Diversity is the Key: Enhancing LLM-based Post-processing for Automated Audio Captioning

Seyed Ali Farokh, Mohammad Mehdi Homayounpour, Ahmad Nickabadi
Department of Computer Engineering, Amirkabir University of Technology, Tehran, Iran
{alifarokh, homayoun, nickabadi}@aut.ac.ir

Abstract

Automated Audio Captioning (AAC) is a multimodal task aimed at generating natural language descriptions of audio content. Previous studies have shown that LLMs can improve AAC performance by summarizing audio events based on a list of candidate captions, which are selected by an external reranker from those generated using Nucleus Sampling. However, the reranking process often selects overly similar captions, disregarding the original diversity of the sampled captions. In this work, we show that this diversity reflects the AAC model's level of certainty and propose a lightweight candidate selection approach that preserves the initial diversity of the generated captions. This, in turn, enables an LLM to summarize the captions while considering the AAC model's certainty in a few-shot setting. Experimental results demonstrate that our method outperforms previous post-processing techniques while being significantly faster.

Keywords: Automated Audio Captioning, Large Language Models, In-context Learning, Post-processing

1 Introduction

Automated Audio Captioning (AAC) is a multimodal task that aims to generate natural language descriptions of the content within audio samples. Unlike Automatic Speech Recognition (ASR), which focuses on transcribing spoken language, AAC primarily targets environmental and non-speech sounds, providing meaningful descriptions of auditory scenes and events.

One of the primary challenges in AAC lies in the inherent ambiguity of audio signals. Unlike image captioning, where objects can be described through concrete attributes such as shape, color, size, and spatial relationships, describing audio clips is significantly more complex (Wu et al., 2019). Acoustic events often exhibit overlapping or similar sound characteristics, leading to varied perceptions among individuals, as shown in Figure 1 (Zhang et al., 2023; Drossos et al., 2020). Consequently, widely used audio captioning datasets, such as Clotho (Drossos et al., 2020), provide multiple ground-truth captions from different annotators for each audio sample, and models are typically trained on one-to-many audio-caption pairs, where each audio clip is randomly paired with a single ground-truth caption in each iteration (Zhang et al., 2023). This can introduce uncertainty in the learned representations and potentially result in performance degradation.

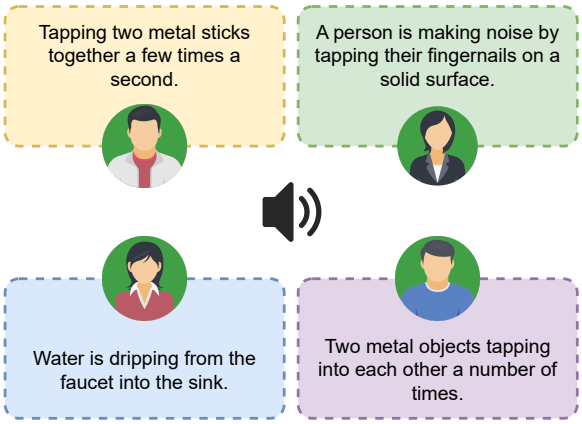

Figure 1: Diverse and occasionally conflicting perceptions of a single audio sample due to the inherent ambiguity of audio signals. The captions correspond to one training sample from the Clotho dataset (`dual metal.wav`).

Thanks to the annual DCASE challenges[1] and the release of open-source audio captioning datasets such as Clotho (Drossos et al., 2020) and AudioCaps (Kim et al., 2019), advancements in AAC research have gained momentum in recent

[1]IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events, available at https://dcase.community

Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025), pages 87–94
November 20-22, 2025

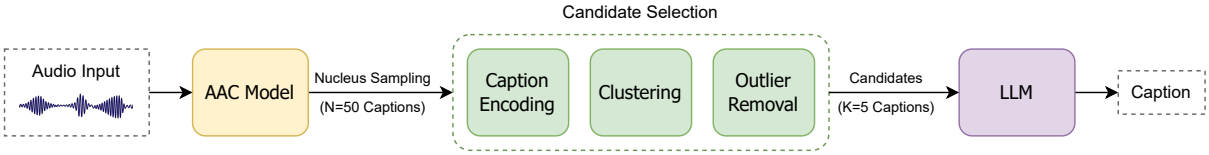

Figure 2: Overview of our proposed method. First, $\boldsymbol{N} = 50$ captions are generated for a given audio input using Nucleus Sampling. Next, in the candidate selection stage, $\boldsymbol{K} = 5$ captions are chosen to preserve the diversity of the generated captions. Finally, these selected captions are processed by an LLM to further enhance diversity and produce the final caption.

