot capabilities, affirming the suitability of our dataset for the audio captioning task. The fine-tuning results further demonstrate that our models surpass existing methods, achieving new SOTA performance on the Clotho and AudioCaps datasets. Notably, our model outperforms even those methods that incorporate the validation set into their training for the Clotho dataset. Compared to the baseline systems, pretraining on WavCaps leads to a significant improvement in the final performance on both datasets. These outcomes suggest that ChatGPT effectively transforms raw descriptions into caption-like sentences, thereby boosting audio captioning performance. In alignment with findings from audio-language retrieval above, the CNN14 audio encoder outperforms the HTSAT audio encoder on the Clotho dataset but exhibits inferior performance on the AudioCaps dataset.

C. Zero-Shot Audio Classification

Audio classification aims at classifying the class of the sound presented in an audio clip. We carry out zero-shot audio classification on three popular audio event datasets to evaluate

TABLE VI: Automated audio captioning results on the test sets of AudioCaps and Clotho. A higher score means better performance. "*" indicates that the validation set of Clotho is used to train the model.

Dataset	Model	BLEU$_1$	BLEU$_4$	ROUGE$_l$	METEOR	CIDEr	SPICE	SPIDEr
Clotho	Prefix [75]	56.0	16.0	37.8	17.0	39.2	11.8	25.5
	MAAC [76]	57.5	17.4	37.7	17.4	41.9	11.9	26.9
	Netease [77]	58.3	17.7	38.8	17.9	45.6	12.8	29.2
	FeatureCut [78]	**60.1**	17.9	38.9	17.6	43.6	12.2	27.9
	CLIP-AAC* [79]	57.2	16.9	37.9	17.1	40.7	11.9	26.3
	Netease* [77]	58.5	**18.2**	**40.0**	18.4	47.4	**13.5**	30.2
	CNN14-BART (zero-shot)	29.9	7.2	29.3	12.0	24.8	8.7	16.7
	CNN14-BART (baseline)	56.0	16.0	37.0	17.1	39.3	11.7	25.5
	CNN14-BART	**60.1**	18.0	**40.0**	**18.5**	**48.8**	13.3	**31.0**
	HTSAT-BART (zero-shot)	36.2	7.8	30.0	12.4	25.0	8.6	16.8
	HTSAT-BART (baseline)	57.6	16.4	38.2	17.5	41.5	11.9	26.7
	HTSAT-BART	58.5	16.8	38.3	18.4	46.2	13.3	29.7
AudioCaps	ACT [48]	64.7	25.2	46.8	22.2	67.9	16.0	42.0
	V-ACT [80]	69.8	28.1	49.4	23.7	71.1	17.2	44.2
	Prefix [75]	**71.3**	**30.9**	50.3	24.0	73.3	17.7	45.5
	BART-tags [12]	69.9	26.6	49.3	24.1	75.3	17.6	46.5
	AL-MixGEN [81]	70.0	28.9	50.2	24.2	76.9	18.1	47.5
	CNN14-BART (zero-shot)	55.1	12.4	37.1	18.6	45.3	11.9	28.6
	CNN14-BART (baseline)	67.0	26.1	48.3	23.1	72.1	16.9	44.5
	CNN14-BART	69.3	27.2	49.9	24.7	75.6	17.9	46.8
	HTSAT-BART (zero-shot)	51.3	11.0	37.8	20.4	39.3	13.8	26.7
	HTSAT-BART (baseline)	67.5	27.2	48.3	23.7	71.1	17.7	44.4
	HTSAT-BART	70.7	28.3	**50.7**	**25.0**	**78.7**	**18.2**	**48.5**

the generalization and robustness of models trained on our WavCaps dataset.

1) Models: We formulate the audio classification task as an audio-language retrieval problem, following the approach as used in CLIP [35]. Initially, we encode all class labels as class embeddings using a text encoder without using any prompts. Each audio clip will be encoded by the audio encoder and then compared with the class embeddings to get a similarity score for each class. Those similarity scores will be normalized to get a final probability distribution over the classes. We use 'HTSAT-BERT-PT' model described in Section IV-A.

2) Experimental Setup: WavCaps and the training sets of AudioCaps and Clotho are merged as a training set, where overlapping samples co-occurred in the evaluation datasets are excluded. The training settings are same as those in audio-language retrieval. Three audio event datasets, ESC-50 [2], UrbanSound8K [89], and VGGSound [90] are employed. ESC-50 is an environmental sound classification dataset consisting of 2000 5-seconds audio clips annotated with 50 classes. UrbanSound8K contains 8732 audio clips less or equal to 4 seconds of urban sounds from 10 classes. VGGSound contains about 200k audio clips for 310 classes sourced from YouTube videos. ESC-50 and UrbanSound8K are officially split into 5-folds and 10-folds for cross-validation. Therefore, we use all audio clips in ESC-50 and UrbanSound8K, and the test set of VGGSound for evaluation. Top-1 accuracy is used as the evaluation metric.

