X-ACE: Explainable and Multi-factor Audio Captioning Evaluation*

Qian Wang[1], **Jia-Chen Gu**[3], **Zhen-Hua Ling**[1,2†]
[1]NERC-SLIP, University of Science and Technology of China
[2]MoE Key Laboratory of Brain-inspired Intelligent Perception and Cognition,
University of Science and Technology of China
[3]University of California, Los Angeles
wangq621@mail.ustc.edu.cn, gujc@ucla.edu, zhling@ustc.edu.cn

Abstract

Automated audio captioning (AAC) aims to generate descriptions based on audio input, attracting exploration of emerging audio language models (ALMs). However, current evaluation metrics only provide a single score to assess the overall quality of captions without characterizing the nuanced difference by systematically going through an evaluation checklist. To this end, we propose the **explainable** and multi-factor **a**udio **c**aptioning **e**valuation (X-ACE) paradigm. X-ACE identifies four main factors that constitute the majority of audio features, specifically *sound event*, *source*, *attribute* and *relation*. To assess a given caption from an ALM, it is firstly transformed into an audio graph, where each node denotes an entity in the caption and corresponds to a factor. On the one hand, graph matching is conducted from part to whole for a holistic assessment. On the other hand, the nodes contained within each factor are aggregated to measure the factor-level performance. The pros and cons of an ALM can be explicitly and clearly demonstrated through X-ACE, pointing out the direction for further improvements. Experiments show that X-ACE[1] exhibits better correlation with human perception and can detect mismatches sensitively.

1 Introduction

Recognizing the pivotal role of auditory perception in human cognition, there is a trend among multimodal large language models (MLLMs) (Liu et al., 2023a; Dai et al., 2023) to broaden their scope into the audio-language domain. It has given rise to audio language models (ALMs) (Chu et al., 2023; Tang et al., 2023; Huang et al., 2023b), with a notable emphasis on automated audio captioning (AAC) for generating comprehensive descriptions of provided audio clips.

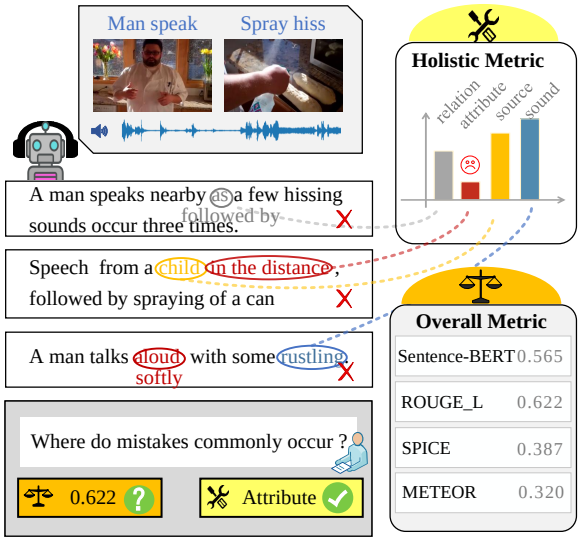


Figure 1: Different types of the errors in AAC task of ALMs. On the right are the quality scores from the different methods of metrics.

With the advancement of ALMs, there is an urgent need for a comprehensive and fine-grained assessment of their generated captions. Since current ALMs generally meet the requirements for depicting salient features of audio, nuanced yet crucial features are still ignored or misdescribed (Takeuchi et al., 2023). As depicted in Figure 1, captions inferred by ALMs have hallucinated in sound activities that do not exist ("*rustling*"), or confused the order of sounds (between "*man speak*" and "*spray hiss*") (Wu et al., 2023; Huang et al., 2023a), demonstrating the significant shortcomings. Metrics used to evaluate the quality of captions include the conventional ones such as ROUGE (Lin, 2004), BLEU (Papineni et al., 2002) and METEOR (Banerjee and Lavie, 2005) on word overlaps, and BERTScore (Zhang et al., 2020) and Sentence-BERT (Reimers and Gurevych, 2019) on semantic similarity. However,

* This work was supported in part by the National Natural Science Foundation of China under Grant U23B2053
† Corresponding author.
[1] The open-source dataset and code are available on https://github.com/wangqian621/X-ACE

Findings of the Association for Computational Linguistics: ACL 2024, pages 12273–12287
August 11-16, 2024

these metrics cannot be used to measure these shortcomings of existing work, since only a single score is provided to assess the overall quality of captions, thus failing to answer the question: "Where do mistakes commonly occur?

To this end, we explore a comprehensive direction of building an automatic evaluation paradigm for audio caption, and propose the **explainable** and multi-factor **a**udio **c**aptioning **e**valuation (X-ACE) paradigm. X-ACE consists of four factors tailored for audio caption including *sound event*, *source*, *attribute* and *relation*, each of which reflects the quality of caption in terms of a specific perspective, rather than providing a single score. To explicitly model the association between these factors, an audio caption to be assessed is transformed into an audio graph, where each node denotes an entity in the audio caption and corresponds to a designed factor. To calculate the score for each factor, the nodes from the audio graph corresponding to this factor are extracted for comparison with the reference through designed graph matching. In this way, we can assess the quality of a caption from multiple perspectives, and enhance the interpretability of the evaluation process.

