AudioSetCaps: Enriched Audio Captioning Dataset Generation Using Large Audio Language Models

Jisheng Bai[123*] **Haohe Liu**[4*] **Mou Wang**[5] **Dongyuan Shi**[1]
Wenwu Wang[4] **Mark D. Plumbley**[4] **Woon-Seng Gan**[3] **Jianfeng Chen**[1]
[1]Northwestern Polytechnical University [2]Xi'an Lianfeng Acoustic Technologies Co., Ltd.
[3]SNTL, Nanyang Technological University [4]CVSSP, University of Surrey
[5]Institute of Acoustics, Chinese Academy of Sciences

Abstract

Building large-scale audio-language datasets is crucial yet challenging for training audio-language models, primarily due to its time-consuming and labour-intensive nature. Although large language models (LLMs) have greatly enhanced the efficiency of this process, current LLM-based pipelines for generating audio-text data still lack the capability to incorporate detailed audio information. In this paper, we propose a novel pipeline leveraging large audio-language models to automatically generate large-scale, fine-grained audio captions. Based on this approach, we create AudioSetCaps, a dataset comprising 1.9 million audio-caption pairs derived from recordings in AudioSet. We evaluate AudioSetCaps on two downstream tasks: audio-text retrieval and automated audio captioning. Models trained with AudioSetCaps achieve state-of-the-art performance on both tasks, demonstrating the high quality of the generated captions. Notably, our proposed data-labelling pipeline employs open-source APIs and can run on a consumer-grade GPU. To facilitate further advancements in this field, we have made our code, audio-caption paired data, and pre-trained models on downstream tasks publicly available at `https://github.com/JishengBai/AudioSetCaps`.

1 Introduction

Audio-language modeling has gained significant attention in recent years, substantially advancing the field of audio perception [23, 24, 28, 3, 22]. This progress primarily relies on the development of comprehensive audio-language datasets [37, 27]. However, current audio-language models (ALMs) face challenges in achieving robust universal audio-language representations and approximating human-like audio understanding. These limitations stem largely from constraints in both the quantity and quality of audio-text data available for training. Creating large-scale, high-quality audio-language datasets is hindered by substantial time and labour costs, impeding progress in this field.

Recent efforts leverage LLMs to automate the construction of large-scale audio-language datasets, as summarized in Table 1. LAION-Audio-630K [38] collects raw data from online sources and uses sentence templates or keyword-to-caption methods to transform tags into captions. WavCaps [29] employs ChatGPT to refine unprocessed manual audio descriptions into more structured captions. Auto-ACD [33] and Sound-VECaps [40] incorporate visual cues with audio information as input for LLMs in caption generation. While these approaches make significant progress, they each face specific challenges. LAION-Audio-630K [38] contains noisy web-harvested descriptions and suffers from imbalanced data distribution, potentially affecting caption quality. WavCaps [29] generates captions that primarily rephrase basic audio labels, lacking comprehensive acoustic information [33].

[*] Equal contribution

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

Dataset	Quantity	Length	Vocabulary	Source
AudioCaps [17]	57K	9	5K	H
Clotho [10]	30k	11	4K	H
LAION-Audio-630K [38]	630K	7	311K	L
WavCaps [29]	400K	8	24K	L
Auto-ACD [33]	1.5M	18	20K	A+V+L
Sound-VECaps [40]	1.6M	40	50K	A+V+L
AudioSetCaps	1.9M	28	21K	A+L

Table 1: The statistics comparison with popular audio-language datasets. Length: average caption length; Vocabulary: vocabulary size of caption. Caption source: H (human), A (audio models), V (visual models), L (language models).

Auto-ACD [33] and Sound-VECaps [40] require additional visual data for audio caption generation, potentially complicating, and introducing noise to the caption-generating process.