3) Results and Analysis: Table VII presents the results, with the top row showing the supervised SOTA performance on each dataset, while the remaining rows show zero-shot results. Compared to our model, LAION and BLAT [95] both use more data to train their models. Our models achieved SOTA zero-shot results on all three datasets, significantly outperforming other models on ESC-50 and UrbanSound8K. In the case of VGGSound, Wu et al. [53] (LAION) reported a 46.2% accuracy, but their use of AudioSet for model training

TABLE VII: Results of the top-1 accuracy on zero-shot audio classification.

Model	ESC-50	UrbanSound8K	VGGSound
Supervised SOTA	98.1 [91]	90.0 [92]	75.4 [53]
Wav2CLIP [93]	41.4	40.4	10.0
AudioCLIP [94]	69.4	65.3	-
CLAP [73]	82.6	73.2	-
BLAT [95]	80.6	77.3	14.9
LAION [53]	91.0	77.0	29.1 (46.2)
Ours	**94.8**	**80.6**	**29.6**

without excluding overlapping samples between AudioSet and VGGSound led to a data leakage issue. Overall, in comparison with LAION and BLAT, our approach yielded superior results using less data, which highlights the effectiveness of our proposed WavCaps dataset.

The zero-shot results we obtained were close to the supervised SOTA on ESC-50 and UrbanSound8K, but exhibited a considerable margin on the VGGSound dataset. ESC-50 and UrbanSound8K datasets are composed of 50 and 10 classes, respectively, while VGGSound dataset includes 310 classes. However, the level of granularity in VGGSound's classification scheme may be too high, which could lead to models trained on WavCaps struggling to generalize effectively across all 310 classes.

D. Text-based Sound Generation

Text-based sound generation is a task that generates sound, including speech, music, and sound effect, with textual information [17]. We follow previous studies [16]–[18] and perform training and evaluation on the AudioCaps [37] dataset.

1) Models: AudioLDM [17] is a text-to-sound generation model that builds upon contrastive language-audio pre-trained (CLAP) encoders. When trained effectively, the CLAP encoders have the capability to capture the relationships between different modalities and ease the training challenges of

TABLE VIII: Performance comparison on Text-based sound generation. AudioGen is marked with[†] because the pretrained model and the evaluation data of AudioGen are not open-sourced, which may lead to unreliable comparison.

Model	Train Condition	FAD ↓	IS ↑	KL ↓	FD ↓
DiffSound [18]	-	7.75	4.01	2.52	47.68
AudioGen-base[†] [16]	-	3.13	-	2.09	-
AudioLDM$_{LAION}$ [17]	Audio	2.98	**7.12**	2.2	24.04
	Text	2.47	6.91	2.25	24.84
AudioLDM$_{WavCaps}$	Audio	2.85	5.42	2.31	24.89
	Text	**2.29**	7.05	2.3	26.3
AudioLDM$_{WavCaps-FT}$	Audio	2.98	5.79	**2.17**	**21.9**
	Text	2.58	6.38	2.23	25.27

audio generative models. The original AudioLDM, denoted by AudioLDM$_{LAION}$, adopts a CLAP model developed by [53] with a dataset of around 2.6 million audio-text pairs. We re-implement AudioLDM with a CLAP model trained with our proposed WavCaps, denoted by AudioLDM$_{WavCaps}$. We further finetuned our CLAP model on AudioCaps training set for text-to-sound generation, denoted by AudioLDM$_{WavCaps-FT}$. Specifically, we use the pretrained Variational Autoencoder (VAE) and vocoder in the open-source implementation of Audi-oLDM[10] and trained a new latent diffusion model (LDM). The LDM in AudioLDM is trained with the re-weighted training objective [96], given by

$$L_n(\theta) = \mathbb{E}_{\boldsymbol{z}_0, \boldsymbol{\epsilon}, n} \|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\boldsymbol{z}_n, n, \boldsymbol{E}^x)\|_2^2, \quad (5)$$

$$q(\boldsymbol{z}_n | \boldsymbol{z}_0) = \mathcal{N}(\boldsymbol{z}_n; \sqrt{\bar{\alpha}_n} \boldsymbol{z}_0, (1 - \bar{\alpha}_n)\boldsymbol{\epsilon}), \quad (6)$$

where $\epsilon \sim \mathcal{N}(0, \boldsymbol{I})$ denotes the standard Gaussian noise, $\bar{\alpha}_n$ is the noise schedule [17], $\boldsymbol{z}_0$ is the original VAE latent extracted from audio, and $\boldsymbol{z}_n$ is the output of the n-th forward diffusion step. To benchmark our models, we also included two state-of-the-art audio generation models, namely DiffSound [18] and AudioGen [16].