years. Most state-of-the-art AAC models employ an encoder-decoder architecture (Xu et al., 2022; Ye et al., 2022; Narisetty et al., 2021; Wu et al., 2024), where the encoder is typically a pre-trained audio encoder, such as PANN (Kong et al., 2020) or BEATs (Chen et al., 2023), that extracts audio features from the input signal. These features are then passed to an autoregressive text decoder, such as BART (Lewis et al., 2020) or GPT-2 (Radford et al., 2019), which generates the corresponding caption. The decoders normally generate sequences using greedy decoding and beam decoding.

In addition to these conventional decoding methods, recent research has demonstrated that a hybrid sampling and reranking strategy, which leverages external pre-trained models for reranking, can improve the outputs of AAC models by exploring a broader search space than beam search (Wu et al., 2024; Jung et al., 2024). Furthermore, inspired by the success of Large Language Models (LLMs) in a zero-shot setting across a variety of tasks and their ability to generate human-like text (Radford et al., 2019), recent studies in AAC have incorporated zero-shot LLM-based caption summarization (Jung et al., 2024) and error correction (Liu et al., 2024) as post-processing steps, demonstrating the effectiveness of these techniques in refining the generated captions.

In this work, we hypothesize that the diversity of sampled captions reflects the AAC model's level of certainty regarding a given input. We demonstrate that reranking is not the most effective approach for candidate caption selection, as the resulting captions lack sufficient diversity to both capture the model's uncertainty and serve as input for LLM-based summarization. To address this limitation, we propose a method that preserves the original diversity of the sampled captions and employs an LLM in a few-shot setting to generate a final caption while considering the AAC model's uncertainty. Experimental results show that the proposed method outperforms previous post-processing techniques while being significantly simpler and faster. Our contributions can be summarized as follows: (1) we propose a lightweight candidate caption selection method as an alternative to the hybrid sampling and reranking strategy, (2) we enhance AAC performance through an LLM-based post-processing approach that leverages in-context learning and accounts for the AAC model's certainty, and (3) we introduce a simple technique to identify high-quality captions generated by AAC models, enabling selective LLM-based refinement that improves performance while minimizing unnecessary processing overhead. The source code, experiments, and results from this work will be publicly released upon publication.

2 Related Work

2.1 Sampling and Reranking

Nucleus Sampling (Holtzman et al., 2020), also known as Top-$\boldsymbol{p}$ Sampling, is a stochastic decoding method commonly used in modern LLMs to generate more diverse text compared to conventional beam decoding (Shi et al., 2024). Inspired by this, Wu et al. (2024), the winners of the DCASE 2023 AAC challenge, observed that approximately one-third of the captions generated using Nucleus Sampling achieve higher scores on AAC metrics compared to those produced with beam decoding. To leverage this advantage, they proposed a sampling and reranking strategy that first generates a list of $\boldsymbol{N}$ captions (50 in their original work) using Nucleus Sampling, followed by a hybrid reranking method to select the most suitable caption from the list by computing two reranking scores. The *decoder score* is obtained by feeding the input waveform into the encoder and the generated captions into the decoder to directly compute the caption log-likelihood on the decoder outputs. The *encoder score* is calculated as the cosine similarity between the audio embedding, obtained by feeding the input