To facilitate the proposed evaluation paradigm and address the granularity deficiency in the current test dataset, this paper presents a novel dataset for assessment. As for the conventional dataset AudioCaps (Kim et al., 2019) widely used in assessment, each caption includes much less information than the audio itself, which leads to omissions of audio characteristics. To tackle this, a dataset *AudioCaps-F* is constructed based on the AudioCaps, with **f**ine-grained annotations from domain experts. This dataset provides specific sound events along with corresponding sound sources and attributes, rather than descriptions in sentence form, promising completeness for further evaluation.

To demonstrate the correlation of the proposed evaluation paradigm with human judgement, X-ACE is compared with current automatic evaluation metrics including Sentence-BERT (Reimers and Gurevych, 2019), BERTScore (Zhang et al., 2020), and CIDEr (Vedantam et al., 2015). Experimental results of pair-wise tests show that X-ACE exhibits remarkable correlation with human subjective evaluation. Furthermore, extensive experiments also show that X-ACE exhibits better ability to detect the inconspicuous mismatches of a caption, as well as better capability of temporal relation reasoning.

Our main contributions can be summarized as follows: (1) An explainable and multi-factor evaluation paradigm X-ACE is proposed for audio captioning, demonstrating better correlation with human judgments and mismatch detection ability. (2) A dataset *AudioCaps-F* with refined annotations is constructed to facilitate further research in this field. (3) A comprehensive and empirical evaluation of existing ALMs based on X-ACE has been conducted, underscoring and analysing the problems with current ALMs.

2 Related Work

Audio Captioning Metrics Presently, there are three types of metrics employed in evaluating AAC. 1) Word Overlapping Metrics: These include ROUGE (Lin, 2004), METEOR(Banerjee and Lavie, 2005), and BLEU (Lin, 2004), which measure the overlap of n-gram words between the generated caption and the reference.2) Semantic Similarity Metrics: Metrics like BERTScore (Zhang et al., 2020; Zhou et al., 2022), Sentence-BERT (Reimers and Gurevych, 2019) assess the semantic similarity between captions. 3) Image Caption Metrics: CIDEr (Vedantam et al., 2015) focuses on n-grams of TF-IDF (Jones, 2004). SPICE (Anderson et al., 2016) transforms caption into a scene graph, and then calculating the graph F1 score. These metrics only offer a final score, overlooking crucial yet elementary error.

Aspect-level Evaluations In the domains of vision and language, evaluation methods now aim to discern performance across multiple aspects, rather than providing a single overall score. For image captioning, some hallucination evaluations for MLLMs (Li et al., 2023; Zhou et al., 2023) focus on visible object, AMBER (Wang et al., 2023) comprehensively assesses existence, attribute and relation hallucination. For video captioning, COAHA (Ullah and Mohanta, 2022) detects object and action hallucination. FactVC (Liu and Wan, 2023) classifies factual errors into categories (person, adjective, etc.). X-EVAL (Liu et al., 2023b) employs text evaluation on naturalness, coherence aspects. These methods systematically assess MLLMs but are inherited from visual or language concepts, making it hard to translation to the audio domain. Furthermore, they merely detect non-existent entities, neglecting the critical omission.

Explainable Evaluations Recent demand for explainability in evaluation metrics has grown significantly. INSTRUCTSCORE (Xu et al., 2023) estab-

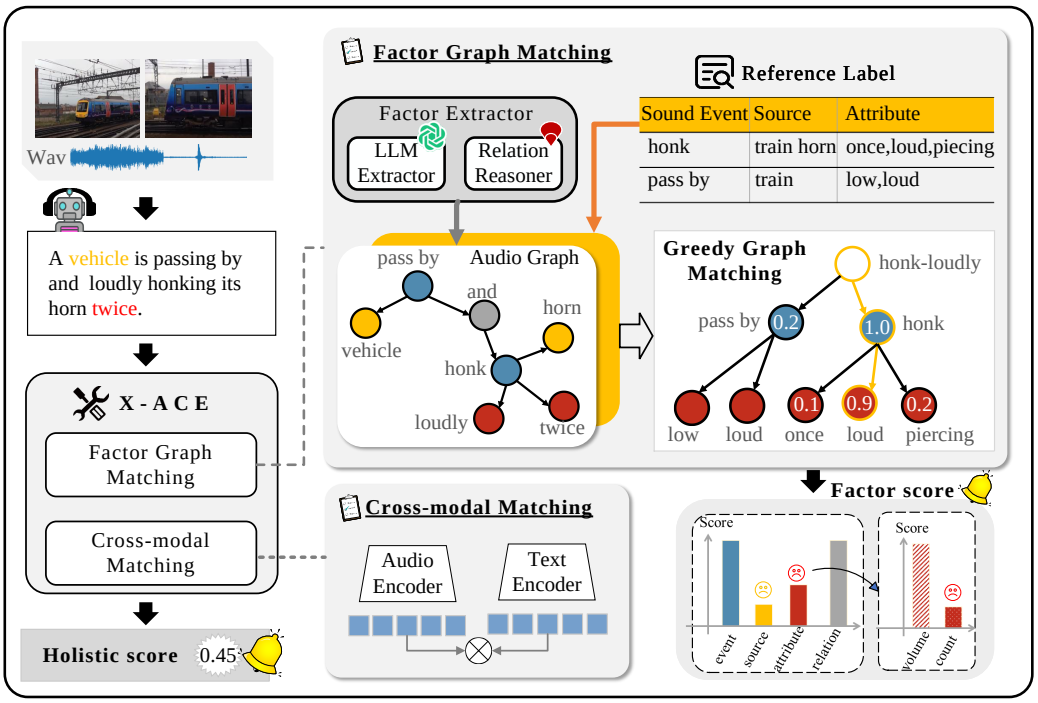


Figure 2: The overview of X-ACE benchmark, which is primarily composed of greedy graph matching and the cross-modal matching. The blue/yellow/red/gray circular nodes in the audio graph represent sound event, source, attribute, and relation respectively.