2) Experimental Setup: The training and testing set split of AudioCaps are the same as the Audio Captioning experiments in Section IV-B. The LDM is optimized on the AudioCaps training set by an Adam optimizer with a learning rate of 3×10^{-5}. We adopt a batch size of 8 and train LDM for a total of 400k steps. We evaluate the performance on the AudioCaps test set at intervals of 50k training steps and choose the outcome with the most optimal Frechet Audio Distance (FAD) as the final result for reporting purposes. The training data from AudioCaps are resampled to 16kHz before the model training. The setting on spectrogram calculation follows exactly the setting of [17]. Regarding the AudioLDM-based models, we conduct experiments with two distinct modalities as training conditions: audio embedding and text embedding obtained from CLAP. When conditioned on audio, the model undergoes self-supervised training for audio genera-tion, whereas conditioning on text involves supervised training using paired audio-text data.

In accordance with [17] and their methodology, we assess the performance of our models using a range of metrics, including the FAD, Inception Score (IS), KL divergence (KL), and PANNS-based Frechet Distance (FD). The FAD and FD

---

metrics evaluate the similarity between two audio data distri-butions, while IS measures the diversity of the generated audio data and its similarity to the target audio data distribution. Additionally, KL provides a sample-level measure of similarity between generated and target samples.

3) Results and Analysis: Table VIII shows the evaluation result of AudioLDM paired with different CLAP models as well as the baseline methods. Even with a much smaller dataset size (15%) compared with AudioLDM$_{LAION}$, our AudioLDM$_{WavCaps}$ still achieves a comparable performance. AudioLDM$_{WavCaps}$ even performs better on FAD and the inception score when we use text as training embedding, indicating our dataset has better text labelling quality. When our CLAP model is finetuned on AudioCaps, KL and FD by AudioLDM$_{WavCaps-FT}$ is improved significantly, which are even better than the KL and FD of AudioLDM$_{LAION}$. Nevertheless, our model does not perform well on IS when conditioned on audio embedding. One possible explanation for this is that the scale of our training data is relatively small, which could restrict the model's ability to generalize. However, we can address this issue by augmenting the CLAP training data with label-to-caption augmented audioset data. Compared with the baseline methods, all our implementation of AudioLDM outperforms DiffSound and FAD of AudioGen by a large margin.

E. Ablation Study

We carried out ablation studies to evaluate the effectiveness of each component in our processing pipeline. Recognizing that both step 2 (ChatGPT-based Transformation) and step 3 (Post-Processing) include caption refinement using ChatGPT, we merged these steps in our analysis. This approach enabled a direct comparison between the efficacy of the original raw descriptions and the ChatGPT-augmented captions. We specifically selected the FreeSound subset for our in-depth study, considering its notable complexity, and the fact that the first step of our pipeline primarily focuses on filtering data from FreeSound.

Table IX presents the experimental results of zero-shot audio language retrieval. The model used is 'HTSAT-BERT' and the training settings are the same as those in the zero-shot setting in Section IV-A. Because we only used the FreeSound subset, the number of training samples is 561 124 and 257 040 before and after the 1st step of processing, respectively. It can be observed that all scores are improved after applying pre-filtering with less than half of the data. After the 3rd step, where the raw descriptions are augmented to captions by ChatGPT, slight improvements can be observed across all scores. When including data from other resources, all the scores are further improved, especially on AudioCaps.

The impact of caption refinement using ChatGPT in the audio-language retrieval task was relatively modest. Our pri-mary motivation was to enhance raw descriptions into more comprehensive audio captions. To validate this approach, we conducted zero-shot automated audio captioning experiments. We trained the 'CNN14-BART' model separately with the raw descriptions and the ChatGPT-augmented captions, using

---

[10]https://github.com/haoheliu/AudioLDM

TABLE IX: Ablation study results of zero-shot audio-language retrieval on the test sets of AudioCaps and Clotho.

| Settings | AudioCaps | | | | | | Clotho | | | | | |
| | Text-to-Audio | | | Audio-to-Text | | | Text-to-Audio | | | Audio-to-Text | | |
	$R@1$	$R@5$	$R@10$	$R@1$	$R@5$	$R@10$	$R@1$	$R@5$	$R@10$	$R@1$	$R@5$	$R@10$
Before 1st step (FreeSound subset)	12.9	37.0	51.8	13.8	38.2	55.1	12.5	36.2	46.9	15.4	36.8	48.3
After 1st step (FreeSound subset)	15.6	43.2	59.6	19.4	46.8	62.8	14.3	36.8	49.4	16.5	39.0	52.3
After 3rd step (FreeSound subset)	17.3	46.3	61.8	20.9	49.1	64.7	14.8	37.6	51.1	17.0	39.8	52.7
WavCaps	28.6	61.1	75.8	40.2	69.4	80.3	16.5	38.8	50.9	20.0	43.3	56.6

TABLE X: Ablation study results of zero-shot automated audio captioning on the test sets of AudioCaps and Clotho.

