Sound-VECaps: Improving Audio Generation With Visual Enhanced Captions

Yi Yuan[1], Dongya Jia[2], Xiaobin Zhuang[2], Yuanzhe Chen[2], Zhengxi Liu[2], Zhuo Chen[2], Yuping Wang[2], Yuxuan Wang[2], Xubo Liu[1], Xiyuan Kang[1], Mark D. Plumbley[1], and Wenwu Wang[1]

[1]Centre for Vision, Speech and Signal Processing, University of Surrey
[2]ByteDance
{yi.yuan,xubo.liu,xk00063,m.plumbley,w.wang}@surrey.ac.uk
{jiadongya, zhuangxiaobin, chenyuanzhe, zhuo.chen1, wangyuping, wangyuxuan.11}@bytedance.com

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. Our dataset and models are available at https://yyua8222.github.io/Sound-VECaps-demo/.

1 Introduction

Generative models have recently achieved substantial success for text-to-audio generation. In particular, the development of language models [21, 16] and diffusion models [1, 17] have enabled the creation of powerful systems [14, 7] on generating high-fidelity audio clips.

Despite success in generating audio with simple captions, current models struggle with complex prompts containing detailed information, which is referred to the challenge as “prompt following” [1]. A potential reason for this limitation is that existing datasets often lack in both 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 [15]), 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 [18], where video features are used to improve the description of the

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

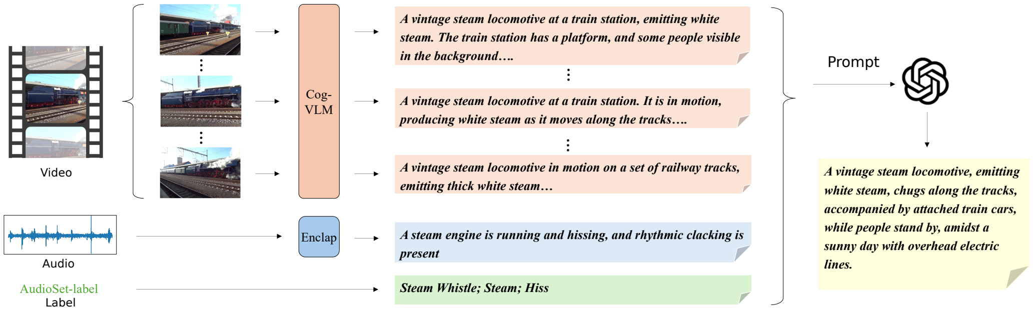
Video
Cog-VLM
A vintage steam locomotive at a train station, emitting white steam. The train station has a platform, and some people visible in the background….
A vintage steam locomotive at a train station. It is in motion, producing white steam as it moves along the tracks….
A vintage steam locomotive in motion on a set of railway tracks, emitting thick white steam…
Audio
Enclap
A steam engine is running and hissing, and rhythmic clacking is present
AudioSet-label
Label
Steam Whistle; Steam; Hiss
Prompt
A vintage steam locomotive, emitting white steam, chugs along the tracks, accompanied by attached train cars, while people stand by, amidst a sunny day with overhead electric lines.

Figure 1: The caption generation pipeline of the Sound-VECaps

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.

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 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 [6], 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 [13], 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 [9] 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 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$_{\mathrm{A}}$) is also provided for different purposes.

2 Audio Dataset

Our Sound-VECaps dataset is built on AudioSet [6], by following 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.

2.1 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 [18] 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 [3, 22] 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 [5] and apply the CogVLM [20] 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.

2.2 Captions from Audio

We found that captions directly from video sometimes may not reflect the correct information, such as background and invisible sound. Hence, two constraints are provided to guide the LLM to understand actual auditory information. One is the label provided by the original AudioSet dataset, another is a simple audio caption generated by audio captioning models. In our system, we applied the SoTA captioning model, EnCLAP [10], to generate the concise and brief captions of each audio clip.

