Sound-VECaps: Improving Audio Generation with Visually Enhanced Captions

Yi Yuan[†], Dongya Jia[‡], Xiaobin Zhuang[‡], Yuanzhe Chen[‡], Zhuo Chen[‡], Yuping Wang[‡], Yuxuan Wang[‡], Xubo Liu[†], Xiyuan Kang[†], Mark D. Plumbley[†], Wenwu Wang[†]
[†]Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey [‡]ByteDance

***Abstract*—Generative models have shown significant achievements in audio generation tasks. However, existing models struggle with complex and detailed prompts, leading to potential performance degradation. We hypothesize that this problem stems from the simplicity and scarcity of the training data. This work aims to create a large-scale audio dataset with rich captions for improving audio generation models. We first develop an automated pipeline to generate detailed captions by transforming predicted visual captions, audio captions, and tagging labels into comprehensive descriptions using a Large Language Model (LLM). The resulting dataset, Sound-VECaps, comprises 1.66M high-quality audio-caption pairs with enriched details including audio event orders, occurred places and environment information. We then demonstrate that training the text-to-audio generation models with Sound-VECaps significantly improves the performance on complex prompts. Furthermore, we conduct ablation studies of the models on several downstream audio-language tasks, showing the potential of Sound-VECaps in advancing audio-text representation learning. Details of our dataset and demos are available here.**

***Index Terms*—Audio generation, audio retrieval, diffusion model, audio-language dataset**

I. Introduction

Generative models have recently achieved substantial success for text-to-audio generation. In particular, the development of language models [1], [2] and diffusion models [3], [4] have enabled the creation of powerful systems [5], [6] on generating high-fidelity audio clips.

Despite their success in generating audio with simple captions, current models struggle with complex prompts containing detailed information, which referred to the challenge as "prompt following" [3]. A potential reason for this limitation is that existing audio-caption datasets often lack in quantity and quality (detailed information) of the captions. In most of these datasets, each audio is matched with simple and short captions, typically, fewer than 10 words. As a result, the captions in these datasets may not contain fine-grained information that could be useful for highly controllable audio generation.

In addition, the simplicity of the caption often results in situations where the same caption corresponds to multiple audio files (e.g., there are 2.5K audio clips match with the caption " Music is playing" in WavCaps [7]), causing the system to avoid learning specific audio feature and lead to more instability in the generated outputs. A possible way to address this issue is to incorporate additional information, such as visual features, which have been shown to provide more detailed insights. One of the previous attempts is the Auto-ACD [8], where video features are used to improve the description of the event-occurring scene. However, Auto-ACD only takes the visual feature of the middle frame, and the caption has been designed to ignore the visual-only contents, losing more detailed information.

TABLE I
The analysis of audio-caption datasets. **Loc** and **Env** are the captions including locations and environmental information.

Dataset	Number	Avg. Len	Loc. Inf	Env. Inf
AudioSet [9]	2.1M	3	Label	Label
Clotho [10]	5K	11	1.2K	0.9K
AudioCaps [11]	46K	9	4K	3K
WavCaps [7]	400K	8	51K	37K
Auto-ACD [8]	1.9M	18	1.23M	69K
Sound-VECaps$_{\mathrm{A}}$	1.66M	31	1.44M	1.36M
Sound-VECaps$_{\mathrm{F}}$	1.66M	40	1.46M	1.38M

In this paper, we aim to leverage external visual guidance to enhance the audio captions. With improved captions, we can provide better alignment between the prompt and the sound, thereby improving text-to-audio generation systems. Specifically, we propose new pipelines to construct a large-scale audio-language dataset with vision-enhanced captions. Our approach first involves collecting external visual information using state-of-the-art (SoTA) image captioning models. These visual captions, combined with simple audio information, are then used to create new, enriched captions through Large Language Models (LLMs). By incorporating additional visual information, our method ensures the accuracy of audio details while enhancing the captions with comprehensive content, including temporal, spatial, and contextual elements related to the environment. Building on AudioSet [9], we introduce Sound-VECaps, a large-scale dataset comprising over 1.66M audio-caption pairs.

