Graph-Score: A Graph-grounded Metric for Audio Captioning

Manh Luong[1], **Gholamreza Haffari**[1], **Dinh Phung**[1],**Lizhen Qu**[1]
[1]Monash University, Australia,
{tien.luong,gholamreza.haffari,dinh.phung,lizhen.qu}@monash.edu

Abstract

Evaluating audio captioning systems is a challenging problem since the evaluation process must consider numerous semantic alignments of candidate captions, such as sound event matching and the temporal relationship among them. The existing metrics fail to take these alignments into account as they consider either statistical overlap (BLEU, SPICE, CIDEr) or latent representation similarity (FENSE). To tackle the aforementioned issues of the current metrics, we propose the graph-score, which grounds audio captions to semantic graphs, for better measuring the performance of AAC systems. Our proposed metric achieves the highest agreement with human judgment on the pairwise benchmark datasets. Furthermore, we contribute high-quality benchmark datasets to make progress in developing evaluation metrics for the audio captioning task.

1 Introduction

Automated audio captioning (AAC) aims to generate textual descriptions of a given audio. There has been significant progress in the AAC task due to framework development (Kim et al., 2024), data curation (Mei et al., 2023), and prefix-tuning language model (Deshmukh et al., 2023; Kim et al., 2023). However, there is little progress in developing reliable evaluation metrics for the AAC task. Furthermore, evaluating AAC systems is challenging due to the diversity of reference captions in terms of style and content.

To assess the quality of AAC systems, the most popular metrics are BLEU (Papineni et al., 2002), SPICE (Anderson et al., 2016), and FENSE (Zhou et al., 2022). However, these metrics are not able to reflect the alignment between audio and candidate captions. The BLEU score is designed to measure n-gram overlap between candidate and reference sentences. Therefore, it is incapable of measuring semantic similarity. SPICE score is proposed to evaluate the semantic similarity for image captioning via semantic graph matching. However, it only focuses on object's attributes and their spatial relationships, which are not vital for audio captioning evaluation. To tackle the shortcomings of prior metrics, FENSE is developed to measure the semantic similarity for AAC systems by combining the Sentence-Bert score (Reimers and Gurevych, 2019) with a fluency penalization score. Although FENSE is effective and well-aligned with human judgment, it struggles to determine genuine temporal relations among sound events. As shown in Figure. 1, the caption **C1** is well-aligned with reference captions. While even the caption **C2** refers to the same sound events, it might describe a different audio due to the difference in temporal relations among sound events. Human raters are able to recognize the difference in temporal relations described in these two candidate captions and give a genuine judgment.

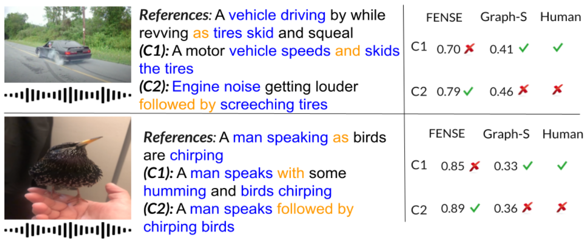

Figure 1: Several qualitative examples from the AudioCaps benchmark. The sound-events are highlighted in blue, and temporal relations are highlighted in orange.

To better compare the performance of AAC systems, we propose a new evaluation metric for the AAC task, coined *graph-score*. Our proposed metric first extracts semantic graphs from the candidate and reference captions to measure their dissimilarity. The semantic graph consists of a list of triplets, each triplet represents a pair of sound events and

Proceedings of The 23rd Annual Workshop of the Australasian Language Technology Association, pages 192–201
November 26-28, 2025

their temporal relationship. For example, the candidate caption *"A man speaks followed by chirping birds"* in Figure. 1 can be represented as *<man speaking, following by, bird chirping>*. There are several ways to express an acoustic event, such as paraphrasing or using a generic/specific expression. The diversity of acoustic event expression causes difficulty in measuring the discrepancy between two graphs. We map the extracted sound events to a predefined list of 527 audio events extracted from the AudioSet dataset (Gemmeke et al., 2017) which is a comprehensive ontology for acoustic events. Therefore, the semantic graph is a better representation of captions for measuring the alignment in terms of sound events and their temporal relationships. By leveraging semantic graphs, we utilize the optimal transport framework to measure the dissimilarity between candidate and reference captions, moreover, we also leverage the pretrained CLAP model (Elizalde et al., 2023) to compute semantic dissimilarity across audio and textual description. Finally, our graph-score metric is the convex combination of both semantic graph and cross-modal dissimilarity. To sum up, our key contributions are two-fold:

1. We propose a new evaluation metric, coined graph-score, for the AAC task to better assess the alignment between candidate caption and audio and a list of reference captions.

2. Due to the lack of high-quality benchmark datasets for developing automatic metrics, we extend subsets of AudioCaps and Clotho test sets with high-quality human judgements.

2 Methodology

2.1 Semantic Graph Construction

The semantic graph of a given audio describes the temporal relationship among audio events that occur in the audio. The semantic graphs consist of a list of triplets $G(c) = \{< e_1, r, e_2 >\}_{i=1}^{n}$, where e_1 and e_2 are two audio events occurring in a caption c that have the temporal relationship $r = \{$*following by*, *concurrent with*$\}$. As discussed in (Xie et al., 2023), the temporal relationships of audio events can be narrowed down to sequential or concurrent relationships to understand the audio content correctly. We formulate the semantic graph extraction from the caption as an open information extraction task. Given an audio caption, we can extract a corresponding semantic graph reflecting the temporal relationship among audio events in the caption. We conduct experiments on using either GPT-4 (Achiam et al., 2023) or LLaMa3.1-8B (Dubey et al., 2024) to construct semantic graphs from the caption of audio. The prompt design can be found in the Appendix C.

Grounded audio events. The extracted semantic graphs from GPT-4 consist of open-ended audio events from the given captions. Due to node mismatching, estimating the distance between a pair of graphs is challenging. Therefore, we propose to use a predefined audio events list to ground open-audio events from GPT-4 to assist in graph comparison. The predefined audio events list consists of 527 audio events extracted from the AudioSet dataset (Gemmeke et al., 2017). The rationale behind choosing AudioSet's labels is that this dataset covers a wide range of sound events in the wild. The covered sound events range from daily sounds like human and animal sounds to source-ambiguous sounds like surface contact. After extracting semantic graphs from a given audio caption, the extracted sound events from the caption are mapped to the AudioSet's sound events by nearest-neighbor search. We utilize the pretrained BERT model (Devlin et al., 2018) to embed both AudioSet's labels and the extracted sound events into the embedding space and then compute the similarity score between the extracted audio events and AudioSet's labels. Given an extracted audio event from the caption, it is mapped to the most similar semantic audio events in the predefined list as follows

$$b = \arg\max_{b_i \in \mathcal{B}} \mathrm{s}(f(e), f(b_i)) \tag{1}$$

where $\mathcal{B}$ is the list of AudioSet's labels. $s(.)$ and $f(.)$ are the cosine similarity and the embedding functions, respectively.

2.2 Graph-grounded Evaluation Metric

Optimal transport for semantic graph comparison. The size of a candidate graph is always smaller than the size of the reference graph unified from its reference captions. Hence, there is more than one matching between a candidate and a reference graph. Exact matching is not able to consider all matchings between two graphs to measure the distance between them. Therefore, We utilize the optimal transport framework to perform bipartite matching and then measure the discrepancy between semantic graphs. We embed triplets of extracted semantic graphs as described in the Section. 2.1 into an embedding space and then perform matching between two sets of embedding vectors. Each triplet in the semantic graph is verbalized by a template: "*The sound of* <e_1> *is* <r> *the sound of* <e_2>" to generate a textual description for the triplet. Formally, the candidate graph G_c and the unified reference graph $\mathcal{G} = \{G_{r_1}, ..., G_{r_N}\}$ are transcribed into two sets of textual descriptions. After that, textual descriptions are fed into the CLAP text encoder to achieve the embedding of the candidate graph $Z_c = \{z_c^i\}_{i=1}^n$ and the embedding of the unified reference graph $Z_{\mathcal{G}} = \{z_{\mathcal{G}}^j\}_{j=1}^m$. We utilize the optimal transport framework to perform point set matching between two sets of embedding vectors and then use the optimal matching solution to measure the discrepancy between them