The emergence of large ALMs [39, 8, 13] has significantly enhanced the ability to comprehend acoustic content, demonstrating strong performance on various downstream tasks. Nevertheless, the scarcity of large-scale, unified audio-text pairs across different tasks presents an opportunity for further improvement in the audio context understanding capabilities of ALMs. To address this limitation, we propose an automated pipeline for generating enriched audio-language datasets leveraging both ALMs and LLMs. Specifically, our approach uses the Qwen-Audio [7] ALM to extract fine-grained audio content and implements Mistral-7B LLM [16] to merge various audio elements and generate natural, detailed captions. We introduce filtering and refining procedures to improve the quality of generated captions. With the proposed pipeline, we generate over 1.9 million in captions for recordings in AudioSet [12], which we refer to as AudioSetCaps. Models trained on AudioSetCaps demonstrate superior performance in audio-text retrieval (ATR) and automated audio captioning (AAC) tasks compared to models trained on datasets generated by previous LLM-based approaches. These results validate the higher quality of our dataset and the effectiveness of our data labelling pipeline. Furthermore, we have extended our pipeline to caption over 4.0 million audio samples from YouTube-8M [1]. To facilitate advancements in audio-language learning, we have made the pipeline and the audio caption datasets publicly available.

2 Large Audio and Language Models

2.1 Large Audio-language Models

Qwen-Audio [7] is a fundamental multi-task audio-language model that supports a wide range of tasks, serving as a universal audio understanding model. Building upon Qwen-Audio, Qwen-Audio-Chat [7] can follow instructions, enable multi-turn dialogues, and support diverse downstream audio tasks. Qwen-Audio achieves state-of-the-art performance across tasks in various audio tasks [7], including speech recognition (with a word error rate of 1.3% on AISHELL-1 [5]), vocal sound classification (with an accuracy of 93% on VocalSound [14]), and acoustic scene classification (with an accuracy of 79.5% on Cochlscene [15]).

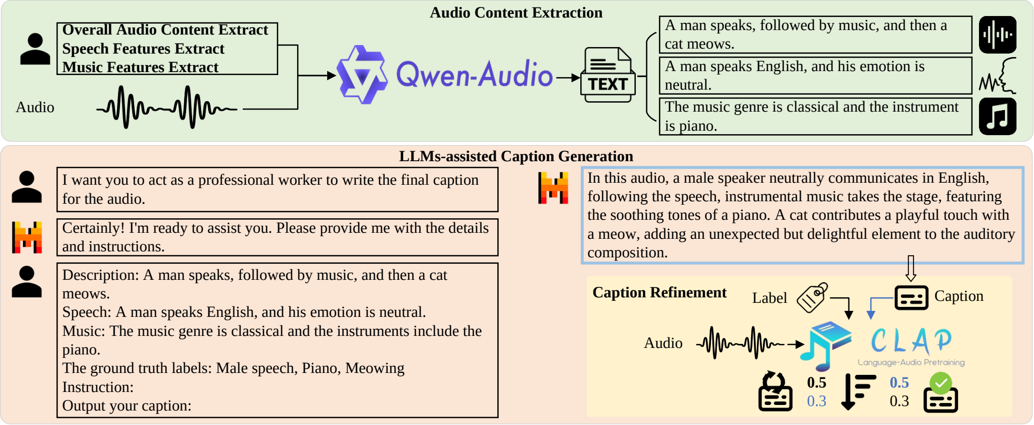
Contrastive Language-Audio Pretraining (CLAP) models [38, 11] have shown significant performance in enhancing audio understanding and retrieval tasks by integrating audio and language modalities. The LAION CLAP [38] model employs feature fusion and keyword-to-caption augmentation to handle variable-length audio inputs. LAION CLAP uses HTSAT [6] as the audio encoder and RoBERTa [26] as the text model to project features into a shared latent space. This approach [38] achieved state-of-the-art performance in text-to-audio retrieval with retrieval at rank 1 of 46.8 on AudioCaps [17], and zero-shot audio classification with accuracies of 91% and 77% on ESC-50 [30] and UrbanSound8K [32], respectively.