Audio File: `Shanghai Traffic Near Peoples Square.wav`		
	CLAP Reranking	**Our Proposed Method**
Candidate Selection	Several cars and trucks are driving by on a busy street. Several cars driving by on a busy street. A busy street with cars driving by on a nearby road. A busy street with vehicles passing by. A large number of cars pass by on a nearby road.	A car drives by and then stops. Several cars are driving by on a busy road. Cars are driving by on the street and a woman is speaking. A person is walking down the street with cars driving by. A car drives by and people are talking.
LLM Output (FENSE)	Several cars and trucks are driving by on a busy street. (53.67%)	Cars are driving by on a busy road while a person is walking and people are talking. (70.35%)
Explanation	CLAP Reranking ranks the sampled captions based on their embedding similarity to a fixed-dimensional audio embedding, capturing only the aspects of the audio emphasized by the CLAP audio encoder. This process may overlook some events detected by the AAC model, such as the sound of people's voices in this example. In contrast, our proposed method selects the most salient concept clusters, better reflecting the full range of events recognized by the AAC model.	
Audio File: `Fountain Trompenburg 090928.wav`		
	CLAP Reranking	**Our Proposed Method**
Candidate Selection	A stream is flowing over rocks as people chatter and walk. Water is flowing in a creek as people talk and walk. Water is flowing as people talk and walk by. Water is flowing as people talk and walk through a stream. A stream of water flows while people talk and walk.	Water is flowing down a stream as people talk in the background.
LLM Output (FENSE)	Water is flowing in a stream as people talk and walk by. (45.77%)	Water is flowing down a stream as people talk in the background. (53.31%)
Explanation	The low diversity among the sampled captions in this example indicates that the AAC model was highly confident about the events in the audio. This is further supported by the fact that our method identified only a single salient cluster. As a result, we skip LLM inference and directly use the centroid of this cluster as the final caption. This not only reduces computational overhead but may also improve evaluation scores, as the AAC model is trained to align with the target caption distribution, whereas the LLM, operating in a few-shot setting, is less familiar with the characteristics of AAC-generated captions.	

Table 1: Illustration of how different candidate selection methods affect the LLM's output.

waveform into the encoder, and the caption embedding, derived by feeding the generated caption into a pre-trained text encoder, i.e., INSTRUCTOR (Su et al., 2023). Finally, the generated captions are reranked using a weighted sum of the *decoder* (0.3) and *encoder* (0.7) scores, with the top-ranked caption selected as the system output.

However, our experiments revealed that the *decoder score* has a negligible effect and can be safely omitted without significantly impacting performance. Specifically, the system achieves FENSE scores of 52.13 and 50.17 when using only the *encoder* or *decoder score* for reranking, respectively, while the fused scores yield a performance of 52.28. This suggests that the success of the proposed reranking method relies heavily on the additional supervision signal provided by INSTRUCTOR during training, which prevents it from being applied to other pre-trained AAC models.

In DCASE 2024, Jung et al. (2024) introduced a model-independent reranking approach based on CLAP (Wu et al., 2023), a multimodal audio and text encoder that uses contrastive learning techniques to jointly embed these two modalities. Their approach is similar to the previous sampling and reranking method, with the key difference that they encode both the generated captions and the input audio using CLAP. Additionally, beyond utilizing CLAP for reranking, they proposed incorporating it as an additional filtering stage prior to the previously described hybrid reranking method. This filtering step removes half of the generated captions that are not sufficiently aligned with the audio embedding.

2.2 LLM-based Summarization

Given that LLMs have been proven effective across a range of zero-shot tasks, Jung et al. (2024) adopt an LLM-based caption summarization method. In this approach, a sampling and reranking strategy is first used to rank a set of sampled captions. Next, rather than selecting the top-ranked caption, the top-$\boldsymbol{K}$ captions are fed into an LLM with a zero-shot caption summarization prompt to generate the final caption. This method aims to enrich the final caption by combining key phrases that may be scat-

Prompt Template

You are provided with several candidate captions generated by an Automated Audio Captioning system for a specific audio file. These captions may contain repetitions, inaccuracies, or illogical details. Each caption may describe one or more main events. Identify the most frequent and relevant events from all the captions, and generate a single caption, logically describing the most probable events present in the original audio. Ensure the caption is free of punctuation marks, including commas.

Captions:
A car is driving down a road with the window open.
The rain is falling as a car passes by.
Water is flowing as a car passes by.
The rain is falling and the wind is blowing.
Generated Caption:
A car is passing by while the rain is falling and the wind is blowing.