$$D_{OT}(\mu^{G_c}, \nu^{\mathcal{G}}) = \min_{\pi \in \Pi(\mu^{G_c}, \nu^{\mathcal{G}})} \sum_{i=1}^{n} \sum_{j=1}^{m} \pi_{i,j}.c(z_c^i, z_{\mathcal{G}}^j) \tag{2}$$

where $\mu^{G_c} = \frac{1}{n}\sum_{i=1}^{n} \delta_{G_c}$ and $\nu^{\mathcal{G}} = \frac{1}{m}\sum_{j=1}^{m} \delta_{\mathcal{G}}$ are two discrete probability measure of candidate and reference embeddings, $\Pi(\mu^{G_c}, \nu^{\mathcal{G}}) = \{\pi \in R^{n\times m} | \pi 1_n = 1_n/n, \pi 1_m = 1_m/m\}$ denotes the set of transportation plans or coupling between μ^{G_c} and $\nu^{\mathcal{G}}$. The distance metric $c(z_c^i, z_{\mathcal{G}}^j)$ is defined as $1 - \frac{<z_c^i, z_{\mathcal{G}}^j>}{||z_c^i||.||z_{\mathcal{G}}^j||}$ to measure the distance between two embedding vectors.

Semantic graph-based score. Given a triplet (a, c, R) of an audio, a candidate caption, and a list of reference captions, we first compute the dissimilarity between the given audio and the candidate caption as $D(a, c) = 1 - \frac{<f(c), g(a)>}{||f(c)||.||g(a)||}$, where $f(\cdot)$ and $g(\cdot)$ are the pretrained text and audio encoder from CLAP model. Then, we compute the distance between the candidate caption and the list of references $D_{OT}(c, R)$ as in Eq. 2. The final score is the convex combination of audio-candidate caption distance and candidate-references caption distance

$$\text{GRAPH-S}(a, c, R) = \alpha D(a, c) + (1-\alpha) D_{OT}(c, R) \tag{3}$$

where $0 \leq \alpha \leq 1$. if $\alpha = 1$, the score is similar with CLIPScore (Hessel et al., 2021). If $\alpha = 0$, the score is based on the semantic graph distance.

3 Experiments

3.1 Eperimental settings

Evaluation datasets. We evaluate our proposed metric, graph-score, on two benchmark datasets. The first benchmark is the FENSE's benchmark (Zhou et al., 2022), which are two subsets of AudioCaps and Clotho test data. The FENSE's benchmark consists of 1,750 pairs on Clotho and 1,671 pairs on AudioCaps human judgments regarding audio caption quality. Although the FENSE's benchmark is the first curated dataset for evaluating audio captioning metrics, its inter-rater reliability is low, 0.28 on AudioCaps and 0.24 on Clotho. Therefore, we curate a new high-quality benchmark for better evaluating audio captioning metrics based on two AudioCaps and Clotho test data subsets. There are 400 samples of three human raters's preferences for AudioCaps and Clotho in our new benchmark. The data annotation detail is described in the Appendix D. As shown in the Table 1, our new benchmark datasets are more high-quality than the previous benchmark datasets (Zhou et al., 2022) in terms of rater-inter reliability due to the filtering procedure and a rigorous quality control process using the guideline in Appendix D.

	Fleiss Kappa	
	AudioCaps	Clotho
(Zhou et al., 2022)	0.28*(0.48)	0.24(0.33)
Our benchmark	0.53	0.42

Table 1: The inter-rater reliability for audio captioning benchmark based on human rating. *We recompute the inter-rater reliability of benchmarks in the FENSE paper and report the reliability in this table, the numbers in parentheses are the ones reported in the original paper.