2.2 Large Language Model

Mistral-7B [16] is a 7.3 billion parameter language model that demonstrates good performance across various benchmarks, outperforming larger models like Llama 2 13B [35] and Llama 1 34B [34] in many areas. Mistral-7B employs grouped-query and sliding window attention methods to improve the inference efficiency of long sequences, facilitating its deployment across various tasks.

3 Data Collection

This section introduces our caption collection pipeline, shown in Figure 1. The pipeline comprises three parts: audio content extraction, LLMs-assisted caption generation, and caption refinement.



A man speaks, followed by music, and then a cat meows.

A man speaks English, and his emotion is neutral.

The music genre is classical and the instrument is piano.

I want you to act as a professional worker to write the final caption for the audio.

Certainly! I'm ready to assist you. Please provide me with the details and instructions.

Description: A man speaks, followed by music, and then a cat meows.
Speech: A man speaks English, and his emotion is neutral.
Music: The music genre is classical and the instruments include the piano.
The ground truth labels: Male speech, Piano, Meowing
Instruction:
Output your caption:

In this audio, a male speaker neutrally communicates in English, following the speech, instrumental music takes the stage, featuring the soothing tones of a piano. A cat contributes a playful touch with a meow, adding an unexpected but delightful element to the auditory composition.

Figure 1: Overview of the proposed automated pipeline for audio caption generation.

3.1 Audio Content Extraction

Audio Content Extraction aims to extract detailed, fine-grained, and diverse audio content from audio recordings. We leverage Qwen-Audio ALM to facilitate this process because Qwen-Audio achieves state-of-the-art performance on speech, music, and audio tasks [7]. Specifically, we utilize Qwen-Audio-Chat, which accepts both audio and language inputs and produces language outputs as answers, thus offering audio understanding and reasoning capabilities [7]. However, Qwen-Audio struggles to process complex prompts. For example, when we ask Qwen-Audio to describe the speech, music, and sound content by using a long and complex prompt, it tends to overlook the requirements shown in the prompt, similar to the instruction following issue observed in the textual large language model [31].

To address this constraint and maximize the effectiveness of Qwen-Audio, we design a structured prompt chain, utilizing a pre-defined set of prompts to perform both general and detailed audio reasoning and understanding, including sound, speech, and music content. Initially, we prompt Qwen-Audio to analyze the overall audio content, imposing a 50-word limit to ensure focus on essential audio elements while avoiding extraneous details. Subsequently, we prompt Qwen-Audio to provide a focused analysis of specific speech features, including emotion, gender, and language, contingent on the presence of speech in the audio. Finally, we adopt a similar prompt used for speech content extraction to instruct Qwen-Audio in identifying music genres and instruments.

3.2 LLMs-assisted Caption Generation

To generate accurate and concise audio captions, we employ the Mistral-7B LLM with a specifically structured prompt. We begin by prompting the LLM to play as a professional audio caption annotator, emphasizing the critical importance of following the provided instructions. We then provide Mistral-7B with detailed contextual information extracted from the audio content with Qwen-Audio, including general content, speech characteristics, and musical analysis results. Additionally, we provide ground truth labels as references for the Mistral-7B.

To improve caption quality, we instruct Mistral-7B to exclude specific speech content and ground truth labels while generating the captions. We also guide Mistral-7B to consolidate the detailed audio content with a 50-word limitation. These approaches aim to generate concise and relevant captions.

3.3 Caption Refinement

While ALMs and LLMs demonstrate strong capabilities in extracting audio content and formulating descriptions into captions, they occasionally fail to generate captions that fully meet all specified requirements. To address this limitation, we have implemented a caption refinement module designed to enhance the quality of the generated captions. We use LAION CLAP to assess caption quality by comparing the audio-language similarity between the audio and generated captions or ground truth labels. Captions with a similarity score lower than their corresponding labels are discarded and regenerated. This iterative process continues until we obtain captions accurately representing the audio content.