Settings	Dataset	$BLEU_1$	$BLEU_4$	$ROUGE_l$	METEOR	CIDEr	SPICE	SPIDEr
After 1st step (FreeSound subset)	Clotho	34.9	3.9	25.0	10.3	16.7	5.7	11.2
After 3rd step (FreeSound subset)		34.6	8.3	31.4	12.7	27.2	9.0	18.1
After 1st step (FreeSound subset)	AudioCaps	30.5	3.0	24.6	10.7	21.2	5.8	13.5
After 3rd step (FreeSound subset)		34.0	6.8	30.0	12.1	36.2	7.7	22.0

the FreeSound subset of the WavCaps dataset. Notably, the FreeSound subset, after the filtering of 1st processing step, comprised 257 040 training samples. Table X shows the results of the ablation study for automated audio captioning. We can observe significant improvements after using the ChatGPT-augmented captions, where the SPIDEr score was improved by 6.9% on the Clotho dataset and by 8.5% on the Audio-Caps dataset. This proves the effectiveness of the proposed ChatGPT-based caption refinement and the suitability of the WavCaps dataset for the audio captioning task.

V. CONCLUSION

Data scarcity presents a significant challenge in audio-language multimodal learning research. In this study, we have introduced WavCaps, a large-scale weakly-labelled audio captioning dataset, created by collecting audio clips and their corresponding raw descriptions from the web. A three-stage processing pipeline is proposed to filter and transform crawled raw descriptions into captions using ChatGPT. Our evaluation of the WavCaps dataset on multiple audio-language multimodal learning tasks resulted in new state-of-the-art performance across all tasks. Our aspiration is that WavCaps can not only facilitate the progress of audio-language multimodal learning research but also showcase how ChatGPT-like LLMs can be utilized to enrich academic research.

ACKNOWLEDGMENTS

APPENDIX

REFERENCES

TABLE XI: Comparison of the official LAION-Audio-630K and our downloaded version in terms of the number of audio clips and the number of captions per audio.

| Data Source	Official	Ours	Captions per audio
FreeSound	515581	501307	2
BBC Sound Effects	15973	15973	1
Free To Use Sounds	6370	6370	1
Sonniss Game Effects	5049	5561	1
We Sound Effects	488	488	1
Paramount Motion Sound Effects	4420	4932	1
Audiostock	10000	0	1
Epidemic Sound	75645	59773	2

[1] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio Set: An ontology and human-labeled dataset for audio events," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, New Orleans, LA, 2017.

[2] K. J. Piczak, "ESC: Dataset for environmental sound classification," in *Proceedings of the 23rd Annual ACM Conference on Multimedia*. ACM Press, pp. 1015–1018.

[3] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, "FSD50K: An open dataset of human-labeled sound events," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, p. 829–852, 2022.

[4] Q. Kong, C. Yu, Y. Xu, T. Iqbal, W. Wang, and M. D. Plumbley, "Weakly labelled AudioSet tagging with attention neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 11, pp. 1791–1802, 2019.

[5] A. Mesaros, T. Heittola, T. Virtanen, and M. D. Plumbley, "Sound event detection: A tutorial," *IEEE Signal Processing Magazine*, vol. 38, no. 5, pp. 67–83, 2021.

[6] Y. Xu, Q. Huang, W. Wang, P. Foster, S. Sigtia, P. J. B. Jackson, and M. D. Plumbley, "Unsupervised feature learning based on deep models for environmental audio tagging," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 6, pp. 1230–1241, 2017.

[7] D. Barchiesi, D. Giannoulis, D. Stowell, and M. D. Plumbley, "Acoustic scene classification: Classifying environments from the sounds they produce," *IEEE Signal Processing Magazine*, vol. 32, no. 3, pp. 16–34, 2015.

[8] H. Xie, O. Räsänen, K. Drossos, and T. Virtanen, "Unsupervised audio-caption aligning learns correspondences between individual sound events and textual phrases," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2022, pp. 8867–8871.

[9] A. S. Koepke, A.-M. Oncescu, J. F. Henriques, Z. Akata, and S. Albanie, "Audio retrieval with natural language queries: A benchmark study," *IEEE Transactions on Multimedia*, vol. 25, pp. 2675–2685, 2023.

[10] X. Mei, X. Liu, J. Sun, M. D. Plumbley, and W. Wang, "On metric learning for audio-text cross-modal retrieval," in *Proc. Interspeech*. ISCA, 2022.

[11] X. Mei, X. Liu, M. D. Plumbley, and W. Wang, "Automated audio captioning: An overview of recent progress and new challenges," *Journal on Audio, Speech, and Music Processing*, no. 26, 2022.