Captions:
Cars are passing by on a busy road.
Cars drive by on a busy highway while a wind blows.
Cars drive by on a wet road.
A car is driving down the road and then the car drives by.
Generated Caption:
Cars are driving down a busy and wet road while the wind blows.

[more demonstrations]

Captions:
[selected candidates]

Table 2: Few-shot prompt template.

tered across different sampled captions, while also leveraging the LLM's ability to generate grammatically accurate sentences. However, in our experiments, we observe that the reranking stage considerably diminishes the diversity of the selected captions, often resulting in many identical captions, thereby reducing the effectiveness of LLM-based summarization.

2.3 LLM-based Error Correction

In their recent work, Liu et al. (2024) used an LLM as a post-corrector to address potential grammatical errors and repetitions in the captions generated by their AAC model, operating in a one-shot setting. In this approach, only a single caption sample from the AAC model is provided to the LLM for error correction.

3 Methodology

A major challenge in AAC arises from the inherent ambiguity of audio signals. Due to the overlapping and similar sound characteristics of many acoustic events, individuals may perceive the same audio differently, sometimes even with conflicting interpretations (Figure 1). To address this variability, popular audio captioning datasets, such as Clotho, provide multiple ground-truth captions from various annotators for each audio sample (Drossos et al., 2020). During training, models are exposed to one-to-many audio-caption mappings, with each audio clip paired with a randomly selected ground-truth caption in each epoch. This randomness can introduce uncertainty into the learned representations and degrade model performance (Zhang et al., 2023).

To examine how this uncertainty affects the output of AAC models, we randomly selected 50 audio samples from the Clotho dataset and generated 50 captions per audio sample with Nucleus Sampling using two pre-trained AAC models. A careful manual analysis of the generated captions revealed that the AAC model's confidence in the acoustic events of a given input audio is strongly reflected in the diversity of the sampled captions. Specifically, when the AAC model is confident about the audio content, nearly all sampled captions describe the same events, differing only in word choice and ordering. Conversely, when the input audio is ambiguous or challenging, the sampled captions display greater diversity, describing a range of possible events.

Thus, we hypothesize that the diversity of sampled captions can serve as an indicator of an AAC model's confidence level. Based on this hypothesis, we propose a post-processing method for AAC models with the following steps (Figure 2): First, we generate $\boldsymbol{N}$ captions for each input audio using Nucleus Sampling and encode them with a lightweight sentence encoder. Next, the encoded captions are clustered into $\boldsymbol{K}$ groups to identify the primary event clusters. The $\boldsymbol{K}$ cluster centroids, representing the primary possible events, are then fed into an LLM along with a few demonstrations to generate the final caption. When the selected captions describe similar events, the LLM is expected to produce a consistent caption with its inputs. However, when the diversity among the selected captions is high, the LLM should incorporate different possible events, resulting in a more diverse and comprehensive output. The following subsections provide a detailed explanation of each step, and Table 1 presents two illustrative examples.

3.1 Sampling and Candidate Selection

For each given input audio, we use Nucleus Sampling to generate a set of $\boldsymbol{N}$ diverse captions. We

AAC Model	Decoding & Post-Processing	FENSE (%)
CoNeTTE (Labbé et al., 2024)	Beam Decoding (width=5)	51.96
CoNeTTE	Beam Decoding (width=5) + LLM-based Error Correction (Liu et al., 2024)	51.60
CoNeTTE	Sampling + CLAP Reranking (Jung et al., 2024)	49.86
CoNeTTE	Sampling + CLAP Reranking + LLM-based Summarization (Jung et al., 2024)	53.32
CoNeTTE	Sampling + Ours	**53.76**
BEATs-Conformer-BART (Wu et al., 2024)	Beam Decoding (width=5)	50.35
BEATs-Conformer-BART	Beam Decoding (width=5) + LLM-based Error Correction	50.15
BEATs-Conformer-BART	Sampling + Hybrid Reranking (Wu et al., 2024)	52.28
BEATs-Conformer-BART	Sampling + Hybrid Reranking + LLM-based Summarization	52.63
BEATs-Conformer-BART	Sampling + CLAP Reranking	51.49
BEATs-Conformer-BART	Sampling + CLAP Reranking + LLM-based Summarization	52.71
BEATs-Conformer-BART	Sampling + CLAP Filtering + Hybrid Reranking	52.75
BEATs-Conformer-BART	Sampling + CLAP Filtering + Hybrid Reranking + LLM-based Summarization	52.89
BEATs-Conformer-BART	Sampling + Ours	**53.49**