4 Experiments

4.1 Dataset Statistics

As Table 1 shows, we compare AudioSetCaps with other popular audio-language datasets based on quantity, average caption length, caption vocabulary size, and caption source. AudioSetCaps contains about 1.9 million audio-caption pairs, based on AudioSet [12] recordings, which is more than the other datasets in Table 1. The average caption in AudioSetCaps is 28 words long, which is less than Sound-VECaps (40 words) but longer than most others. While its vocabulary of 21K words is smaller than LAION-Audio-630K (311K), it still offers good variety. Notably, AudioSetCaps stands as the only audio-language dataset that leverages audio and textual language models in its creation process. In comparison to LAION-Audio-630K [38] and WavCaps [29], AudioSetCaps offers more informative captions with speech and music content. Unlike Auto-ACD [33] and Sound-VECaps [40], which incorporate visual models, AudioSetCaps focuses on solely incorporating audio information, thereby avoiding the potential introduction of complex visual models and associated noisy content.

4.2 Audio-text Retrieval

Method	Training Set	Model	T2A		A2T	
			R@1	R@10	R@1	R@10
LAION [38]	LA+AC+Clotho	HTSAT+RoBERTa-FT	36.1	83.9	46.8	90.7
WavCaps [29]	AC+Clotho	HTSAT+BERT	39.2	86.5	49.5	91.5
	WC+AC+Clotho	HTSAT+BERT-PT	39.7	86.1	51.7	90.6
	WC+AC+Clotho	HTSAT+BERT-FT	42.2	87.1	54.6	92.4
Auto-ACD [33]	ACD+AC+Clotho	HTSAT+RoBERTa-PT	39.5	85.4	53.7	91.7
	ACD+AC+Clotho	HTSAT+RoBERTa-FT	42.7	**88.5**	56.3	**93.9**
Sound-VECaps [40]	Sound-VECapsF	HTSAT+RoBERTa	39.2	85.0	54.0	93.2
	Sound-VECapsA	HTSAT+RoBERTa	41.2	85.3	53.3	93.0
AudioSetCaps	ASC+AC+Clotho	HTSAT+BERT-PT	39.8	85.7	54.2	92.4
	ASC+AC+Clotho	HTSAT+BERT-FT	**43.4**	88.2	**57.3**	93.2

Table 2: Performance comparison of audio-text retrieval on the AudioCaps test set. “LA”, “AC”, “WC”, “ACD”, and “ASC” denote LAION-Audio-630K, AudioCaps, WavCaps, Auto-ACD, and AudioSetCaps, respectively. “PT” represents pre-training on the training set and “FT” represents fine-tuning on the AudioCaps training set. “T2A” and “A2T” refer to text-to-audio and audio-to-text retrieval, respectively. “R@1” and “R@10” denote recall at ranks 1 and 10.

We adopt the models used in WavCaps [29] to build our model for audio-text retrieval (ATR) task. This model combines a pre-trained HTSAT [6], a pre-trained BERT-base network[9], and a 2-layer multilayer perceptron. We first pre-train the model on a merged dataset comprising AudioSetCaps and the training sets of AudioCaps and Clotho. The ATR model is trained for 40 epochs with a batch size of 196 and a learning rate of 5×10^{-5}, and fine-tuned on AudioCaps with a learning rate of 1×10^{-5} for 20 epochs. During training, the validation and test sets of AudioCaps are excluded. The evaluation metric for ATR is recall at rank k (R@k), averaged across the dataset.

Table 2 presents the ATR results on the AudioCaps test dataset. Our HTSAT and BERT-based ATR model, pre-trained on a combination of AudioCaps, Clotho, and AudioSetCaps, outperforms models pre-trained using WavCaps [29] or Auto-ACD [33] combined with AudioCaps and Clotho. We further fine-tuned our ATR model on the AudioCaps training dataset. This fine-tuned model achieves a text-to-audio R@1 of 43.4 and an audio-to-text R@1 of 57.3, surpassing the performance of ATR models trained on other datasets.