2.3 Proposed Caption Generation

Combining three textual information mentioned above, an LLM is applied to generate the final caption, where we use Llama3-7B [19] to assemble re-caption the comprehensive description of each audio. Details of the prompts for generation are provided in the Appendix A.

2.4 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$_\mathrm{A}$ removes visual-only information and contains only audible contents or environmental-descriptive information, while Sound-VECaps$_\mathrm{F}$ describes full detailed information including visual features, e.g., texts, names, shapes, and colours.

3 Audio Generation System

To evaluate the impact of the proposed dataset, we conduct experiments on text-to-audio generation using AudioLDM [13] models, a SoTA audio generation model. 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 [11] to produce the waveform as the final output.

Instead of using CLAP [21] for computing the audio and text embedding as conditions during the training and inference stages respectively, our experiment only takes the text embedding for conditioning throughout the whole stage. Thus, the CLAP encoder is replaced with a T5 [16] encoder for condition embedding, and an across-attention module [24] is applied to process the T5 embedding instead of the previous film conditioning module [13]. 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 vocoder for audio feature reconstruction.

Table 1: The comparison between different audio generation frameworks, evaluation on AudioCaps (previous benchmarks) and AudioCaps-Enhanced (proposed benchmarks). Both $\mathrm{CLAP}_{\mathrm{score}}(\%)$ and MOS are only evaluated on the best results of each system, where $\mathrm{CLAP}_{\mathrm{score}}(\%)$ is calculated based on the system developed in Section 4.3. AC and AS are short for AudioCaps [9] and AudioSet [6] respectively.

Model	**Training Dataset**	**AudioCaps**			**AudioCaps-Enhanced**			**Best Result**	
		KL ↓	IS ↑	FAD ↓	KL ↓	IS ↑	FAD ↓	$\mathrm{CLAP}_{\mathrm{score}}(\%)\uparrow$	MOS↑
AudioGen [12]	AC+AS+8 others	1.49	**9.93**	1.82	2.63	6.66	4.53	40.30	3.56
AudioLDM [13]	AC+AS+2 others	2.22	7.54	2.98	2.48	5.63	5.65	40.17	3.08
Tango2 [7]	AudioCaps	1.32	9.12	2.03	2.19	6.84	4.99	43.39	3.85
AudioLDM2-Large [14]	AC+AS+6 others	**1.22**	7.86	1.83	1.65	7.61	2.92	38.05	3.47
AudioLDM-T5	Sound-VECaps$_\mathrm{F}$	1.68	6.8	1.78	1.44	6.29	1.45	41.20	3.92
AudioLDM-T5-L	Sound-VECaps$_\mathrm{F}$	1.49	8.77	**1.49**	**1.17**	**7.96**	**1.06**	**43.59**	**4.05**

4 Experiments

4.1 Evaluation Dataset

We first follow previous baseline models [13, 12] and evaluate the performance of text-to-audio generation on the AudioCaps testing set. However, AudioCaps only includes simple and audio-only textual information, to better evaluate the system on complex and extended prompts, we introduce a novel benchmark with enriched and enhanced captions (same audio with better captions). For the same testing audio samples, we apply the proposed re-captioning pipeline in Section 2 to generate improved captions. Specifically, human supervision is applied during the captioning process to check the accuracy and relevance of each caption and ensure the quality of LLM outputs. Same as the AudioCaps testing set, 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-Enhanced$_{\text{F}}$) and captions that exclude visual-only contents (AudioCaps-Enhanced$_{\text{A}}$) for various evaluation purposes.