lishes a fine-grained explainable evaluation in natural language generation (NLG) task. EAPrompt (Lu et al., 2023), AUTOMQM (Fernandes et al., 2023) attempt to use error analysis on human-like translation evaluation. Metrics from VIEScore (Ku et al., 2023) explain the reason for scoring in image systhesis evaluation. InfoMetIC (Hu et al., 2023) reports fine-grained scores with explainable proof with specific incorrect words or image regions. However, they overly emphasize pinpointing the exact issues rather than categorizing them as a deficiency in a particular aspect. Furthermore, these identified issues fail to full coverage modal information, hindering the establishment of a unified and comprehensive metric.

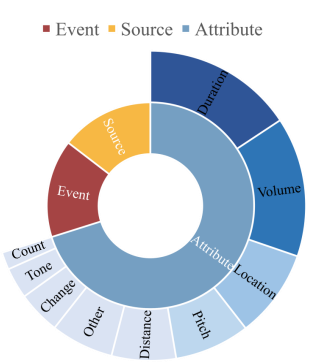
3 X-ACE

In this section, we commence with defining the audio factors in Section 3.1, which are key components of X-ACE for assessment in Section 3.2.

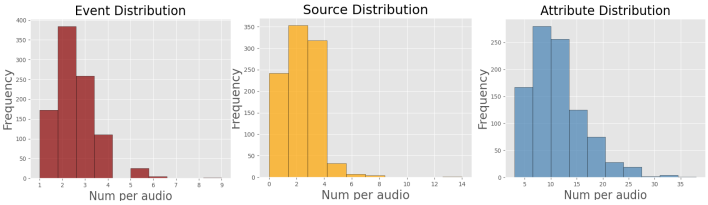
3.1 Audio Factors

Definition of Audio Factors As depicted in Figure 1, errors in captions arise from different perspectives. In this paper, four factors covering nearly all error issues and audio information are defined including *sound event*, *source*, *attribute* and *relation*. First of all, *sound event* denotes the specific sound activity occurring in an audio like "crying" and "speaking". We have observed that the most common and vital omission occurs in sound event. Models tend to describe prominent sounds while overlooking less significant ones or background noises. Secondly, *source* denotes the object producing the sound. By assessing this factor, we can detect subtle yet significant differences when the same event is emitted by different objects. Thirdly, *attribute* denotes auditory characteristics of the sound. This factor highlights nuanced audio features that often overlooked in descriptions. Last but not least, *relation* denotes temporal order between sounds, which is as crucial as spatial relation in the vision domain. In this way, four factors complement and interdepend on each other, collectively forming a comprehensive description of the audio content.

Definition of Audio Graph Based on audio factors, a caption is transformed into an audio graph, akin to scene graph (Johnson et al., 2015; Schuster et al., 2015) in vision. The graph in the center of Figure 2 comprises two tiers of nodes of *parent* and *child* nodes. On the one hand, a parent node is formed by an entity belonging to sound event, denoted as E_i representing the i-th sound event in an audio, where $i = 0, .., I-1$ and I denotes the number of sound events in this caption.



(a) Annotation Total Distribution



(b) Factor Distribution /per audio

	Sound Event	Source	Attribute
Average Num	2.4 /per audio	0.96 /per event	4.58 /per event
Total Num	2153 entities	2067 entities	9856 entities

Change	Count	Duration	Volume	Tone	Distance	Other	Pitch	Location
631	257	2206	2049	454	941	917	1100	1295

(c) Annotation Statics

Figure 3: The distribution of annotation data in the AudioCaps-F dataset: (a) depicts the proportions of each factor and subcategories of attribute, (b) showcases the distribution of factors describing a clip of audio, while (c) presents the total count of various labels.

On the other hand, the corresponding child nodes include source $S_{i,j}$ and attribute $A_{i,j}$, representing the j-th source/attribute of the i-th sound event, and the child node $R_{i,k}$ on the edge connecting the i-th event with the k-th event represents the temporal relation between them.

Factor Annotation To use audio graphs as labels and further enhance the completeness of references, a fine-grained dataset *AudioCaps-F* is constructed base on AudioCaps, which is widely used in AAC task. We hired domain experts to meticulously annotate entities corresponded to three factors grounded each audio, comprising sound events and their respective sources and attributes. Audio was annotated with an average of three annotators per sample. Subsequently, an automated program and manual review were employed for checking and refinement, the output format is shown in Appendix A. Different from the past practice of using only a few sentences of human descriptions as labels, this detailed annotation is the first attempt on the AAC task, supporting a more comprehensive and robust evaluation. The occurrence distribution of different audio factors are depicted in the Figure 3. The attributes are divided into nine subcategories for finer assessment on ALMs.

3.2 Evaluation Steps

Overview Our X-ACE pipeline, as illustrated in Figure 2, to evaluate a given caption inferred by an ALM, it is firstly decomposed into an audio graph using a factor extractor, which consists of an LLM extractor and a relation reasoner. In the assessment computation phase, the process diverges is divided into two streams. In one stream,

Schematic	Relation Rule
X ⇄ Y (before/after; after/before)	X occurs earlier/later than Y ↔ Y occurs later/earlier than X
X and Y	X and Y occur simultaneously
X → Z → Y	Employing the intermediary Z, X⇒Y is deduced through the steps X⇒Z and Z⇒Y.