[12] F. Gontier, R. Serizel, and C. Cerisara, "Automated audio captioning by fine-tuning BART with AudioSet tags," in *Proceedings of the 6th Detection and Classification of Acoustic Scenes and Events 2021 Workshop*, Barcelona, Spain, November 2021, pp. 170–174.

[13] Y. Koizumi, R. Masumura, K. Nishida, M. Yasuda, and S. Saito, "A Transformer-based audio captioning model with keyword estimation," in *Proc. Interspeech*. ISCA, 2020, pp. 1977–1981.

[14] S. Lipping, P. Sudarsanam, K. Drossos, and T. Virtanen, "Clotho-AQA: A crowdsourced dataset for audio question answering," in *30th European Signal Processing Conference*, 2022, pp. 1140–1144.

[15] H. M. Fayek and J. Johnson, "Temporal reasoning via audio question answering," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, p. 2283–2294, Aug 2020.

[16] F. Kreuk, G. Synnaeve, A. Polyak, U. Singer, A. Défossez, J. Copet, D. Parikh, Y. Taigman, and Y. Adi, "AudioGen: Textually guided audio generation," in *The Eleventh International Conference on Learning Representations*, 2023. [Online]. Available: https://openreview.net/forum?id=CYK7RfcOzQ4

[17] H. Liu, Z. Chen, Y. Yuan, X. Mei, X. Liu, D. Mandic, W. Wang, and M. D. Plumbley, "AudioLDM: Text-to-audio generation with latent diffusion models," in *Proceedings of the 40th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 202. PMLR, 23–29 Jul 2023, pp. 21 450–21 474.

[18] D. Yang, J. Yu, H. Wang, W. Wang, C. Weng, Y. Zou, and D. Yu, "Diffsound: Discrete diffusion model for text-to-sound generation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 1720–1733, 2023.

[19] 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.

[20] X. Liu, H. Liu, Q. Kong, X. Mei, J. Zhao, Q. Huang, M. D. Plumbley, and W. Wang, "Separate what you describe: Language-queried audio source separation," in *Proc. Interspeech*. ISCA, 2022.

[21] S. Uppal, S. Bhagat, D. Hazarika, N. Majumder, S. Poria, R. Zimmermann, and A. Zadeh, "Multimodal research in vision and language: A review of current and emerging trends," *Information Fusion*, vol. 77, pp. 149–171, 2022.

[22] D. Qi, L. Su, J. Song, E. Cui, T. Bharti, and A. Sacheti, "ImageBERT: Cross-modal pre-training with large-scale weak-supervised image-text data," *arXiv preprint arXiv:2001.07966*, 2020.

[23] L. Zhou, H. Palangi, L. Zhang, H. Hu, J. Corso, and J. Gao, "Unified vision-language pre-training for image captioning and VQA," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 13 041–13 049.

[24] X. Li, X. Yin, C. Li, X. Hu, P. Zhang, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei, Y. Choi, and J. Gao, "OSCAR: Object-semantics aligned pre-training for vision-language tasks," *European Conference on Computer Vision*, 2020.

[25] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with CLIP latents," *arXiv preprint arXiv:2204.06125*, 2022.

[26] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2022, pp. 10 684–10 695.

[27] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "VQA: Visual question answering," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2425–2433.

[28] Z. Yu, J. Yu, Y. Cui, D. Tao, and Q. Tian, "Deep modular co-attention networks for visual question answering," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6281–6290.

[29] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[30] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[31] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[32] J. Li, R. Selvaraju, A. Gotmare, S. Joty, C. Xiong, and S. C. H. Hoi, "Align before fuse: Vision and language representation learning with momentum distillation," *Advances in Neural Information Processing Systems*, vol. 34, pp. 9694–9705, 2021.

[33] J. Li, D. Li, C. Xiong, and S. Hoi, "BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *International Conference on Machine Learning*. PMLR, 2022, pp. 12 888–12 900.

[34] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig, "Scaling up visual and vision-language representation learning with noisy text supervision," in *International Conference on Machine Learning*. PMLR, 2021, pp. 4904–4916.

[35] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8748–8763.

[36] Y. Gong, Y.-A. Chung, and J. Glass, "AST: Audio Spectrogram Transformer," in *Proc. Interspeech*, 2021, pp. 571–575.

[37] 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*, 2019, pp. 119–132.

[38] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2009.

[39] 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.

[40] P. Sharma, N. Ding, S. Goodman, and R. Soricut, "Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*. Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 2556–2565.

[41] S. Changpinyo, P. Sharma, N. Ding, and R. Soricut, "Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3558–3568.

[42] Z.-H. Zhou, "A brief introduction to weakly supervised learning," *National Science Review*, vol. 5, no. 1, pp. 44–53, 08 2017.

[43] Y.-F. Li, L.-Z. Guo, and Z.-H. Zhou, "Towards safe weakly supervised learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 1, pp. 334–346, 2021.