4.2 Results

Effectiveness on Audio Generation. Audio generation systems are trained on Sound-VECaps to evaluate the effectiveness, where all the models are trained using the same hyperparameters of AudioLDM. Specifically, AudioLDM-T5 maintains the same size as AudioLDM [13], while AudioLDM-T5-L is a larger system with increased hidden sizes. As shown in Table 1, AudioLDM-T5 achieves SoTA performance on the AudioCaps testing sets. Moreover, the larger model, AudioLDM-T5-L trained on Sound-VECaps$_{\text{F}}$, outperforms baseline models trained on other datasets by a large margin. In addition, current audio generation models struggle with complex and extended prompts, resulting in notable performance degradation on AudioCaps-Enhanced (e.g., the FAD score increases from 1.83 to 2.92 on AudioLDM2-Large). By applying Sound-VECaps as the training dataset, AudioLDM-T5 models successfully overcome this limitation, achieving a FAD score of 1.06 and a MOS score of 4.05 on 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 2 left, systems utilizing Sound-VECaps$_{\text{F}}$ as the training dataset demonstrates enhanced performance across all three evaluation metrics. For the evaluation, using AudioCaps as the prompt presents a higher quality with an IS score of 8.77, while the audio outputs generated through the prompts with visual content (AudioCaps-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-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$_{\text{F}}$) provides better results, the additional visual information may increase data complexity during inference. Therefore, the system that uses prompts without visual-only features (AudioCaps-Enhanced$_{\text{A}}$) generates the best result with an FAD score of 0.96.

Table 2: Results for visual-only experiments on the left and temporal-feature experiments on the right

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

Model	Text-to-Audio	Audio-to-Text
CLAP$_M$ [4]	45.7	44.1
CLAP$_L$ [21]	56.2	53.2
WavCaps [15]	58.5	49.7
Sound-VECaps$_{\text{F}}$	61.2	57.3
Sound-VECaps$_{\text{A}}$	**63.6**	**59.0**

Table 3: Performance comparison between different systems, CLAP$_M$ and CLAP$_L$ are models trained by Microsoft [4] and LAION [21]. For the training set, "AC", "CL" and "LA" are short for AudioCaps, Clotho and LAION-630k respectively. AudioCaps presents the results of the original testing set and AudioCaps-Enhanced for the proposed caption-enhanced testing set with full features.

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$_L$ [21]	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$_M$ [4]	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 [15]	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 [18]	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
Sound-VECaps$_{\mathrm{A}}$	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
Sound-VECaps$_{\mathrm{F}}$	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**

4.3 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 [15], which uses BERT [8] as the text encoder and HTSAT [2] as the audio encoder, to train and evaluate CLAP-based models in audio-text retrieval tasks. As illustrated by Table 3, the CLAP-based models trained on the Sound-VECaps dataset matched the performance of the baseline models on AudioCaps testing set. However, the experiment shows a notable performance decline across current SoTA systems on AudioCaps-Enhanced, highlighting the challenges posed by longer and more detailed textual information. Conversely, the systems trained with enriched captions, such as Auto-ACD [18] and Sound-VECaps, present improvements in retrieval capabilities, where the system on Sound-VECaps$_{\mathrm{F}}$ achieves the best performance. The results show the enhancement of Sound-VECaps through visual information on both the accuracy and robustness of the system. Additionally, the CLAP model trained with Sound-VECaps$_{\mathrm{F}}$ 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 distinguishing aspect of Sound-VECaps is the temporal information. Since visual guidance is provided by frame, temporal information (events ordering) is also included. We applied the T-Classify method from T-CLAP [23] to evaluate the performance on temporal feature retrieval. Results in Table 2 right demonstrate a stronger capability to identify temporal information in the system on Sound-VECaps, illustrating the improvement of temporal features. 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 the proposed dataset for several other audio-related tasks. However, due to the rich content in our captions, particularly regarding environmental or 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.

5 Conclusion

We have presented 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 the AudioLDM models 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. In addition, we demonstrated that using Sound-VECaps can offer substantial improvements in audio retrieval and temporal feature identification. Nevertheless, the results between AudioCaps and AudioCaps-Enhanced testing sets highlight the limitations of previous benchmarks that rely on simple prompts and emphasize the potential of the better prompts in advancing the performance of audio-language models. We developed two versions of the proposed datasets with captions that include and exclude visual-only content for different purposes and tasks and we hope these datasets will generate more profound impacts on audio-language learning.