Table 3: Results on the evaluation subset of Clotho.

then select a set of $\boldsymbol{K} = 5$ candidate captions that preserve the original diversity of events present in the generated captions (the first example in Table 1). To achieve this, we use a lightweight pre-trained off-the-shelf text encoder, Sentence-BERT (Reimers and Gurevych, 2019), to encode the captions into vector embeddings, and then apply Agglomerative clustering with complete link to group them into $\boldsymbol{K}$ clusters. For each cluster, we compute the center point by averaging the embeddings of the captions within the cluster, and select the caption with the closest embedding to this center as the cluster representative. In this work, cosine similarity was used consistently across all embedding-based steps.

Moreover, to prevent the selection of too infrequent events that could mislead the LLM, we incorporate an outlier removal step during this phase, removing clusters with fewer than $\boldsymbol{R} = 5$ embeddings. Additionally, when the majority of embeddings fall into a single cluster (at least $\boldsymbol{C} = 72\%$ of the embeddings), indicating high confidence from the AAC model, we bypass the LLM phase and directly use the cluster representative as the system output (the second example in Table 1). This approach not only reduces the computational overhead of LLM inference but also enhances performance, as the AAC model is specifically trained to generate captions and is more adept at producing outputs that align with the target distribution. This simple yet effective step is also extendable to other LLM-based post-processing methods.

3.2 Few-shot Caption Diversity Enhancement

The selected captions are then processed by an LLM using a few-shot prompt to generate the final caption. When there is high diversity among the input candidate captions, the LLM is anticipated to generate a more diverse caption. Conversely, when the diversity is low, the LLM is expected to produce a caption that closely matches the inputs. Table 2 contains the prompt template used for this task. Since the primary goal of this study is to evaluate the impact of diversity-enhanced candidate selection, we did not focus on optimizing the number or content of demonstrations used in the LLM prompt. Instead, a fixed set of five manually crafted demonstrations was used across all inputs. This choice was supported by preliminary experiments, which indicated that four to six demonstrations are generally sufficient for reasonable LLM performance, depending on the model. Given that manually creating this small number of examples is straightforward, we leave the exploration of automatic demonstration optimization for future work. The complete list of demonstrations can be found in the accompanying source code.

4 Experimental Setup

4.1 Models

Our proposed post-processing method is independent of the AAC model. Thus, we conduct our experiments using two open-source models: CoNeTTE[2] (Labbé et al., 2024) and BEATs-Conformer-BART[3] (Wu et al., 2024). Additionally, GPT-4o-mini is used as the LLM in our experiments, accessed through the OpenAI API.

[2]https://github.com/Labbeti/conette-audio-captioning

[3]https://github.com/slSeanWU/beats-conformer-bart-audio-captioner

4.2 Hyperparameters

During the sampling phase of all methods, Nucleus Sampling was performed with a temperature of 0.5 and a top-p value of 0.95. Greedy decoding was used for all LLM-based stages. The max tokens parameter was set to 50 for both Nucleus Sampling and LLM generations.

The parameters $K = 5$ and $R = 5$ were selected based on intuition and preliminary experiments. We observed that moderate changes to these values do not significantly affect the results, and the chosen values offer a good balance that works well across a wide range of AAC models and LLMs. In contrast, $C = 0.72$ was determined through grid search on Clotho's validation subset.

4.3 Dataset

We conduct our experiments using the Clotho v2.1 dataset (Drossos et al., 2020), which served as the standard benchmark in previous DCASE scientific challenges. The dataset consists of four subsets. The *development* and *validation* subsets are intended solely for optimizing AAC models, while the *evaluation* subset is used for assessing and comparing results. The *testing* subset is reserved exclusively for scientific challenges, such as the DCASE challenge. To conform with this standard, we use only the *evaluation* subset of the dataset to compare and report our results.