Evaluation metrics. We measure the performance of audio captioning metrics by evaluating their correlation with human judgment. The evaluation metrics are evaluated in four scenarios: human-human caption correct (HC), human-human caption incorrect (HI), human-machine caption (HM), and machine-machine caption (MM). We consider the caption rated by the majority of human raters to be correct and measure how frequently the evaluation metrics assign a higher score to the correct caption of the pair.

3.2 Agreement with human judgment

Table 8 in the Appendix illustrates the agreement of evaluation metrics with human judgment for four evaluation scenarios on (Zhou et al., 2022) benchmarks. The graph-score achieves the highest agreement with human judgment on the human-human incorrect and human-machine caption scenarios on the AudioCaps benchmark, therefore, it aligns well

Metrics	AudioCaps					Clotho				
	HC	HI	HM	MM	Avg	HC	HI	HM	MM	Avg
BERTScore	60	51	49	49	52.25	58	53	53	60	56
BLEURT	62	84	61	72	69.75	58	91	67	60	69
Sentence-BERT	61	94	61	76	73	63	91	**71**	68	73.25
FENSE	61	94	63	**76**	73.5	63	91	71	67	73
Graph-score + GPT4	**65**	98	**67**	74	**76**	**64**	**98**	70	67	**74.75**
Graph-score + LLaMa3.1 8B	64	**99**	63	75	75.25	60	98	67	**68**	73.25

Table 2: Correlation with human judgment on our curated benchmark datasets sampled from AudioCaps and Clotho test sets. $\alpha = 0.6$ for both AudioCaps and Clotho benchmarks.

Metric	Clotho				
	HC	HI	HM	MM	Avg
BERTScore	57.1	95.5	70.3	61.3	67.5
BLEURT	59	93.9	75.4	67.4	71.6
Sentence-BERT	60	95.5	75.9	**66.9**	71.8
FENSE	**60.5**	94.7	75.8	66.8	**74.4**
Graph-score+ GPT4	56.9	**97.1**	**77.1**	64.6	73.9

Table 3: Correlation with human judgment on Clotho benchmark from (Zhou et al., 2022) with $\alpha = 0.6$. See Table. 8 in the Appendix. E for experiment on both AudioCaps and Clotho.

Figure 2: Failure cases on AudioCaps benchmark in which Graph-score failed to align with human judgment.

with human judgment. On the other hand, an identical finding is observed in the Clotho dataset, and our proposed metric achieves comparable performance with the state-of-the-art metric, FENSE.

We also recognize an issue for the previous benchmark datasets: low inter-rater reliability. Furthermore, the ranking of evaluation metrics on the Clotho benchmark is different, as shown in Table 2 and Table 3. Previous benchmarks used outdated audio captioning systems to generate annotation data with numerous grammatical errors. Due to model development, these types of errors rarely occur in state-of-the-art AAC systems, but the hallucination issue is a more critical issue for the current AAC systems. Therefore, we annotate new high-quality benchmarks using state-of-the-art AAC systems to generate annotation data for better evaluating the new metrics. As shown in Table 2, the graph-score significantly outperforms all baseline methods on the AudioCaps benchmark, while our metric is 1.75 points better than the FENSE metric on average measured on the Clotho benchmark. We further provide Spearman's Correlation shown in Table. 4. The graph-score is the most correlated metric with human preference, therefore, it can better evaluate machine-generated captions.

	AudioCaps(Spearman's ρ)	Clotho(Spearman's ρ)
BertScore	0.04	0.122
BLEURT	0.392	0.375
Sentence-BERT	0.46	0.46
FENSE	0.462	0.445
Graph-score	**0.512**	**0.492**

Table 4: Spearman Correlation between human preferences and metric preferences. All p-values < 0.05.

3.3 Failure analysis

Figure. 2 demonstrates cases in which the Graph-score fails to align with human judgment on AudioCaps benchmark. The major failure is due to the inability to comprehend the importance of sound events. Some sound events are referred to multiple times in reference captions, thereby, they are more crucial and should be weight with higher score. In the failure examples, the graph-score failed to take the importance of "*bird chirping*" in the top-left example into account, therefore, it is not well aligned with human judgment.