Table 1: The types of temporal relations with their rules.

the predicted graph undergoes a greedy matching with the reference graph. This branch provides both factor-level and overall graph assessment. Simultaneously, in the other stream, a cross-modal similarity is calculated and then integrated with the graph score to produce a holistic evaluation.

Factor Extraction. To obtain factor nodes in our audio graph, we concatenate the given caption with a designed prompt and a template, and feed it into an LLM extractor. ChatGPT (GPT3.5-Turbo) serves as the main module for factor extraction. Furthermore, to consider user cost-effectiveness, we also offer Llama3-8b[2], which is fine-tuned using output from ChatGPT to achieve the same function. This process results in the structured output of all sound events described in the caption, along with their sources and attributes. The details of the prompt instructions are provided in Appendix A.

Subsequently, a relation reasoner module is employed. It firstly locates sound events and extracts intermediate temporal relation $R_{i,i+1}$ between

[2] https://github.com/meta-llama/llama3

adjacent events, and then extends the inference to deduce the relation $R_{i,k}$ between arbitrary events. To elaborate, we summarize relations into three types: $X \xrightarrow{before} Y$, $X \xleftrightarrow{and} Y$, and $X \xrightarrow{after} Y$ as depicted in the Table 1. To deduce the relation $R_{i,k}$ between the i-th event and the k-th event, the recursive formula follows the chain reasoning rule outlined in the last row of Table 1 as:

$$R_{i,k} = G\left(R_{k-1,k}, R_{i,k-1}\right), \tag{1}$$

$R_{i,k}$ = <before> represents the i-th event occurring before the k-th event, and

$$G(x, \hat{x}) = \begin{cases} N_a(x, \hat{x}) & \text{<and>} \in [\hat{x}, x], \hat{x} \neq x, \\ x & \hat{x} = x, \\ \text{<unknown>} & \text{others.} \end{cases} \tag{2}$$

Here, $\hat{x}$ represents a previously inferred relation, and x represents the current relation between adjacent events, $N_a(x, \hat{x})$ outputs the one between x and $\hat{x}$ which is not equal to <and>.

Greedy Graph Matching Existing binary matching method (Anderson et al., 2016) compares predicted and reference graphs as sets of tuples, scoring each tuple as 1/0, which may overlook potential candidates. To address this, we propose a matching method inspired from the greedy search algorithm, finding the best match candidate node for each node. We formulate matching probability $P(x)$ of each node, which involves computing the maximum similarity between an anchor node x and the candidate set $Y = \{y_i\}_{i=0}^{N}$ at each level of graph, as illustrated in the following equation:

$$P(x) = \max_{y \in Y} S(x, y), \quad 0 \leq P(x) \leq 1. \tag{3}$$

Here, $S(x, y)$ represents the similarity between x and y. When calculating precision, the prediction node is considered as x, with the reference node as y. When calculating recall, the roles are reversed.

Contrary to previous methods (Gontier et al., 2023; Anderson et al., 2016) that treated nodes equally, here sources, attributes, and relations are dependent on their respective sound events, and their significance as child nodes relies on their parent node. Thus, we define the matching probability of a child node based on the matching probability of its parent node, as follows:

$$P(C_{i,j}) = P(C_{i,j}, E_i) = P(C_{i,j}|E_i)P(E_i), \tag{4}$$

where E_i represents the i-th event as a parent node, and $C_{i,j}$ (source or attribute) represents its j-th child node. Notably, $C_{i,j}$ is contained within E_i, leading to $P(C_{i,j}) = P(C_{i,j}, E_i)$.

As for the temporal relation between sound events, the formula is as follows:

$$\begin{aligned} P\left(R_{i,k}\right) &= P\left(R_{i,k}, E_i, E_k\right) \\ &= P\left(R_{i,k} \mid E_i, E_k\right) P\left(E_i\right) P\left(E_k\right), \end{aligned} \tag{5}$$

where $R_{i,k}$ denotes the temporal relation between the i-th event and the k-th event, and $P(E_i)$ and $P(E_k)$ are independent from each other.

Subsequently, the average matching probabilities S_{fac} for each factor are summarized, with $\boldsymbol{E}$, $\boldsymbol{A}$, $\boldsymbol{S}$ and $\boldsymbol{R}$ denoting the factor sets of sound events, attributes, sources, and relations, respectively:

$$S_{fac} = Avg\left[P(fac)\right], fac \in [\boldsymbol{E}, \boldsymbol{A}, \boldsymbol{S}, \boldsymbol{R}]. \tag{6}$$

If the input variable x in $P(x)$ represents a hypothesis/reference, approximate precision $\hat{P}_{fac}$ /recall $\hat{R}_{fac}$ can replace score S_{fac} above. This yields the factor-level F-value F_{fac}, which serves as the score for the E/S/A/R-ACE metrics.