Method	Training Set	Model	METEOR	CIDEr	SPICE	SPIDEr
AL-MixGen [25]	AC	ACT [28]	24.2	76.9	18.1	47.5
CoNeTTE [20]	AudioSet	CNext-Trans	-	-	-	49.5
EnCLAP [19]	AC	EnCLAP-Base	24.7	78.0	18.6	48.3
	AC	EnCLAP-Large	25.5	80.3	**18.8**	49.5
Wavcaps [29]	AC+Clotho	HTSAT+BART	23.7	71.1	17.7	44.4
	WC+AC+Clotho	HTSAT+BART	25.0	78.7	18.2	48.5
AudioSetCaps	ASC+AC+Clotho	HTSAT+BART	**25.7**	**84.8**	18.4	**51.6**

Table 3: The performance of different methods for automated audio captioning on AudioCaps test set, where "ACT" and "CNext-Trans" refer to Audio Captioning Transformer and ConvNeXt-Transformer.

4.3 Automated Audio Captioning

We also adopt the automated audio captioning (AAC) models used in [29], which are based on the pre-trained HTSAT and BART [21] models. We first pre-trained the AAC model on AudioSetCaps combined with the training sets of Clotho and AudioCaps, using a learning rate of 5×10^{-5} and a batch size of 24 for 15 epochs. The pre-trained AAC model is further fine-tuned on AudioCaps for 20 epochs with a learning rate of 5×10^{-6}. The performance is evaluated using conventional AAC metrics, including METEOR [4], CIDEr [36], SPICE [2] and SPIDEr [25].

Table 3 presents the AAC results on the AudioCaps test dataset. Our AAC model achieves state-of-the-art performance with a METEOR of 25.7, CIDEr of 84.8, SPICE of 18.4, and SPIDEr of 51.6. Compared to training using only AudioCaps and Clotho, the AAC model trained with the additional combination of AudioSetCaps achieves better performance than when trained with the additional combination of WavCaps. Moreover, our method outperforms other SOTA approaches such as AL-MixGen [18], CoNeTTE [20], and EnCLAP [19], which use different model architectures.

4.4 Subjective Evaluation

Caption Data	Label	AudioCaps	WavCaps	AudioSetCaps	Auto-ACD
Mean Score	3.81	3.79	3.70	3.70	3.60

Table 4: Mean Scores of human evaluation across datasets. The scores indicate how well the text annotation reflects the audio content based on the following scale: 1-Bad, 2-Poor, 3-Fair, 4-Good, 5-Excellent.

To further assess the quality of AudioSetCaps, we perform subjective evaluation to evaluate captions corresponding to 79 randomly selected overlapping audio samples across AudioCaps [17], WavCaps [29], AudioSetCaps, and Auto-ACD [33], as well as the ground truth labels in AudioSet [12]. The question we ask for the rater is *Please listen to the provided audio samples and rate the quality of the text annotation based on its accuracy, completeness, and presence of false information.* Each caption or label is scored by at least 5 evaluators on how well the text annotation reflects the audio content. The score is on a scale from 1 to 5, where 1 represents "Bad" and 5 represents "Excellent". The mean scores for each dataset and the ground truth labels are presented in Table 4. AudioSetCaps achieves a mean score of 3.70, which closely approaches the quality of both the ground truth labels and human-labelled captions from AudioCaps.

5 Conclusion

In this paper, we have presented an automated pipeline for generating enriched audio-text data with fine-grained audio content, leveraging audio-language models and large language models. Using this pipeline, we presented AudioSetCaps, a dataset consisting of over 1.9 million audio captioning data, aiming to advance the field of audio-language learning. We conducted experiments on audio-text retrieval and automated audio captioning tasks using AudioSetCaps. The experimental results demonstrate that the quality of our data surpasses that of previous pipelines. To further encourage the development of audio-language learning, we will release the pipeline codes, audio-caption datasets, and pre-trained downstream models to the research community.