References

[1] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8, 2023.

[2] Ke Chen, Xingjian Du, Bilei Zhu, Zejun Ma, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. HTSAT: A hierarchical token-semantic audio transformer for sound classification and detection. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 646–650. IEEE, 2022.

[3] Sihan Chen, Xingjian He, Longteng Guo, Xinxin Zhu, Weining Wang, Jinhui Tang, and Jing Liu. Valor: Vision-audio-language omni-perception pretraining model and dataset. *arXiv preprint arXiv:2304.08345*, 2023.

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

[5] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024.

[6] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. AudioSet: An ontology and human-labeled dataset for audio events. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 776–780, 2017. doi: 10.1109/ICASSP.2017.7952261.

[7] Deepanway Ghosal, Navonil Majumder, Ambuj Mehrish, and Soujanya Poria. Text-to-audio generation using instruction tuned LLM and latent diffusion model. *arXiv preprint arXiv:2304.13731*, 2023.

[8] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2. Minneapolis, Minnesota, 2019.

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

[10] Jaeyeon Kim, Jaeyoon Jung, Jinjoo Lee, and Sang Hoon Woo. Enclap: Combining neural audio codec and audio-text joint embedding for automated audio captioning. *arXiv preprint arXiv:2401.17690*, 2024.

[11] Jungil Kong, Jaehyeon Kim, and Jaekyoung 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*, pages 17022–17033, 2020.

[12] Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi. AudioGen: textually guided audio generation. In *International Conference on Learning Representations*, 2023.

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

[14] 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*, 2024.

[15] 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.

[16] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(1):5485–5551, 2020.

[17] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10674–10685. IEEE Computer Society, 2022.

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

[19] 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 arXiv:2302.13971*, 2023.

[20] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. CogVLM: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*, 2023.

[21] 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, ICASSP*, 2023.

[22] Haiyang Xu, Qinghao Ye, Ming Yan, Yaya Shi, Jiabo Ye, Yuanhong Xu, Chenliang Li, Bin Bi, Qi Qian, Wei Wang, et al. mPLUG-2: A modularized multi-modal foundation model across text, image and video. In *International Conference on Machine Learning*, pages 38728–38748. PMLR, 2023.

[23] Yi Yuan, Zhuo Chen, Xubo Liu, Haohe Liu, Xuenan Xu, Dongya Jia, Yuanzhe Chen, Mark D Plumbley, and Wenwu Wang. T-CLAP: Temporal-Enhanced Contrastive Language-Audio Pretraining. *arXiv preprint arXiv:2404.17806*, 2024.

[24] Yi Yuan, Haohe Liu, Xubo Liu, Qiushi Huang, Mark D Plumbley, and Wenwu Wang. Retrieval-augmented text-to-audio generation. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 581–585. IEEE, 2024.

A Appendix

A.1 LLMs Prompts

We provide the prompts used as the input for the Llama3 model to generate our proposed captions. As shown in the Figure 2, the prompt is a combination of three different features. In the system section, both the caption from Enclap and the audio label are provided, while the frame captions are presented as the user input. Two different contents are also provided for both the full-featured caption (section in green boundaries) and the caption that filtered all the visual-only contents (section in red boundaries). For the AudioCaps-Enhanced dataset, we apply the same prompting pipeline, while changing the caption of enclap into the actual caption provided by the AudioCaps testing set. Nevertheless, all the captions for AudioCaps-Enhanced are generated under human-involved supervision, to ensure the correctness and relevance of the prompts.

Prompts for Llama3 to generate the caption

Role-System:

You are a helpful, assistant for identifying audio events and generating sentences. Please combine three different features of a 10-second audio and help the user to generate a single sentence of caption.

The caption feature is a sentence generated by an audio-caption model: **{enclap_caption}**.

The label feature is several audio events that happened in the audio: **{audio_label}.**

Lastly, the user is given several sentences which are the image description of the scene for each second, connected by "and then".