Lastly, the overall graph score S_G is calculated leveraging the macro-Precision/Recall as follows:

$$P_{\text{macro}} = \frac{1}{4} \sum_{fac} \hat{P}_{fac}, \tag{7}$$

$$R_{\text{macro}} = \frac{1}{4} \sum_{fac} \hat{R}_{fac}, \tag{8}$$

$$S_G = F_{\text{macro}} = \frac{2 \cdot P_{\text{macro}} \cdot R_{\text{macro}}}{P_{\text{macro}} + R_{\text{macro}}}. \tag{9}$$

Cross-modal Similarity This module is employed to identify predictions that may correspond to a part of the audio but are not explicitly mentioned in the text. The text global vector $\boldsymbol{V}_t \in \mathbb{R}^{Dim}$ is extracted from predicted caption using BERT (Devlin et al., 2019), and the audio global vector $\boldsymbol{V}_a \in \mathbb{R}^{Dim}$ is extracted from reference audio using the HT-SAT encoder (Chen et al., 2022). Subsequently, cross-modal similarity is derived from the cosine similarity as follows,

$$S_C = \boldsymbol{V}_a^{\top} \boldsymbol{V}_t. \tag{10}$$

The holistic score of X-ACE is defined as:

$$S_{X-ACE} = (S_G + S_C)/2. \tag{11}$$

	HC	HI	HM	MM	Total
Random judgement	45.8	45.7	51.1	52.0	50.0
BLEU_4 (Papineni et al., 2002)	55.2	85.8	77.3	50.7	61.5
SPICE (Anderson et al., 2016)	50.7	83.4	76.5	49.3	59.6
CIDEr (Vedantam et al., 2015)	56.7	96.0	89.1	61.0	70.8
METEOR (Banerjee and Lavie, 2005)	66.0	96.4	89.1	60.2	71.7
ROUGE_L (Lin, 2004)	61.1	91.5	81.1	52.4	64.8
BERTScore (Zhang et al., 2020)	61.1	97.2	91.2	65.4	74.3
Sentence-BERT (Reimers and Gurevych, 2019)	64.0	**99.6**	92.0	73.7	79.7
ChatGPT*					
X-ACE	65.7	**99.6**	93.7	**76.8**	**81.8**
X-ACE w/o. cm	63.7	93.5	90.0	73.1	78.0
X-ACE w/o. anno&cm	66.7	90.2	84.8	74.2	77.6
Llama3-8B*					
X-ACE	64.2	**99.6**	**95.4**	75.9	81.4
X-ACE w/o. cm	64.7	94.3	91.6	72.6	78.2
X-ACE w/o. anno&cm	**69.7**	93.1	86.1	72.7	77.8

Table 2: Correlation with human judgement on the AudioCaps dataset. The "w/o. cm" represents X-ACE without the cross-modal similarity stream, while "anno" denotes our human annotation from *AudioCaps-F*. "ChatGPT*" and "Llama3-8B*" refer to the performance of different LLMs as factor extractors.

Factor	Caption Perturbation
Sound	Rain and thunder occurs. Rain and thunder occurs with blowing wind.
Source	A woman speaks with a boy cries. A boy speaks with a woman cries.
	A bird is chirping. A buzz is chirping.
Attribute	Several people are talking aloud. Several people are talking lightly.
Relation	An motor is operating with rhythmic whirring. An motor is operating followed by rhythmic whirring.

Table 3: The samples of perturbation on gold captions.

4 Evaluation on X-ACE

4.1 Correlation with Human Judgement

In our experiment on the AudioCaps dataset, a pairwise comparison was employed to measure the correlation between metrics and human judgments. For each pair of candidate captions, humans label which caption in the pair is closer to the given audio. The evaluation is categorized into four splits, as utilized in (Zhou et al., 2022): "HC": two human-written captions matching the audio, "HI": two human-written captions with only one matching the audio, "HM": a matching human-written caption, and a machine-generated caption, and "MM": two machine-generated captions. As depicted in Table 2, X-ACE emerges as the state-of-the-art in correlation with human subjective judgment. X-ACE without cross-modal similarity module and extra annotation exhibits higher correlation in HC split, with further analysis detailed in Appendix C. We also offer an open-source model Llama3-8B, trained using outputs from ChatGPT. Our X-ACE composed of Llama3-8B as the factor extraction component, still closely matches human subjective perception. Interestingly, it even outperforms system built with ChatGPT under ablation.

4.2 Mismatch Detection

To see whether our factor-level score can sensitively detect types of nuanced mismatch. we automatically introduced different perturbation to clean captions as Table 3 to synthesize subtle mismatch. We then compare the fluctuations in factor-level score with other metrics before and after the perturbations to validate their ability to distinguish error types. To measure the degree of fluctuations on metric scores, we employed the Kruskal-Wallis (McKight and Najab, 2010) significance test. The value in the Table 4 is $\hat{p} = -log(p)$, with p denotes significance level. A smaller p value, corresponding to a larger $\hat{p}$, indicates a higher level of difference. When p is less than 0.05 ($\hat{p}$ is greater than 3), it signifies a prominent difference in scores

Perturbated Factor	E-ACE	S-ACE	A-ACE	R-ACE	BLEU_4	METEOR	ROUGE_L	CIDEr	SPICE	S-BERT
Sound Event	26.73	17.16	0.78	1.1	7.88	3.42	10.32	21.72	8.15	59.41
Source	1.78	138.92	2.88	0.43	50.66	40.42	41.24	29.25	34.22	88.56
Attribute	0.43	0.94	21.21	1.48	3.87	2.24	3.09	5.99	4.32	9.64
Relation	0.02	1.2	0.07	215.32	13.7	8.29	13.84	10.65	0.63	4.42

Table 4: Sensitivity degree towards perturbation on factors. E/S/A/R-ACE respectively represent individual factor-level metrics for Sound Event/Source/Attribute/Relation in X-ACE, while S-BERT denotes Sentence-BERT.