Using Sound-VECaps as the training dataset, our experiments with the audio generation model, AudioLDM [12], show substantial improvements over baseline models. To evaluate the performance on complex and extended prompts, we propose a new benchmark for text-to-audio generation by constructing an enhanced AudioCaps [11] testing set (same audio with better captions) named AudioCaps-Enhanced. Specifically, the AudioLDM-Large trained on Sound-VECaps achieves a Frechet Audio Distance (FAD) score of 1.49 on the AudioCaps. It further improves to a score of 1.06 on AudioCaps-Enhanced, significantly outperforming current

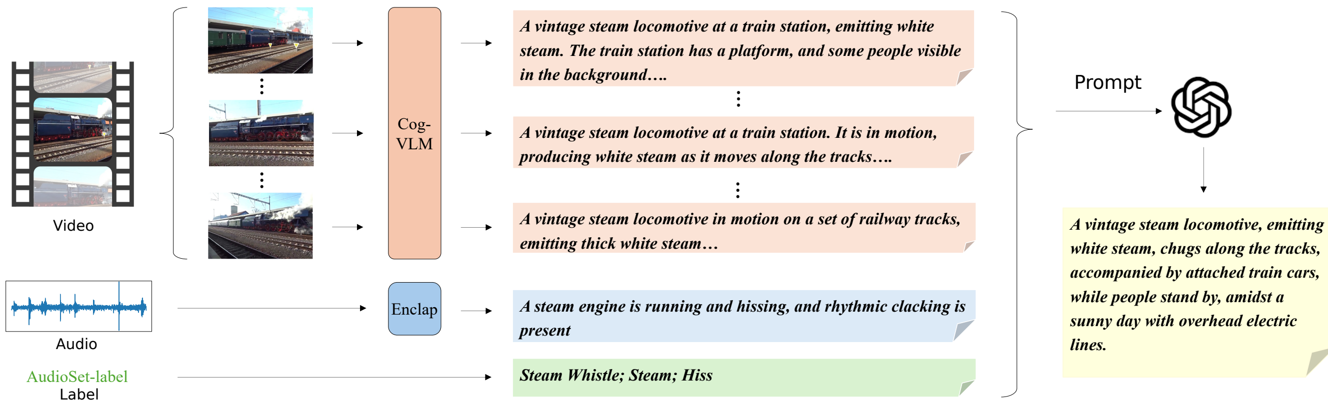

Fig. 1. The caption generation pipeline of the Sound-VECaps

SoTA models. Moreover, we conduct experiments on Sound-VECaps across various audio-language tasks, demonstrating that systems trained on Sound-VECaps achieve SoTA performance in specific audio-domain tasks, such as audio retrieval. We also investigate the effectiveness of the visual-only content within the caption and the impact of these features during inference. In addition, an external version of Sound-VECaps that excludes all the visual-only information (Sound-VECaps$_{\text{A}}$) is also provided for different research purposes.

II. Audio Dataset

Our Sound-VECaps dataset is built on AudioSet [9] with the processing pipeline shown in Figure 1. In particular, LLMs are prompted to generate captions based on three pieces of text information, namely, visual captions from the video, audio captions from the waveform, and the label tags in the original dataset.

A. Captions from Video

One of the novel aspects of the proposed dataset is that we leverage the caption of the corresponding video to provide more detailed information about the audio events. Different from previous visual-related approaches [8] that only apply the visual information of middle frames, the proposed strategy utilizes the captions of complete videos to secure more detailed descriptions. On the other hand, current SoTA video captioning systems [13], [14] mainly pool all the visual information into an aligned feature dimension, losing the temporal information (order of the events). Hence, we capture visual information for multiple frames through image captioning to maintain this temporal information. Specifically, we follow the image caption generation of Stable Diffusion 3 [15] and apply the CogVLM [16] captioning system. To improve the efficiency, the system only captions the frame of each video by second e.g., 11 captions for a 10-second audio.

B. Captions from Audio

We found that captions directly from video sometimes may not reflect the correct events, such as background and invisible sound. Hence, two constraints are provided to guide the LLM to understand the auditory information: the label provided by the original AudioSet dataset, and the simple audio caption generated by audio captioning models. We apply the SoTA captioning model, EnCLAP [17], to generate concise and brief captions of each audio clip.

C. Proposed Caption Generation

Combining three textual information mentioned above, an LLM is applied to generate the final caption, where we use Llama3-7B [18] to assemble re-caption the comprehensive description of each audio.