4.4 Evaluation Metrics

We adopt the FENSE metric (Zhou et al., 2022), the standard evaluation metric of the DCASE 2024 challenge, as our evaluation metric. Prior to FENSE, AAC evaluation metrics were borrowed from machine translation and image captioning and focused on the surface form of the words (Labbé et al., 2024). FENSE, on the other hand, leverages pre-trained models to capture sentence meanings. It also penalizes grammatically incorrect or incoherent sentences.

5 Results and Discussion

The experimental results (Table 3) demonstrate the effectiveness of our proposed method compared to other post-processing approaches when applied to the outputs of two open-source AAC models. These findings underscore the importance of preserving the diversity of sampled candidates, particularly for LLM-based post-processing methods.

Method	FENSE (%)
Random Selection (K=5)	52.46
Random Selection (K=20)	53.05
All Candidates (K=50)	53.07
Clustering (K=5)	52.21
+ Outlier Removal	53.31
+ Skipping LLM Usage	**53.49**

Table 4: Ablation study of the candidate selection stages.

Additionally, Table 1 presents two concrete examples that illustrate how our method works in practice and provide intuition behind its effectiveness. In these examples, the same prompt template was used across different candidate selection methods to enable a fair comparison, ensuring that the observed improvements can be attributed solely to the proposed candidate selection strategy rather than differences in prompt design compared to prior studies.

5.1 Ablation Study

We conduct a comprehensive ablation study on the candidate selection phase, beginning with a random candidate selection method and gradually incorporating the proposed components. Table 4 shows that the clustering phase is significantly affected by outliers, leading to performance that falls behind random candidate selection. However, removing the outliers results in a notable improvement, emphasizing the importance of this step. Additionally, while including more samples in the prompt, up to selecting all generated captions, can slightly improve performance, it still lags behind the proposed clustering method. This is likely due to the large volume of redundant information the LLM must process, as well as the presence of outliers that represent highly unlikely events in the inputs. These findings underscore the importance of targeted candidate selection. Finally, skipping LLM inference when a single cluster contains more than C captions leads to additional performance gains. In this specific scenario, although the overall improvement across the entire Clotho evaluation set may appear modest, the FENSE score increases from 55.15 to 56.11 for the 116 samples where this condition applies (approximately 11% of the subset). This demonstrates that the method effectively identifies cases where the AAC model is confident

Stage	Time (ms)
Beam Decoding (width=5)	487
Nucleus Sampling (N=50)	991
Hybrid Reranking (Wu et al., 2024)	739
CLAP Reranking (Jung et al., 2024)	340
CLAP Filtering + Hybrid Reranking	833
Candidate Selection (Ours)	15
LLM Inference (GPT-4o-mini)	779

Table 5: Average processing time per sample (in milliseconds) for various decoding and post-processing methods.

and avoids unnecessary LLM processing.

5.2 Runtime Analysis

As depicted in Table 5, our proposed candidate selection stage is considerably faster than previous reranking strategies. The processing times were calculated by running the methods on the entire evaluation subset of Clotho v2.1 using a machine with a single Nvidia RTX 3090 GPU. The LLM inference time, which includes the HTTP request and response times as well, was measured on Google Colaboratory servers. Since this time was consistently similar across different inputs, with negligible variations, only the average time is reported. During each stage, parallelism was disabled, and all samples were processed sequentially. The AAC model used throughout all stages was BEATs-Conformer-BART.

6 Conclusion and Future Work

In this work, we explored various post-processing methods for automated audio captioning and proposed a novel LLM-based method for enhancing caption diversity. The proposed approach leverages in-context learning to consider the certainty of the AAC model, reflected in the diversity of its generated captions. Despite being considerably faster, our method demonstrates performance improvements over previous post-processing techniques, as evidenced by experiments conducted on two open-source models.