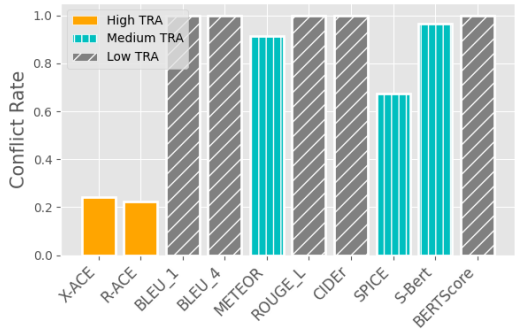


Figure 4: The conflict rate between the metric scores and the prior facts, the lower the value, the stronger the temporal reasoning ability (TRA).

before and after perturbing the captions. As showed in the Table 4, our factor-level score mostly get the greatest sensitivity towards the corresponding perturbation on captions, while maintaining other scores essentially the same. In particular, along with the change occur on the sound, the score of source score also changes. The results demonstrate that the factor score of X-ACE exhibits strong capabilities in detecting subtle mismatches and distinguishing different types of errors.

4.3 Temporal Relation Reasoning

To evaluate the temporal relation reasoning capability of metrics shown in Figure 4, individual sound events were concatenated to form three types of captions: $X \xrightarrow{before} Y$ (reference), $X \xrightarrow{after} Y$ (candidate 1), and $Y \xrightarrow{after} X$ (candidate 2).

Based on prior knowledge, candidate 2 merits a higher score for its equivalence to the reference through reasoning. We observed the conflict rate between the metric scores and the prior facts, the results shows that X-ACE and R-ACE greatly outperform others, with their conflict rates lower than 30%. As the temporal relation reasoning abilities of the other metrics even inferior to random selection (50% conflict rate).

4.4 Ablation Studies

To eliminate interference of cross-modal similarity and annotation, X-ACE w/o. cm&anno serves as our baseline X-ACE$_{base}$ in ablation study.

Setting	HC	HI	HM	MM	Total
X-ACE$_{base}$	**66.7**	90.2	**84.8**	**74.2**	**77.6**
w/o. Sound Event	65.2	75.1	72.6	71.2	71.2
w/o. Source	65.7	87.8	**84.8**	72.5	76.1
w/o. Attribute	63.7	90.2	81.9	67.5	73.1
w/o. Relation	64.2	**91.4**	**84.8**	70.9	75.7

Table 5: Correlation with human judgement after ablating different factors.

Method	HC	HI	HM	MM	Total
Greedy Matching	**66.7**	**90.2**	**84.8**	**74.2**	**77.6**
Binary Matching	60.7	87.3	79.7	49.5	62.2

Table 6: Correlation with human judgement using different matching methods

The impact of each factors The ablation analysis was conducted to investigate the impact of removing individual factor-level scores. The results presented in Table 5 indicate that removing any single factor leads to a significant decrease in correlation, suggesting that all factors are indispensable for correlation with human perception. The removal of sound event leads to a notably decreased performance, which underscores the dominance of sound events within audio descriptions.

Greedy matching vs. binary matching We conducted a comparative experiment to evaluate the greedy graph matching method designed for X-ACE against the conventional binary matching (Anderson et al., 2016). As evident from the Table 6, our approach results in an overall performance improvement of 15.2%, demonstrating substantial enhancements across all splits, especially with 24.7% improvement in the MM split.