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

[45] I. Martin and A. Mesaros, "Diversity and bias in audio captioning datasets," in *Proceedings of the 6th Detection and Classification of Acoustic Scenes and Events 2021 Workshop*, Barcelona, Spain, November 2021, pp. 90–94.

[46] H. Agrawal, K. Desai, Y. Wang, X. Chen, R. Jain, M. Johnson, D. Batra, D. Parikh, S. Lee, and P. Anderson, "NoCaps: Novel object captioning at scale," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 8948–8957.

[47] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *Transactions of the Association for Computational Linguistics*, vol. 2, pp. 67–78, 02 2014.

[48] X. Mei, X. Liu, Q. Huang, M. D. Plumbley, and W. Wang, "Audio captioning Transformer," in *Proceedings of the 6th Detection and Classification of Acoustic Scenes and Events 2021 Workshop*, Barcelona, Spain, November 2021, pp. 211–215.

[49] X. Liu, X. Mei, Q. Huang, J. Sun, J. Zhao, H. Liu, M. D. Plumbley, V. Kilic, and W. Wang, "Leveraging pre-trained BERT for audio captioning," in *30th European Signal Processing Conference*. IEEE, 2022, pp. 1145–1149.

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

[51] M. Wu, H. Dinkel, and K. Yu, "Audio caption: Listen and tell," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2019, pp. 830–834.

[52] S. Deshmukh, B. Elizalde, and H. Wang, "Audio retrieval with Wav-Text5K and CLAP training," in *Proc. Interspeech*. ISCA, 2022.

[53] 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*, 2023.

[54] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of Machine Learning Research*, vol. 21, no. 140, pp. 1–67, 2020.

[55] 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.

[56] S. Hershey, D. P. W. Ellis, E. Fonseca, A. Jansen, C. Liu, R. Channing Moore, and M. Plakal, "The benefit of temporally-strong labels

in audio event classification," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2021, pp. 366–370.

[57] Anonymous, "Listen, think, and understand," in *The Twelfth International Conference on Learning Representations*, 2024.

[58] L. Sun, X. Xu, M. Wu, and W. Xie, "A large-scale dataset for audio-language representation learning," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2024.

[59] S. Doh and J. Nam, "Lp-musiccaps: Llm-based pseudo music captioning," in *24th International Society for Music Information Retrieval Conference*. International Society for Music Information Retrieval Conference, 2023.

[60] D. Yang, J. Tian, X. Tan, R. Huang, S. Liu, X. Chang, J. Shi, S. Zhao, J. Bian, X. Wu *et al.*, "Uniaudio: An audio foundation model toward universal audio generation," *arXiv preprint arXiv:2310.00704*, 2023.

[61] H. Liu, Y. Yuan, X. Liu, X. Mei, Q. Kong, Q. Tian, Y. Wang, W. Wang, Y. Wang, and M. D. Plumbley, "Audioldm 2: Learning holistic audio generation with self-supervised pretraining," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 2871–2883, 2024.

[62] D. Ghosal, N. Majumder, A. Mehrish, and S. Poria, "Text-to-audio generation using instruction guided latent diffusion model," in *Proceedings of the 31st ACM International Conference on Multimedia*. New York, NY, USA: Association for Computing Machinery, 2023, p. 3590–3598.

[63] Anonymous, "SALMONN: Towards generic hearing abilities for large language models," in *The Twelfth International Conference on Learning Representations*, 2024.

[64] Z. Zhao, L. Guo, T. Yue, S. Chen, S. Shao, X. Zhu, Z. Yuan, and J. Liu, "Chatbridge: Bridging modalities with large language model as a language catalyst," *arXiv preprint arXiv:2305.16103*, 2023.

[65] J. Liang, X. Liu, W. Wang, M. D. Plumbley, H. Phan, and E. Benetos, "Acoustic prompt tuning: Empowering large language models with audition capabilities," *arXiv preprint arXiv:2312.00249*, 2023.

[66] F. Font, G. Roma, and X. Serra, "Freesound technical demo," in *Proceedings of the 21st ACM International Conference on Multimedia*, ser. MM '13. New York, NY, USA: Association for Computing Machinery, 2013, p. 411–412.

[67] J. P. Kincaid, R. P. Fishburne Jr, R. L. Rogers, and B. S. Chissom, "Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel," Defense Technical Information Center, Fort Belvoir, VA, Tech. Rep., 1975.

[68] A. Nagrani, P. H. Seo, B. Seybold, A. Hauth, S. Manen, C. Sun, and C. Schmid, "Learning audio-video modalities from image captions," in *European Conference on Computer Vision*. Springer, 2022, pp. 407–426.

[69] Y. Xin, D. Yang, and Y. 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, pp. 1–5.

[70] 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*, 2022, pp. 646–650.

[71] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*, 2019, pp. 4171–4186.

[72] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International Conference on Machine Learning*. PMLR, 2020, pp. 1597–1607.