Please identify all the audio events based on all three features, and try to conclude in one single sentence to describe this scene with audio events or actions that present sound.

Please include some time features to present the order of each event, such as "and then", "followed by", etc for order; "and", "while" etc for happening parallelly.

Based on the first caption feature, you might need to change or alter any wrong audio event, improve the sentence with more features, such as the weather, the emotion of any people, the description of the car and so on.

Remove all the visual features that are too specific and irrelevant to the audio events, such as the colour, shape, any text or label, name and what people are writing, and so on. Please make sure that you keep most of the contents, especially the audio-related events, and their possible correlation, such as the order of occurrence, background, and so on.

Please use the sentences provided by the user to identify the background/forground sounds, and point out the backgrounds sounds in the sentence.

Role-User:

The descriptions of the frames are: **{frame_caption}**

Figure 2: The prompts used for caption generation, where the contents in green section are used for full feature captions and red sections are applied to avoid any visual-only contents,

A.2 Caption Demos

We present the comparison of the captions from Sound-VECaps and other baseline datasets. As shown in Table 4, for each audio sample, we compare the caption from the AudioSet label, Wavcaps, Enclap, Auto-ACD and two versions of the proposed Sound-VECaps.

Nevertheless, we also present a sample of the AudioCaps-Enhanced testing dataset in Table 5

Num	Dataset	Caption
No.1	AudioSet	Honk , Speech
	WavCaps	Crinkling, wind, laughter, ducks, and people speaking are heard.
	Enclap	Wind blows, ducks quack and people speak.
	Auto-ACD	The wind blows as ducks quack and a man speaks.
	VECaps_a	A goose quacks and honks, while the wind blows, and the person speaks, followed by the sound of bread being offered to the goose, amidst the scattered leaves and grass.
	VECaps_f	As the person stands near the car, a goose quacks and honks, while the wind blows, and the person speaks, followed by the sound of bread being offered to the goose, and the goose's orange beak and feet can be seen amidst the scattered leaves and grass.
No.2	AudioSet	Dial tone
	WavCaps	A dial tone is heard.
	Enclap	A telephone rings
	Auto-ACD	A dial tone rings with a probability of 0.66, indicating a telephone call in an indoor setting.
	VECaps_a	A telephone rings in the background, followed by a dial tone, while a man is holding a child in his arms, as a news article plays in the background.
	VECaps_f	A telephone rings in the background, followed by a dial tone, while a man is holding a child in his arms in front of a destroyed building, as a news article about US urging Israel to protect civilians and increase aid to Gaza plays in the background.
No.3	AudioSet	Music, instrument, string
	WavCaps	Music is playing.
	Enclap	A man speaks over a loudspeaker as music plays in the distance
	Auto-ACD	The sitar player strums melodious music on stage, accompanied by instruments in an orchestra pit.
	VECaps_a	A man plays a sitar, accompanied by the sound of a plucked string instrument, followed by the soft hum of a bowed string instrument, in a dimly lit room, with music playing in the distance.
	VECaps_f	A man plays the sitar, a traditional Indian stringed instrument, in a dimly lit room with a projection screen in the background, while music plays in the distance, accompanied by the sound of a plucked string instrument, followed by the soft hum of a bowed string instrument.

Table 4: The comparison between different caption datasets.

Dataset	Caption
AudioCaps	A man talking as water splashes.
$\text{AudioCaps-E}_\text{a}$	Waves crashing onto a calm shore, followed by a man speaking amid a gathering of people, some with cameras, by a coastal backdrop.
$\text{AudioCaps-E}_\text{f}$	Waves gently lap against the shore under an overcast sky, as a man in a grey shirt and glasses addresses a gathering. Surrounding him, a few individuals, possibly security or journalists, hold cameras and microphones, suggesting a public event near a tropical waterfront.

Table 5: The comparison between AudioCaps and proposed AudioCaps-E testing dataset.