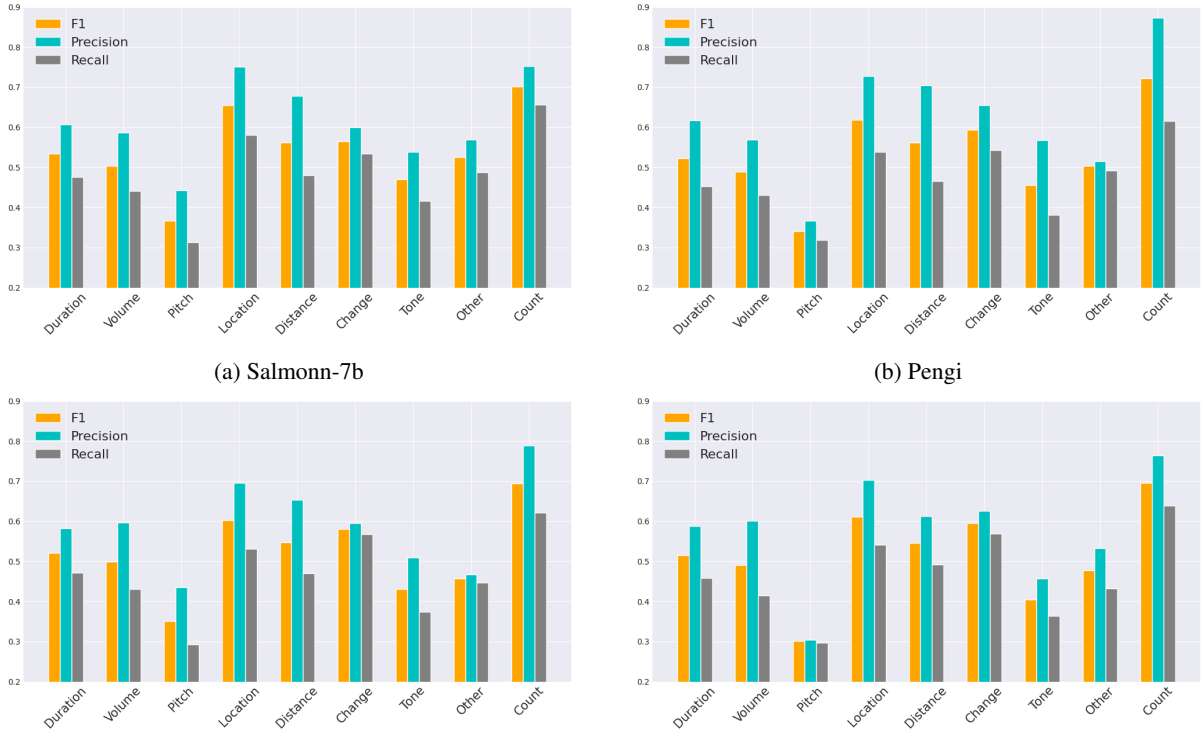
5 Evaluation of ALMs Using X-ACE

5.1 Empirical Evaluation of Current ALMs

In this section, we employed X-ACE for the first time to systematically evaluate current ALMs in audio captioning. We specifically analyzed the areas in which different ALMs excel or fall short in. The ALMs models selected for evaluation include

Models	Sound Event		Source		Attribute		Relation		Total
	P	R	P	R	P	R	P	R	F
Salmonn-13b	66.86	64.52	48.4	45.15	32.71	21.18	24.3	22.99	39.77
Salmonn-7b	70.66	63.88	53.47	47.85	26.82	14.39	24.23	22.83	39.39
Qwen-Audio	70.23	62.3	50.82	44.68	16.66	8.81	21.52	19.49	35.84
Pengi	68.23	55.3	45.11	38.63	19.5	10.42	17.09	15.49	32.53
AudioGPT	63.35	55.47	40.06	36.23	17.11	9.69	17.96	16.76	31.11

Table 7: The X-ACE scores of different ALMs in AAC based on the *AudioCaps-F* dataset.



(a) Salmonn-7b

(b) Pengi

(c) Qwen-Audio

(d) AudioGPT

Figure 5: The fine-grained A-ACE scores of different ALMs in AAC based on the *AudioCaps-F* dataset.

Qwen-Audio (Chu et al., 2023), SALMONN (Tang et al., 2023), Pengi (Deshmukh et al., 2023), and AudioGPT (Huang et al., 2023b). Extensive experiments to assess ALMs with other metrics and investigate the influence of certain variables on metric assessment are shown in Appendix D.

The performance of ALMs evaluated by X-ACE across different factors is depicted in Table 7. Salmonn emerges as the top-performing model, followed by Qwen-Audio, showcasing the best overall performance. The average performance of ALMs on the sound event factor is the highest, while scores for other factors dropped significantly. Notably, even if a model excels in a specific area (e.g., Salmonn-7b performs the best in sound events), the total score may still decrease due to a high incidence of omissions in other factor. Because X-ACE evaluates all factors collectively, and the shortcomings of each factor affect the overall evaluation. Among models, Qwen-Audio, Pengi, and AudioGPT all exhibit subpar performance in attributes and relations. These models necessitate targeted enhancements grounded in their performance in specific factors, highlighting their inadequacy in characterizing audio attributes.

5.2 Analysis of Attribute Factor

As Table 7 clearly shows that the factor attribute constitutes a vulnerability for the popular ALMs, consequently, a finer evaluation of their performance across nine subcategories of attributes was undertaken. The examples of these subcategories and the fine-grained calculation of the A-ACE score can be referred in Appendix B. Figure 5 presents a comparison of precision and recall values, revealing that the primary issue in attributes

lies in insufficient detail rather than a high incidence of incorrect descriptions or hallucination. Pitch is mostly misdescribed, followed by tone and volume. Most ALMs generate accurate descriptions regarding the vocalization count and their location. Among them, Pengi achieves a high precision rate of nearly 90% in count, but it lacks description or contains errors in pitch. There is a noticeable inconsistency in the distribution of strengths and weaknesses among different models, necessitating targeted improvements.

6 Conclusion

The current assessment of ALMs only provide an overall score, presenting challenges for model refinement. In response, this paper introduces explainable and multi-factor evaluation paradigm X-ACE for AAC, defining sound event, source, attribute, and relation as four factors tailored for the audio description. Furthermore, we provided a dataset *AudioCaps-F* to enhance evaluation granularity. X-ACE exhibits remarkable alignment with human perception and shows a nuanced capacity to distinguish model errors. Our analysis, derived from outcomes of X-ACE, illuminates substantial variances among mainstream models in audio attribute, and temporal sequence description. While differences are less pronounced in sound events, considerable room for improvement exists across factors.

Limitation

Despite achieving satisfactory performance, our X-ACE metric also has limitations. Due to the fact that X-ACE requires entity reasoning from LLM during the factor extraction stage, it incurs greater time overhead compared to other metrics that do not rely on large models. We are also dedicated to speeding up inference and training smaller models to replace LLM, thereby achieving more efficient automatic evaluation. Additionally, the cross-modal metrics in X-ACE rely on the robustness of the corresponding cross-modal models, and due to cost constraints, annotations are limited to the AudioCaps dataset. We hope to expand this to other formats of datasets in the future.

References

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*, volume 9909 of *Lecture Notes in Computer Science*, 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 Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization@ACL 2005, Ann Arbor, Michigan, USA, June 29, 2005*, pages 65–72. Association for Computational Linguistics.