[73] 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*, 2023.

[74] D. P. Kingma and J. Ba, "ADAM: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[75] 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*. IEEE, 2023.

[76] Z. Ye, H. Wang, D. Yang, and Y. Zou, "Improving the performance of automated audio captioning via integrating the acoustic and semantic information," in *Proceedings of the 6th Detection and Classification of Acoustic Scenes and Events 2021 Workshop*, Barcelona, Spain, November 2021, pp. 40–44.

[77] Q. Han, W. Yuan, D. Liu, X. Li, and Z. Yang, "Automated audio captioning with weakly supervised pre-training and word selection methods," in *Proceedings of the 6th Detection and Classification of Acoustic Scenes and Events 2021 Workshop*, Barcelona, Spain, November 2021, pp. 6–10.

[78] Z. Ye, Y. Wang, H. Wang, D. Yang, and Y. Zou, "FeatureCut: An adaptive data augmentation for automated audio captioning," in *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, 2022, pp. 313–318.

[79] C. Chen, N. Hou, Y. Hu, H. Zou, X. Qi, and E. S. Chng, "Inter-active audio-text representation for automated audio captioning with contrastive learning," in *Proc. Interspeech*. ISCA, 2022.

[80] X. Liu, Q. Huang, X. Mei, H. Liu, Q. Kong, J. Sun, S. Li, T. Ko, Y. Zhang, L. H. Tang *et al.*, "Visually-aware audio captioning with adaptive audio-visual attention," in *Proc. Interspeech*. ISCA, 2023.

[81] E. Kim, J. Kim, Y. Oh, K. Kim, M. Park, J. Sim, J. Lee, and K. Lee, "Improving audio-language learning with MixGen and multi-level test-time augmentation," *arXiv preprint arXiv:2210.17143*, 2022.

[82] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Jul. 2020, pp. 7871–7880.

[83] 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.

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

[85] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," *Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pp. 65–72, 2005.

[86] 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.

[87] 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.

[88] 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.

[89] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *22nd ACM International Conference on Multimedia*, Orlando, FL, USA, Nov. 2014, pp. 1041–1044.

[90] H. Chen, W. Xie, A. Vedaldi, and A. Zisserman, "VGGSound: A large-scale audio-visual dataset," in *International Conference on Acoustics, Speech, and Signal Processing*, 2020.

[91] S. Chen, Y. Wu, C. Wang, S. Liu, D. Tompkins, Z. Chen, W. Che, X. Yu, and F. Wei, "BEATs: Audio pre-training with acoustic tokenizers," in *Proceedings of the 40th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., vol. 202. PMLR, 23–29 Jul 2023, pp. 5178–5193.

[92] A. Gazneli, G. Zimerman, T. Ridnik, G. Sharir, and A. Noy, "End-to-end audio strikes back: Boosting augmentations towards an efficient audio classification network," *arXiv preprint arXiv:2204.11479*, 2022.

[93] H.-H. Wu, P. Seetharaman, K. Kumar, and J. P. Bello, "Wav2CLIP: Learning robust audio representations from CLIP," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2022, pp. 4563–4567.

[94] A. Guzhov, F. Raue, J. Hees, and A. Dengel, "AudioCLIP: Extending CLIP to image, text and audio," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2022, pp. 976–980.

[95] X. Xu, Z. Zhang, Z. Zhou, P. Zhang, Z. Xie, M. Wu, and K. Q. Zhu, "BLAT: Bootstrapping language-audio pre-training based on audioset tag-guided synthetic data," in *Proceedings of the 31st ACM International Conference on Multimedia*. New York, NY, USA: Association for Computing Machinery, 2023, p. 2756–2764.

[96] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.
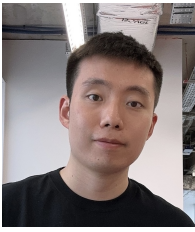
Xinhao Mei received the B.Eng. degree from the Southwest Petroleum University, Chengdu, China, in 2019, and the M.S. degree from the University of Surrey, Guildford, U.K., in 2020. He is currently working toward the Ph.D. degree with the Centre for Vision, Speech, and Signal Processing (CVSSP), University of Surrey, Guildford, U.K. His research interests include understanding and generation of audio using deep learning methods.