4 Conclusion

To better assess the quality of AAC systems, we propose a new evaluation metric, graph score, grounded in audio and semantic graphs. The experimental results on benchmark datasets demonstrate the superior agreement of our proposed metric with human judgments. The graph-score is able to measure the discrepancy in terms of sound-events and their temporal relations, therefore provide a better metric for evaluating the quality of machine-generated captions. We also contribute new high-quality benchmark datasets to facilitate the development of more reliable automatic evaluation metrics for the AAC task.

Limitation

Our proposed metric, the graph-score, has a few limitations. First, the graph-score metric is a model-based evaluation metric, therefore, its performance heavily depends on the quality of the pretrained CLAP model. Second, we primarily utilize ChatGPT, an API LLM, to extract semantic graphs from audio captions. It is worth exploring the other open-source LLMs such as Llama3 or Vicuna to reduce the inference costs and latency. We plan to use open-source LLMs to develop a totally transparent evaluation metric for audio captioning.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pages 382–398. Springer.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Ke Chen, Xingjian Du, Bilei Zhu, Zejun Ma, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. 2022. Hts-at: 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.

Soham Deshmukh, Benjamin Elizalde, Rita Singh, and Huaming Wang. 2023. Pengi: An audio language model for audio tasks. *Advances in Neural Information Processing Systems*, 36:18090–18108.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. 2023. Clap learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *Proc. IEEE ICASSP 2017*, New Orleans, LA.

Félix Gontier, Romain Serizel, and Christophe Cerisara. 2023. Spice+: Evaluation of automatic audio captioning systems with pre-trained language models. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*.

Anwen Hu, Shizhe Chen, Liang Zhang, and Qin Jin. 2023. Infometic: An informative metric for reference-free image caption evaluation. *arXiv preprint arXiv:2305.06002*.

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.

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

Minkyu Kim, Kim Sung-Bin, and Tae-Hyun Oh. 2023. Prefix tuning for automated audio captioning. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Xinhao Mei, Chutong Meng, Haohe Liu, Qiuqiang Kong, Tom Ko, Chengqi Zhao, Mark D Plumbley, Yuexian Zou, and Wenwu Wang. 2023. Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research. *arXiv preprint arXiv:2303.17395*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Ji Qi, Chuchun Zhang, Xiaozhi Wang, Kaisheng Zeng, Jifan Yu, Jinxin Liu, Jiuding Sun, Yuxiang Chen, Lei Hou, Juanzi Li, et al. 2023. Preserving knowledge invariance: Rethinking robustness evaluation of open information extraction. *arXiv preprint arXiv:2305.13981*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

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. *arXiv preprint arXiv:1908.10084*.

O Sainz, I García-Ferrero, R Agerri, OL de Lacalle, G Rigau, and E Agirre. Gollie: Annotation guidelines improve zero-shot information-extraction, corr abs/2310.03668,(2023). doi: 10.48550. *arXiv preprint ARXIV.2310.03668*.

Thibault Sellam, Dipanjan Das, and Ankur P Parikh. 2020. Bleurt: Learning robust metrics for text generation. *arXiv preprint arXiv:2004.04696*.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Zhen Wan, Fei Cheng, Zhuoyuan Mao, Qianying Liu, Haiyue Song, Jiwei Li, and Sadao Kurohashi. 2023. Gpt-re: In-context learning for relation extraction using large language models. *arXiv preprint arXiv:2305.02105*.

Xingyao Wang, Sha Li, and Heng Ji. 2022. Code4struct: Code generation for few-shot event structure prediction. *arXiv preprint arXiv:2210.12810*.

Zeyu Xie, Xuenan Xu, Mengyue Wu, and Kai Yu. 2023. Enhance temporal relations in audio captioning with sound event detection. *arXiv preprint arXiv:2306.01533*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

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. IEEE.

A Related Works

Statistics-based evaluation. This line of evaluation compares statistical overlap of candidate and reference captions to determine alignment between them, such as BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), ROUGE$_L$ (Lin, 2004), and CIDEr (Vedantam et al., 2015). SPICE (Anderson et al., 2016) is a scence-graph-based evaluation metric for the captioning tasks.