Future work could investigate the effectiveness of alternative embedding and clustering methods in the proposed candidate selection phase. Additionally, since the demonstrations in our prompt were manually crafted and remained fixed across all inputs, future research could improve performance by exploring automatic example generation techniques or employing more advanced prompting strategies.

7 Limitations

This study is limited to experiments conducted with a single LLM (GPT-4o-mini) due to resource limitations. A broader evaluation involving multiple LLMs could offer deeper insights into the strengths and limitations of LLM-based post-processing methods for AAC.

References

Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, Wanxiang Che, Xiangzhan Yu, and Furu Wei. 2023. BEATs: Audio pre-training with acoustic tokenizers. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 5178–5193. PMLR.

Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. 2020. Clotho: an audio captioning dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 736–740. IEEE.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Jee-weon Jung, Dong Zhang, HCH Yang, Shih-Lun Wu, David M Chan, Zhifeng Kong, D Ruifan, Z Yaqian, V Rafael, and Shinji Watanabe. 2024. Automatic audio captioning with encoder fusion, multi-layer aggregation, and large language model enriched summarization. Technical report, DCASE Challenge, Tech. Rep.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. 2019. 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, Volume 1 (Long and Short Papers)*, pages 119–132.

Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D Plumbley. 2020. PANNs: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2880–2894.

Étienne Labbé, Thomas Pellegrini, and Julien Pinquier. 2024. CoNeTTE: An efficient audio captioning system leveraging multiple datasets with task embedding. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 32:3785–3794.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. 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, Online. Association for Computational Linguistics.

Jizhong Liu, Gang Li, Junbo Zhang, Chenyu Liu, Heinrich Dinkel, Yongqing Wang, Zhiyong Yan, Yujun Wang, and Bin Wang. 2024. Leveraging CED encoder and large language models for automated audio captioning. Technical report, DCASE Challenge, Tech. Rep.

Chaitanya Prasad Narisetty, Tomoki Hayashi, Ryunosuke Ishizaki, Shinji Watanabe, and Kazuya Takeda. 2021. Leveraging state-of-the-art ASR techniques to audio captioning. In *Proc. Conf. Detection Classification Acoust. Scenes Events*, pages 160–164.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Chufan Shi, Haoran Yang, Deng Cai, Zhisong Zhang, Yifan Wang, Yujiu Yang, and Wai Lam. 2024. A thorough examination of decoding methods in the era of LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8601–8629, Miami, Florida, USA. Association for Computational Linguistics.

Hongjin Su, Weijia Shi, Jungo Kasai, Yizhong Wang, Yushi Hu, Mari Ostendorf, Wen-tau Yih, Noah A. Smith, Luke Zettlemoyer, and Tao Yu. 2023. One embedder, any task: Instruction-finetuned text embeddings. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1102–1121, Toronto, Canada. Association for Computational Linguistics.

Mengyue Wu, Heinrich Dinkel, and Kai Yu. 2019. Audio caption: Listen and tell. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 830–834. IEEE.

Shih-Lun Wu, Xuankai Chang, Gordon Wichern, Jee-weon Jung, François Germain, Jonathan Le Roux, and Shinji Watanabe. 2024. Improving audio captioning models with fine-grained audio features, text embedding supervision, and llm mix-up augmentation. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 316–320. IEEE.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. 2023. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Xuenan Xu, Zeyu Xie, Mengyue Wu, and Kai Yu. 2022. The SJTU system for DCASE2022 challenge task 6: Audio captioning with audio-text retrieval pre-training. *Tech. Rep., DCASE2022 Challenge*.

Zhongjie Ye, Yuexian Zou, Fan Cui, and Yujun Wang. 2022. Automated audio captioning with multi-task learning. In *Proc. Conf. Detection Classification Acoust. Scenes Events*, pages 1–3.

Yiming Zhang, Hong Yu, Ruoyi Du, Zheng-Hua Tan, Wenwu Wang, Zhanyu Ma, and Yuan Dong. 2023. ACTUAL: Audio captioning with caption feature space regularization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Zelin Zhou, Zhiling Zhang, Xuenan Xu, Zeyu Xie, Mengyue Wu, and Kenny Q. Zhu. 2022. Can audio captions be evaluated with image caption metrics? In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 981–985.