Haohe Liu is a PhD student at the Centre for Vision Speech and Signal Processing (CVSSP), University of Surrey, UK, and previously obtained his B.Eng degree from Northwestern Polytechnical University, Xi'an, China in 2020. His past research has contributed to the fields of audio quality enhancement, audio generation, source separation, and audio recognition. Haohe is mostly known for developing AudioLDM for text-to-audio generation, generating wide attention in the open-source community. His work as the primary author has been accepted in top journals and conferences such as TPAMI, TASLP, ICML, AAAI, ICASSP, and INTERSPEECH. Notable recent projects include AudioLDM, VoiceFixer, AudioSR, and NaturalSpeech.
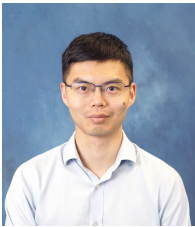
Qiuqiang Kong is currently an assistant professor at the EE department of the Chinese University of Hong Kong. He received his Ph.D. degree from the University of Surrey, Guildford, UK, in 2019. Following his Ph.D., he joined ByteDance as a research scientist. His research topic includes the classification, detection, separation, and generation of general sounds and music. He was the top 2% scientist in 2021 in "Updated science-wide author databases of standardized citation indicators. He was known for developing pretrained audio neural networks (PANNs) for audio tagging and was awarded the IEEE SPS Young Author Best Paper in 2023. He won the detection and classification of acoustic scenes and events (DCASE) challenge in 2017. He was known for transcribing the largest piano MIDI dataset GiantMIDI-Piano in the world. He has co-authored over 50 papers in journals and conferences, including IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP), ICASSP, INTERSPEECH, IJCAI, DCASE, EUSIPCO, LVA-ICA. He has been cited 3156 times, with an H-index of 28 till Sep. 2023. He was a frequent reviewer for world well known journals and conferences, including TASLP, TMM, SPL, TKDD, JASM, EURASIP, Neurocomputing, Neural Networks, ISMIR, CSMT. He assisted with organizing the LVA-ICA 2018 in Guildford, UK and the DCASE 2018 Workshop in Woking, UK.
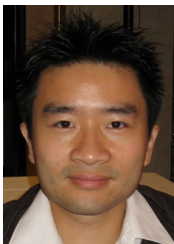
Tom, Ko Yu Ting received the Ph.D. degree in computer science and engineering from the Hong Kong University of Science and Technology in 2014. After his graduation, he joined Huawei Noah's Ark Lab as a research scientist. In 2019, he worked as an assistant professor at Southern University of Science and Technology in China. In 2021, he joined ByteDance AI as a research scientist. His research interests include speech processing and natural language processing.

Mark D. Plumbley (S'88-M'90-SM'12-F'15) received the B.A.(Hons.) degree in electrical sciences and the Ph.D. degree in neural networks from University of Cambridge, Cambridge, U.K., in 1984 and 1991, respectively. He is Professor of Signal Processing at the Centre for Vision, Speech, and Signal Processing at the University of Surrey, Guildford, U.K. His current research concerns AI, machine learning and signal processing for analysis, recognition and generation of sound. He led the first international data challenge on Detection and Classification of Acoustic Scenes and Events, and currently holds an Engineering and Physical Sciences Research Council Fellowship on "AI for Sound". He is a Member of the IEEE Signal Processing Society Technical Committee on Audio and Acoustic Signal Processing, and a Fellow of the IET and IEEE.

Yuexian Zou (Senior Member, IEEE) received the B.S. degree from the University of Electronic Science and Technology of China in 1985 and the Ph.D. degree from The University of Hong Kong in 2001. She is currently a Full Professor with Peking University and the Director of the Advanced Data and Signal Processing Laboratory, Peking University. She serves as the Deputy Director for the Shenzhen Association of Artificial Intelligence (SAAI). Since 2010, she has been actively involved in teaching and research on machine learning and its applications in video and audio analysis. She has conducted more than 20 research projects including NSFC and 863 projects, published more than 300 academic papers in famous journals and flagship conferences, and issued nine invention patents. Her current research interests are mainly in cross-modality learning and AI education.

Wenwu Wang was born in Anhui, China. He received the B.Sc., M.E., and the Ph.D. degrees 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, 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 300 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 and 2023 Judge's Award, DCASE 2019 and 2020 Reproducible System Award, and LVA/ICA 2018 Best Student Paper Award. He is an Associate Editor (2020-2025) for IEEE/ACM Transactions on Audio Speech and Language Processing and 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. He is the elected Chair (2023-2024) of IEEE Signal Processing Society (SPS) Machine Learning for Signal Processing Technical Committee, a Board Member (2023-2024) of IEEE SPS Technical Directions Board, the elected Chair (2025-2027) and Vice Chair (2022-2024) of the EURASIP Technical Area Committee on Acoustic Speech and Music Signal Processing, an elected Member (2021-2026) of the IEEE SPS Signal Processing Theory and Methods Technical Committee. He was a Satellite Workshop Co-Chair for IEEE ICASSP 2024 and INTERSPEECH 2022, a Publication Co-Chair for IEEE ICASSP 2019, Local Arrangement Co-Chair of IEEE MLSP 2013, and Publicity Co-Chair of IEEE SSP 2009. He is a Special Session Co-Chair of IEEE MLSP 2024, and Technical Program Co-Chair of IEEE MLSP 2025.

Chutong Meng biography not available at the time of publication.

Chengqi Zhao biography not available at the time of publication.