Model-based evaluation. This type of evaluation leverages pretrained models to assess the quality of generated captions. ClipScore (Hessel et al., 2021) and InfoMetIC (Hu et al., 2023) are reference-free evaluation metrics for audio captioning by utilizing pretrained CLIP encoders (Radford et al., 2021). Sentence-Bert (Reimers and Gurevych, 2019), BertScore (Zhang et al., 2019), and BLEURT (Sellam et al., 2020) are evaluation metric for text-generation tasks by fine-tuning the pretrained BERT model. Recently, FENSE (Zhou et al., 2022) proposed a state-of-the-art evaluation metric to assess the quality of audio caption by combining the Sentence-Bert score and the fluency score. SPICE+ (Gontier et al., 2023) is a modification of SPICE which leverages a pretrained language model for semantic graph extraction and soft-matching for sound-events comparison. All the prior evaluation metrics are either text-generation metrics or text-based evaluation metrics, therefore, they lack an understanding of the alignment between audio and generated captions. In addition, there have not been evaluation metrics considering the temporal relationship of sound events in audio to assess the quality of candidate captions.

B Model details

CLAP (Elizalde et al., 2023) is a cross-modal audio-text retrieval model trained on 128k audio-text pairs from 4 datasets. The CLAP model consists of two encoders, text and audio encoders, trained using the contrastive learning method to bridge the modality gap between audio and captions. The audio encoder is HTSAT model (Chen et al., 2022), which is pretrained on 2M audio clips from the AudioSet dataset for sound event tagging. The text encoder is GPT2 model (Radford et al., 2019), which is pretrained on text data for language modeling.

C Prompt Design

Prompt to extract semantic graph from captions

Instruction: Given an audio caption consisting of some sound events. You are able to extract sound events and temporal relations among sound events using the following template: [sound event| relationship| sound event]. There are two possible temporal relations: concurrent with and following by.

Examples: {In-context examples}
Input: {caption}

Prompting for semantic graphs extraction. Recently, large language models (LLMs) have achieved a great performance in open information extract tasks (Qi et al., 2023; Sainz et al.). The given caption is concatenated with a predefined prompt to input to LLMs to extract a corresponding semantic graph. We give detailed instructions on information extraction from a given audio caption to construct a semantic graph from the caption. The prompt is used to extract the semantic graph from the caption shown in the above table. Although GPT-4 is capable of extracting relevant information from a given audio caption, there are two problems with information extraction using LLMs: inconsistent responses and incomparable performance with fine-tuning models. We resolve the aforementioned issues of LLMs by adopting the in-context learning technique for information extraction (Wan et al., 2023; Wang et al., 2022). We provide representative examples as a part of the input prompt to assist GPT-4 in better understanding the information extraction task for the audio caption. The final prompt is utilized to extract the semantic graph of the audio caption is

$$\mathcal{P} = \mathcal{I} \cup \mathcal{D} | \mathcal{D} = d_1, ..., d_k \tag{4}$$

, where $\mathcal{I}$ and $\mathcal{D} = d_1, ..., d_k$ are the instruction of information extraction for audio caption and k-representative examples for the task.

D Dataset construction and annotation guideline

AudioCaps (Kim et al., 2019) and Clotho (Drossos et al., 2020) are two popular datasets for training