Acknowledgment

We acknowledge the support provided by the China Scholarship Council during a visit of Jisheng Bai to Nanyang Technological University. This research was partly supported by Engineering and Physical Sciences Research Council (EPSRC) Grant EP/T019751/1 “AI for Sound”, a Research Gift from Adobe, and a PhD scholarship from the Centre for Vision, Speech and Signal Processing (CVSSP) at the University of Surrey and BBC R&D.

References

[1] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. YouTube-8M: A Large-Scale Video Classification Benchmark. *arXiv preprint:1609.08675*, 2016.

[2] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. SPICE: Semantic Propositional Image Caption Evaluation. In *European Conference on Computer Vision*, pages 382–398, 2016.

[3] Jisheng Bai, Han Yin, Mou Wang, Dongyuan Shi, Woon-Seng Gan, Jianfeng Chen, and Susanto Rahardja. AudioLog: LLMs-Powered Long Audio Logging with Hybrid Token-Semantic Contrastive Learning. *arXiv preprint:2311.12371*, 2023.

[4] Satanjeev Banerjee and Alon Lavie. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, 2005.

[5] Hui Bu, Jiayu Du, Xingyu Na, Bengu Wu, and Hao Zheng. AISHELL-1: An open-source Mandarin speech corpus and a speech recognition baseline. In *20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA)*, pages 1–5. IEEE, 2017.

[6] Ke Chen, Xingjian Du, Bilei Zhu, Zejun Ma, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. HTS-AT: A Hierarchical Token-Semantic Audio Transformer for Sound Classification and Detection. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 646–650. IEEE, 2022.

[7] Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. Qwen-Audio: Advancing Universal Audio Understanding via Unified Large-Scale Audio-Language Models. *arXiv preprint:2311.07919*, 2023.

[8] Soham Deshmukh, Benjamin Elizalde, Rita Singh, and Huaming Wang. Pengi: An Audio Language Model for Audio Tasks. In *Proceedings of the 37th Conference on Neural Information Processing Systems*, 2023.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina 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*, pages 4171–4186, 2019.

[10] Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. Clotho: an Audio Captioning Dataset. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 736–740. IEEE, 2020.

[11] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. CLAP Learning Audio Concepts from Natural Language Supervision. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1–5. IEEE, 2023.

[12] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio Set: An ontology and human-labeled dataset for audio events. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 776–780. IEEE, 2017.

[13] Yuan Gong, Hongyin Luo, Alexander H. Liu, Leonid Karlinsky, and James R. Glass. Listen, Think, and Understand. In *Proceedings of the 12th International Conference on Learning Representations*, 2024.

[14] Yuan Gong, Jin Yu, and James Glass. Vocalsound: A Dataset for Improving Human Vocal Sounds Recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 151–155. IEEE, 2022.

[15] Il-Young Jeong and Jeongsoo Park. CochlScene: Acquisition of acoustic scene data using crowdsourcing. In *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 17–21. IEEE, 2022.

[16] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7B. *arXiv preprint:2310.06825*, 2023.

[17] Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee 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*, pages 119–132, 2019.

[18] Eungbeom Kim, Jinhee Kim, Yoori Oh, Kyungsu Kim, Minju Park, Jaeheon Sim, Jinwoo Lee, and Kyogu Lee. Exploring Train and Test-Time Augmentations for Audio-Language Learning. *arXiv preprint:2210.17143*, 2022.

[19] Jaeyeon Kim, Jaeyoon Jung, Jinjoo Lee, and Sang Hoon Woo. EnCLAP: Combining Neural Audio Codec and Audio-Text Joint Embedding for Automated Audio Captioning. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6735–6739. IEEE, 2024.

[20] Étienne Labbé, Thomas Pellegrini, and Julien Pinquier. CoNeTTE: An Efficient Audio Captioning System Leveraging Multiple Datasets With Task Embedding. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:3785–3794, 2024.

[21] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke 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*, pages 7871–7880, July 2020.