D. Dataset Processing

Due to the issues of some videos being too old (not accessible anymore), we collected a total of 1.81M videos from the AudioSet. In addition, around 10k video clips are skipped due to the sensitive policy of the LLMs (e.g., violence). Furthermore, we found that some video clips present static visual information with complete background sounds, leading the caption focusing on visual events but ignoring the actual audio events. To ensure the correctness of the visual guidance and improve the data quality, a filtering strategy is applied to detect and exclude the captions of static video which presents more than 80% same frames. Overall, we obtain the Sound-VECaps datasets containing 1.66M audio-caption pairs. The Sound-VECaps provides two different versions of captions for various purposes, specifically, Sound-VECaps$_{\text{A}}$ removes visual-only information and contains only audible contents or environmental-descriptive information, while Sound-VECaps$_{\text{F}}$ describes full detailed information including visual features, e.g., texts, names, shapes, and colours.

III. Audio Generation System

We conduct experiments on text-to-audio generation using AudioLDM [12] models, a SoTA audio generation model, to evaluate Sound-VECaps. For instance, AudioLDM is divided into four sections: a CLAP encoder for condition embedding, a latent diffusion-based model to generate audio features within the latent space, a variational autoencoder (VAE) decoder to reconstruct the information into a mel spectrogram, and a generative adversarial network(HiFi-GAN) vocoder [20] to produce the waveform as the final output.

TABLE II

THE COMPARISON BETWEEN GENERATION FRAMEWORKS EVALUATED ON AUDIOCAPS (PREVIOUS BENCHMARKS) AND AUDIOCAPS-ENHANCED (PROPOSED BENCHMARKS). BOTH $\text{CLAP}_{\text{score}}(\%)$ AND MOS ARE ONLY EVALUATED ON THE BEST RESULTS OF EACH SYSTEM. AC AND AS ARE SHORT FOR AUDIOCAPS [11] AND AUDIOSET [9] RESPECTIVELY.

Model	Training Dataset	AudioCaps			AudioCaps-Enhanced			Best Result	
		KL ↓	IS ↑	FAD ↓	KL ↓	IS ↑	FAD ↓	$\text{CLAP}_{\text{score}}(\%)$↑	MOS↑
AudioGen [19]	AC+AS+8 others	1.49	**9.93**	1.82	2.63	6.66	4.53	40.30	3.56
AudioLDM [12]	AC+AS+2 others	2.22	7.54	2.98	2.48	5.63	5.65	40.17	3.08
Tango2 [6]	AudioCaps	1.32	9.12	2.03	2.19	6.84	4.99	43.39	3.85
AudioLDM2-Large [5]	AC+AS+6 others	**1.22**	7.86	1.83	1.65	7.61	2.92	38.05	3.47
AudioLDM-T5	Sound-VECaps_{F}	1.68	6.8	1.78	1.44	6.29	1.45	41.20	3.92
AudioLDM-T5-L	Sound-VECaps_{F}	1.49	8.77	**1.49**	**1.17**	**7.96**	**1.06**	**43.59**	**4.05**

Instead of using CLAP [1] for computing the audio and text embedding, our experiments replaced the encoder with a T5 [2] encoder for condition embedding, and an cross-attention module [21] is applied to process the T5 embedding. We name the system as AudioLDM-T5. For the remaining modules, we follow the same design of AudioLDM and our system takes the pre-trained VAE decoder and Hifi-GAN [20] vocoder for audio feature reconstruction.

IV. EXPERIMENTS

A. Evaluation Dataset

We first follow previous baseline models [12], [19] and evaluate the performance of text-to-audio generation on the AudioCaps testing set. Since AudioCaps only includes simple and audio-only textual information, we introduce a novel benchmark with enriched and enhanced captions (same audio with better captions). We apply the proposed re-captioning pipeline in Section II to generate improved captions for AudioCaps testing audio samples. Specifically, human supervision is applied during the captioning process to check the accuracy and relevance of each caption (ensure the quality of LLM outputs). The proposed AudioCaps-Enhanced testing dataset includes five different captions for each audio clip, totalling 4430 captions for 886 audio samples. Similar to the Sound-VECaps dataset, we provide both full-feature captions (AudioCaps-$\text{Enhanced}_{\text{F}}$) and captions excluding visual-only contents (AudioCaps-$\text{Enhanced}_{\text{A}}$) for various purposes.