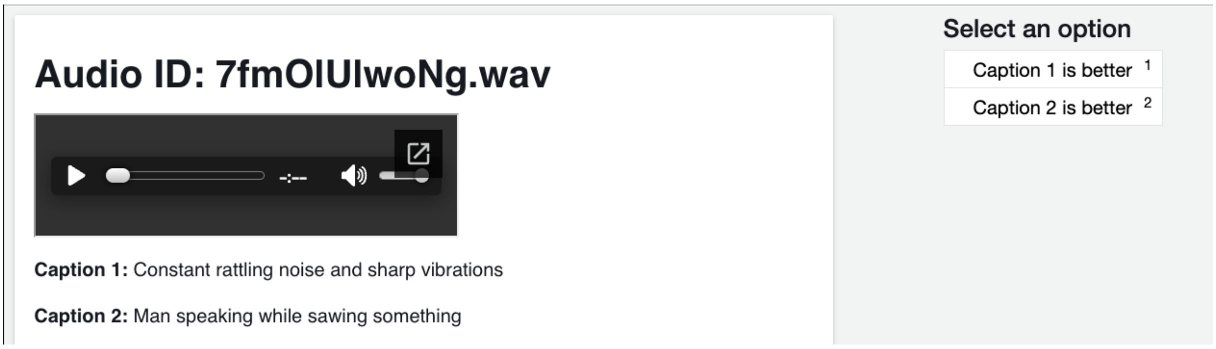

Figure 3: An overview of annotation platform

and evaluating audio captioning systems. There are currently 959 and 1045 available audio for the AudioCaps and Clotho test data, respectively. To construct a high-quality benchmark dataset for the audio captioning task, we annotate a representative subset of test data for each dataset since annotating the whole test data for these datasets is time-consuming and costly. We first cluster all audio in the test set of two datasets into 20 clusters using the K-mean algorithm on the audio embedding from the audio encoder from the CLAP model (Elizalde et al., 2023). Then, there are 5 audio samples extracted from each cluster as representative samples for the cluster. To avoid ambiguity for the human raters, we choose representative audio based on the criteria: the semantics of ground-truth captions of representative audio should be diverse, meaning the cosine similarity between reference caption embeddings should not be high. Eventually, we sample 100 audio and their captions from both AudioCaps and Clotho test data to build two benchmark datasets.

We follow the previous work (Zhou et al., 2022) to construct the audio captioning evaluation dataset based on human judgment. Given a triplet, an audio and two candidate captions, three annotators are asked to pick which candidate caption describes sound events in the audio better in terms of correctness and fluency. The annotation guideline and interface are demonstrated in the appendix D. We build four pair caption groups: human-human correct (HC), human-human incorrect (HI), human-machine (HM), and machine-machine (MM). The HC consists of two out of five audio reference captions. The HI also contains two human-written captions, one sampled from the given audio's reference captions and the other randomly sampled from a pool of reference captions. The HM is built from a human-written caption sampled from the audio's reference captions and a machine-generated caption for the same audio. The MM is built from two machine-generated captions describing the same audio. Two state-of-the-art audio captioning systems, Enclap (Kim et al., 2024) and Pengi (Deshmukh et al., 2023), are utilized to generate machine-generated captions. To give a clear instruction guideline and avoid disagreement during the annotation stage, we perform a dry-run for 20 samples for each dataset and discuss with annotators regarding our guideline and the dry-run annotation.

Annotation guideline. We give a detailed instruction to help human raters annotate benchmark data with high quality and agreement. The overview of the annotation platform is demonstrated in Figure. 3. Given an audio with a pair of candidate captions, human raters are asked to identify sound events described in each candidate caption and then identify their temporal relation. There are two valid temporal relations: sequential and concurrent relations. For example, the caption *"Constant rattling noise and sharp vibrations"*, there are two sound-events described in the given caption: rattling noise and sharp vibration, and their relationship is concurrent. After that, human raters listen to the audio at least twice to determine which candidate caption is more aligned with the audio.

E Additional ablation studies

We conduct an ablation study on a range of $0 \leq \alpha \leq 1$ to choose the best value of α for two benchmark datasets. As shown in Figure. 4. The highest performance of the graph score metrics is 77.5% with $\alpha = 0.8$ on the AudioCaps benchmark and 74.75% with $\alpha = 0.6$ on the Clotho benchmark. The experimental results show that both audio-caption and graph distance are vital for evaluating audio caption. The audio-caption distance is capa-