[22] Dongting Li, Chenchong Tang, and Han Liu. Audio-LLM: Activating the Capabilities of Large Language Models to Comprehend Audio Data. In *International Symposium on Neural Networks*, pages 133–142, 2024.

[23] Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley. AudioLDM: Text-to-Audio Generation with Latent Diffusion Models. In *Proceedings of the 40th International Conference on Machine Learning*, pages 21450–21474, 2023.

[24] Haohe Liu, Yi Yuan, Xubo Liu, Xinhao Mei, Qiuqiang Kong, Qiao Tian, Yuping Wang, Wenwu Wang, Yuxuan Wang, and Mark D. Plumbley. AudioLDM 2: Learning Holistic Audio Generation With Self-Supervised Pretraining. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:2871–2883, 2024.

[25] Siqi Liu, Zhenhai Zhu, Ning Ye, Sergio Guadarrama, and Kevin Murphy. Improved Image Captioning via Policy Gradient optimization of SPIDEr. In *IEEE International Conference on Computer Vision*, pages 873–881, 2017.

[26] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint:1907.11692*, 2019.

[27] Irene Martin Morato and Annamaria Mesaros. Diversity and bias in audio captioning datasets. In *Proceedings of the 6th Workshop on Detection and Classication of Acoustic Scenes and Events*, pages 90–94, November 2021.

[28] Xinhao Mei, Xubo Liu, Qiushi Huang, Mark D Plumbley, and Wenwu Wang. Audio Captioning Transformer. *arXiv preprint:2107.09817*, 2021.

[29] Xinhao Mei, Chutong Meng, Haohe Liu, Qiuqiang Kong, Tom Ko, Chengqi Zhao, Mark D. Plumbley, Yuexian Zou, and Wenwu Wang. WavCaps: A ChatGPT-Assisted Weakly-Labelled Audio Captioning Dataset for Audio-Language Multimodal Research. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:3339–3354, 2024.

[30] Karol J Piczak. ESC: Dataset for Environmental Sound Classification. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 1015–1018, 2015.

[31] Yiwei Qin, Kaiqiang Song, Yebowen Hu, Wenlin Yao, Sangwoo Cho, Xiaoyang Wang, Xuansheng Wu, Fei Liu, Pengfei Liu, and Dong Yu. Infobench: Evaluating instruction following ability in large language models. *arXiv preprint:2401.03601*, 2024.

[32] Justin Salamon, Christopher Jacoby, and Juan Pablo Bello. A Dataset and Taxonomy for Urban Sound Research. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 1041–1044, 2014.

[33] Luoyi Sun, Xuenan Xu, Mengyue Wu, and Weidi Xie. Auto-ACD: A Large-scale Dataset for Audio-Language Representation Learning. In *ACM Multimedia*, 2024.

[34] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. LLaMA: Open and Efficient Foundation Language Models. *arXiv preprint:2302.13971*, 2023.

[35] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv preprint:2307.09288*, 2023.

[36] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. CIDEr: Consensus-based Image Description Evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4566–4575, 2015.

[37] Gijs Wijngaard, Elia Formisano, Michele Esposito, and Michel Dumontier. Audio-Language Datasets of Scenes and Events: A Survey. *arXiv preprint:2407.06947*, 2024.

[38] Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-Scale Contrastive Language-Audio Pretraining with Feature Fusion and Keyword-to-Caption Augmentation. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1–5. IEEE, 2023.

[39] Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun Lv, Zhou Zhao, Chang Zhou, et al. AIR-Bench: Benchmarking Large Audio-Language Models via Generative Comprehension. *arXiv preprint:2402.07729*, 2024.

[40] Yi Yuan, Dongya Jia, Xiaobin Zhuang, Yuanzhe Chen, Zhengxi Liu, Zhuo Chen, Yuping Wang, Yuxuan Wang, Xubo Liu, Mark D Plumbley, et al. Improving Audio Generation with Visual Enhanced Caption. *arXiv preprint:2407.04416*, 2024.