B. Results

Effectiveness on Audio Generation. Audio generation systems are trained on Sound-VECaps to evaluate their effectiveness, where models are trained using the same hyperparameters of AudioLDM. Specifically, AudioLDM-T5 maintains the same size as AudioLDM [12], while AudioLDM-T5-L is a larger system with increased hidden sizes. As shown in Table II, AudioLDM-T5 achieves SoTA performance on the AudioCaps testing sets. Moreover, the larger model (AudioLDM-T5-L), trained on Sound-VECaps_{F}, outperforms baseline models by a large margin. In addition, current audio generation models struggle with complex and extended prompts, resulting in performance degradation on AudioCaps-Enhanced (e.g., the FAD score increases from 1.83 to 2.92 on AudioLDM2-L). By applying Sound-VECaps for training, AudioLDM-T5 models successfully overcome this limitation, achieving a FAD score of 1.06 and a MOS score of 4.05 on larger AudioLDM-T5-L.

Effectiveness of Visual-Only Content. To evaluate the effectiveness of the visual information in the captions, we compare the performance of different AudioLDM-T5-L systems trained and evaluated on various datasets that include and exclude visual-only content. Notably, all three versions of the testing dataset share the same group of audio clips (same target audio samples while using different prompts for generation), providing reliability assurance for the comparison. As shown in Table III, systems utilizing Sound-VECaps_{F} as the training dataset demonstrates enhanced performance across all three evaluation metrics. For the evaluation, using AudioCaps as the prompt presents a higher quality (IS score of 8.77), while the audio outputs generated through the prompts with visual content (AudioCaps-$\text{Enhanced}_{\text{F}}$) show minor degradation. However, audio samples generated through enriched prompts lead to significant improvements in the fidelity of generated audio, with the prompts excluding visual-only content (AudioCaps-$\text{Enhanced}_{\text{A}}$) showing SoTA performance. Through these experiments, we have summarized three key findings: 1) Training on captions with visual features can improve the capability of the system to handle auditory information and identify features across different modalities, leading to significant improvement in the overall performance. 2) The simplicity of the prompts in current evaluation benchmarks (e.g. AudioCaps) limits the presentation of detailed audio features. The proposed benchmark testing on AudioCaps-Enhanced enriches the information with more controllable features and offers greater potential for enhancing the output quality. 3) Although training with external visual features (Sound-VECaps_{F}) provides better

TABLE III

AUDIOLDM-T5-L MODELS TRAINED AND EVALUATED ON DIFFERENT DATASETS, WHERE E_{F} IS SHORT FOR $\text{ENHANCED}_{\text{F}}$ WITH FULL FEATURE CAPTIONS AND AUDIOCAP-E_{A} WITHOUT VISUAL-ONLY CONTENTS.

Training Dataset	Testing Dataset	KL↓	IS↑	FAD↓
Sound-VECaps_{A}	AudioCaps	1.22	7.31	1.65
Sound-VECaps_{A}	AudioCaps-E_{F}	1.33	6.27	1.67
Sound-VECaps_{A}	AudioCaps-E_{A}	1.38	7.18	1.64
Sound-VECaps_{F}	AudioCaps	1.49	**8.77**	1.49
Sound-VECaps_{F}	AudioCaps-E_{F}	**1.17**	7.96	1.06
Sound-VECaps_{F}	AudioCaps-E_{A}	1.19	8.13	**0.96**

TABLE IV
PERFORMANCE COMPARISON BETWEEN DIFFERENT SYSTEMS ON AUDIOCAPS AND AUDIOCAPS-ENAHCNED(PROPOSED BENCHMARK), CLAP_M AND CLAP_L ARE MODELS TRAINED BY MICROSOFT [22] AND LAION [1], USING DIFFERENT STRUCTURES AND DATASETS RESPECTIVELY. FOR THE TRAINING SET, "AC", "CL" AND "LA" ARE SHORT FOR AUDIOCAPS, CLOTHO AND LAION-AUDIO-630K DATASETS RESPECTIVELY.