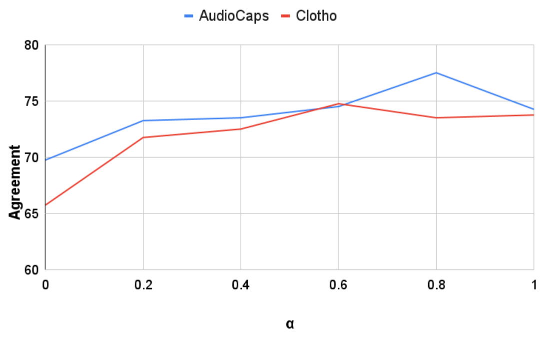

Figure 4: Ablation study on AudioCaps and Clotho benchmark for the graph-score metric with $0 \leq \alpha \leq 1$.

ble of measuring semantic alignment across audio and natural language modalities, while the graph distance is able to take sound-event matching and their temporal relationship into account for measuring the discrepancy between candidate and reference captions We further conduct ablation studies

	AudioCaps	Clotho
1-reference	76.25	73.25
3-reference	77.5	74.75

Table 5: Ablation study on the effectiveness of the unified reference graph by varying the number of reference captions.

on the effectiveness of the unified reference graphs by varying the number of reference captions. The experimental results are shown in Table. 5. There is a drop in terms of performance of our metric, the agreement decreases from 77.5% to 76.25% and from 74.75% to 73.25% on AudioCaps and Clotho benchmark, respectively. We also conducted an extensive study on the quality of extracted semantic graphs to our proposed metrics, the ablation results can be found in the Appendix. E

	Precision	Recall	F1
GPT-3.5	68.2	71.4	69.8
GPT-4	82.5	80.1	81.3

Table 6: The performance of GPT3.5 and GPT-4 on the semantic graph extraction task on 200 random human-written captions on the AudioCaps test set.

We conducted a study on the quality of LLM on semantic graph extraction from audio captions and then examined the effect of the quality of extracted semantic graphs on the performance of our metric. We first randomly sample 200 human-written captions from the AudioCaps test set and then leverage

	AudioCaps	Clotho
GPT-3.5	75	72.5
GPT-4	77.5	74.75

Table 7: Experiment on using GPT-3.5 and GPT-4 for semantic graph extraction from audio captions for the graph-score metric.

GPT-4 to extract semantic graphs, the prompt and procedure are described in the Appendix. C. After that, an expert, one of the authors of this paper, manually checks the extracted semantic graphs and relabels them if needed. We use the human-labeled data as ground-truth for the semantic graph extraction task. The performance of two LLMs, GPT-3.5 and GPT-4, on the semantic graph extraction task, is reported in Table. 6. The performance of GPT-3.5 is significantly lower than GPT-4 in terms of F1 score in extracting semantic graphs from audio captions; thereby, this reflects the lower performance of using GPT-3.5 in graph-score metric than using GPT-4 in Table. 7

To examine the influence of matching methods for semantic graph comparison in the graph-score, we compare the OT matching with exact matching as a baseline. The experimental results are shown in Table. 9. The exact distance is computed as follows $c(z_i, z_j) = 1$, if $z_i = z_j$, otherwise $c(z_i, z_j) = 0$

Metrics	AudioCaps					Clotho				
	HC	HI	HM	MM	Avg	HC	HI	HM	MM	Avg
BERTScore	60.6	97.6	92.9	65	74.3	57.1	95.5	70.3	61.3	67.5
BLEURT	**77.3**	93.9	88.7	72.4	79.3	59.0	93.9	75.4	67.4	71.6
Sentence-BERT	64	99.2	92.5	73.6	79.6	60	95.5	75.9	**66.9**	71.8
FENSE	64.5	98.4	91.6	**76.6**	82.7	**60.5**	94.7	75.8	66.8	**74.4**
Graph-score	72.6	**99.1**	**93.2**	73.1	**84.5**	56.9	**97.1**	**77.1**	64.6	73.9

Table 8: Correlation with human judgment on AudioCaps and Clotho benchmark from (Zhou et al., 2022). $\alpha = 0.6$ for both AudioCaps and Clotho benchmarks.

	AudioCaps	Clotho
Exact matching	52.25	44.25
Optimal transport	69.75	65.75

Table 9: Ablation study on our benchmarks to evaluate the performance of matching methods for graph comparison with $\alpha = 0$. The reported numbers are the average of correlation with human judgment.