Model	Training Set	AudioCaps						AudioCaps-Enhanced					
		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
CLAP_L [1]	AC+CL+LA	34.2	71.1	84.1	43.1	79.5	90.1	21.6	54.9	71.6	34.1	65.4	77.7
CLAP_M [22]	4.6M-Audio	33.5	70.4	80.2	47.8	80.2	90.7	19.5	46.2	60.9	29.3	59.1	70.1
WavCaps [7]	WavCaps+AC+CL	39.7	74.5	86.1	51.7	82.3	90.6	23.0	52.3	66.2	35.5	62.8	75.8
Auto-ACD [8]	Auto-ACD	40.4	**75.3**	**87.4**	51.1	84.0	92.7	46.3	81.8	89.7	55.8	84.1	92.6
CLAP	Sound-VECaps-Audio	**41.2**	74.5	85.3	53.3	83.2	93.0	49.2	83.1	91.7	59.1	87.5	94.3
CLAP	Sound-VECaps-Full	39.2	74.1	85.0	**54.0**	**85.5**	**93.2**	**53.1**	**85.7**	**91.3**	**64.3**	**90.2**	**96.4**

results, the additional visual information may increase data complexity during inference. Therefore, the system that uses prompts without visual-only features ($\text{AudioCaps-Enhanced}_\text{A}$) generates the best result (0.96 on FAD).

C. Studies on Other Audio Tasks

Audio Caption Retrieval. In addition to our experiments on audio generation, we evaluated the effectiveness of Sound-VECaps for improving audio-language retrieval systems. Specifically, we employed the framework in WavCaps [7], which uses RoBERTa as the text encoder and HTSAT as the audio encoder, to train and evaluate CLAP-based models in audio-text cross modal retrieval tasks. As illustrated by Table IV, the evaluation using the AudioCaps testing set demonstrated that the CLAP-based models trained on the Sound-VECaps dataset matched the performance of the baseline models (trained on other datasets). However, when testing with the enhanced captions (AudioCaps-Enhanced), the experiment shows a notable performance decline in current SoTA systems, highlighting the challenges posed by longer and more detailed textual information. Conversely, the systems trained with enriched captions (Auto-ACD [8] and Sound-VECaps) present improvements in retrieval capabilities, while the system on $\text{Sound-VECaps}_\text{F}$ achieves the best performance. The results show the enhancement of captions through visual information, as well as the accuracy and robustness of the system on Sound-VECaps. Additionally, the CLAP model trained with $\text{Sound-VECaps}_\text{F}$ exhibited better performance, particularly on AudioCaps-Enhanced dataset, indicating that the overall performance of the system can be further improved with the visually augmented captions.

Temporal Feature Retrieval. Another aspect of Sound-VECaps is the temporal information. Since visual captions are provided by frame, temporal information (events ordering) is also included. We applied the T-Classify method from T-CLAP [23] to evaluate the performance of temporal feature retrieval. Table V demonstrates stronger capabilities to identify temporal information in systems with Sound-VECaps, illustrating its improvement in temporal features. In addition, the system developed without visual-only contents presents better performance, indicating that extensive visual features might influence the model's understanding of temporal information.

Limitation. We also attempt to use Sound-VECaps for other audio-related tasks. However, due to the rich content in our captions, particularly regarding visual information, the model did not perform well on tasks that are purely audio-targeted content, such as audio captioning and zero-shot tasks. These results demonstrate that Sound-VECaps may not be broadly applied to audio-language tasks. It is mainly effective in a range of tasks that require processing and distinguishing detailed content, such as generation and retrieval.

TABLE V
RESULTS OF TEMPORAL FEATURE RETRIEVAL ON T-CLASSIFY [23].

Model	Text-to-Audio	Audio-to-Text
CLAP_m [22]	45.7	44.1
CLAP_l [1]	56.2	53.2
WavCaps [7]	58.5	49.7
$\text{Sound-VECaps}_\text{F}$	61.2	57.3
$\text{Sound-VECaps}_\text{A}$	**63.6**	**59.0**

V. CONCLUSION

We present Sound-VECaps, a large-scale dataset comprising 1.66M audio clips with captions augmented by video data, to address the challenge of prompt following in audio generation systems. Experiments show that systems trained on Sound-VECaps achieve SoTA performance and outperform baseline models. In addition, a new benchmark using improved captions is proposed to evaluate audio-language systems on complex and extended prompts. Our systems are further improved by a large margin when taking more detailed captions as prompts, reaching a FAD score of 0.96. Nevertheless, we demonstrated that using Sound-VECaps can offer substantial improvements in audio-language and temporal feature retrieval. The results of the proposed AudioCaps-Enhanced testing sets highlight the limitations of previous benchmarks and emphasize the potential of better prompts in advancing the performance of audio-language models. Overall, we develop two versions of the proposed datasets with captions that include and exclude visual-only content for different purposes and tasks and hope these datasets will generate more profound impacts on audio-language learning.

VI. ACKNOWLEDGMENT

This research was partly supported by a research scholarship from the China Scholarship Council (CSC), an internship from ByteDance, and supported by the Engineering and Physical Sciences Research Council (EPSRC) under Grant EP/T019751/1 "AI for Sound". For the purpose of open access, the authors have applied a Creative Commons Attribution (CC BY) license to any Author Accepted Manuscript version arising.

REFERENCES

[1] Y. Wu, K. Chen, T. Zhang, Y. Hui, T. Berg-Kirkpatrick, and S. Dubnov, "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, 2023.

[2] 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. 1, pp. 5485–5551, 2020.

[3] J. Betker, G. Goh, L. Jing, T. Brooks, J. Wang, L. Li, L. Ouyang, J. Zhuang, J. Lee, Y. Guo *et al.*, "Improving image generation with better captions," *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, vol. 2, no. 3, p. 8, 2023.

[4] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE Computer Society, 2022, pp. 10 674–10 685.

[5] 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*, 2024.

[6] D. Ghosal, N. Majumder, A. Mehrish, and S. Poria, "Text-to-audio generation using instruction tuned LLM and latent diffusion model," *arXiv preprint arXiv:2304.13731*, 2023.

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

[8] L. Sun, X. Xu, M. Wu, and W. Xie, "A large-scale dataset for audio-language representation learning," *arXiv preprint arXiv:2309.11500*, 2023.

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

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

[11] C. D. Kim, B. Kim, H. Lee, and G. Kim, "AudioCaps: Generating captions for audios in the wild," in *Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 2019.

[12] 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 *International Conference on Machine Learning*, 2023.

[13] S. Chen, X. He, L. Guo, X. Zhu, W. Wang, J. Tang, and J. Liu, "Valor: Vision-audio-language omni-perception pretraining model and dataset," *arXiv preprint arXiv:2304.08345*, 2023.

[14] H. Xu, Q. Ye, M. Yan, Y. Shi, J. Ye, Y. Xu, C. Li, B. Bi, Q. Qian, W. Wang *et al.*, "mPLUG-2: A modularized multi-modal foundation model across text, image and video," in *International Conference on Machine Learning*. PMLR, 2023, pp. 38 728–38 748.

[15] P. Esser, S. Kulal, A. Blattmann, R. Entezari, J. Müller, H. Saini, Y. Levi, D. Lorenz, A. Sauer, F. Boesel *et al.*, "Scaling rectified flow transformers for high-resolution image synthesis," in *Forty-first International Conference on Machine Learning*, 2024.

[16] W. Wang, Q. Lv, W. Yu, W. Hong, J. Qi, Y. Wang, J. Ji, Z. Yang, L. Zhao, X. Song *et al.*, "CogVLM: Visual expert for pretrained language models," *arXiv preprint arXiv:2311.03079*, 2023.

[17] J. Kim, J. Jung, J. Lee, and S. H. Woo, "Enclap: Combining neural audio codec and audio-text joint embedding for automated audio captioning," *arXiv preprint arXiv:2401.17690*, 2024.

[18] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[19] 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 *International Conference on Learning Representations*, 2023.

[20] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: generative adversarial networks for efficient and high fidelity speech synthesis," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 2020, pp. 17 022–17 033.

[21] Y. Yuan, H. Liu, X. Liu, Q. Huang, M. D. Plumbley, and W. Wang, "Retrieval-augmented text-to-audio generation," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 581–585.

[22] B. Elizalde, S. Deshmukh, M. Al Ismail, and H. Wang, "CLAP learning audio concepts from natural language supervision," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2023.

[23] Y. Yuan, Z. Chen, X. Liu, H. Liu, X. Xu, D. Jia, Y. Chen, M. D. Plumbley, and W. Wang, "T-CLAP: Temporal-Enhanced Contrastive Language-Audio Pretraining," *arXiv preprint arXiv:2404.17806*, 2024.