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 *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022*, pages 646–650. IEEE.

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. 2023. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *CoRR*, abs/2311.07919.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. 2023. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. *CoRR*, abs/2305.06500.

Soham Deshmukh, Benjamin Elizalde, Rita Singh, and Huaming Wang. 2023. Pengi: An audio language model for audio tasks. *CoRR*, abs/2305.11834.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Patrick Fernandes, Daniel Deutsch, Mara Finkelstein, Parker Riley, André Martins, Graham Neubig, Ankush Garg, Jonathan H. Clark, Markus Freitag, and Orhan Firat. 2023. The devil is in the errors: Leveraging large language models for fine-grained machine translation evaluation. In *Proceedings of the Eighth Conference on Machine Translation, WMT 2023, Singapore, December 6-7, 2023*, pages 1066–1083. Association for Computational Linguistics.

Félix Gontier, Romain Serizel, and Christophe Cerisara. 2023. Spice+: Evaluation of automatic audio captioning systems with pre-trained language models. In *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2023, Rhodes Island, Greece, June 4-10, 2023*, pages 1–5. IEEE.

Anwen Hu, Shizhe Chen, Liang Zhang, and Qin Jin. 2023. Infometic: An informative metric for reference-free image caption evaluation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 3171–3185. Association for Computational Linguistics.

Jiawei Huang, Yi Ren, Rongjie Huang, Dongchao Yang, Zhenhui Ye, Chen Zhang, Jinglin Liu, Xiang Yin, Zejun Ma, and Zhou Zhao. 2023a. Make-an-audio 2: Temporal-enhanced text-to-audio generation. *CoRR*, abs/2305.18474.

Rongjie Huang, Mingze Li, Dongchao Yang, Jiatong Shi, Xuankai Chang, Zhenhui Ye, Yuning Wu, Zhiqing Hong, Jiawei Huang, Jinglin Liu, Yi Ren, Zhou Zhao, and Shinji Watanabe. 2023b. Audiogpt: Understanding and generating speech, music, sound, and talking head. *CoRR*, abs/2304.12995.

Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2015. Image retrieval using scene graphs. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 3668–3678. IEEE Computer Society.

Karen Spärck Jones. 2004. A statistical interpretation of term specificity and its application in retrieval. *J. Documentation*, 60(5):493–502.

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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 119–132. Association for Computational Linguistics.

Max Ku, Dongfu Jiang, Cong Wei, Xiang Yue, and Wenhu Chen. 2023. Viescore: Towards explainable metrics for conditional image synthesis evaluation. *CoRR*, abs/2312.14867.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 292–305. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. Visual instruction tuning. *CoRR*, abs/2304.08485.

Hui Liu and Xiaojun Wan. 2023. Models see hallucinations: Evaluating the factuality in video captioning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 11807–11823. Association for Computational Linguistics.

Minqian Liu, Ying Shen, Zhiyang Xu, Yixin Cao, Eunah Cho, Vaibhav Kumar, Reza Ghanadan, and Lifu Huang. 2023b. X-eval: Generalizable multi-aspect text evaluation via augmented instruction tuning with auxiliary evaluation aspects. *CoRR*, abs/2311.08788.

Qingyu Lu, Baopu Qiu, Liang Ding, Liping Xie, and Dacheng Tao. 2023. Error analysis prompting enables human-like translation evaluation in large language models: A case study on chatgpt. *CoRR*, abs/2303.13809.

Patrick E McKight and Julius Najab. 2010. Kruskal-wallis test. *The corsini encyclopedia of psychology*, pages 1–1.

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. *CoRR*, abs/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, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1532–1543. ACL.

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 2019, Hong Kong, China, November 3-7, 2019*, pages 3980–3990. Association for Computational Linguistics.

Sebastian Schuster, Ranjay Krishna, Angel X. Chang, Li Fei-Fei, and Christopher D. Manning. 2015. Generating semantically precise scene graphs from textual descriptions for improved image retrieval. In *Proceedings of the Fourth Workshop on Vision and Language, VL@EMNLP 2015, Lisbon, Portugal, September 18, 2015*, pages 70–80. Association for Computational Linguistics.

Daiki Takeuchi, Yasunori Ohishi, Daisuke Niizumi, Noboru Harada, and Kunio Kashino. 2023. Audio difference captioning utilizing similarity-discrepancy disentanglement. *CoRR*, abs/2308.11923.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. 2023. SALMONN: towards generic hearing abilities for large language models. *CoRR*, abs/2310.13289.

Nasib Ullah and Partha Pratim Mohanta. 2022. Thinking hallucination for video captioning. In *Computer Vision - ACCV 2022 - 16th Asian Conference on Computer Vision, Macao, China, December 4-8, 2022, Proceedings, Part IV*, volume 13844 of *Lecture Notes in Computer Science*, pages 623–640. Springer.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 4566–4575. IEEE Computer Society.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Ming Yan, Ji Zhang, and Jitao Sang. 2023. An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *CoRR*, abs/2311.07397.

Ho-Hsiang Wu, Oriol Nieto, Juan Pablo Bello, and Justin Salamon. 2023. Audio-text models do not yet leverage natural language. In *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2023, Rhodes Island, Greece, June 4-10, 2023*, pages 1–5. IEEE.

Wenda Xu, Danqing Wang, Liangming Pan, Zhenqiao Song, Markus Freitag, William Wang, and Lei Li. 2023. INSTRUCTSCORE: towards explainable text generation evaluation with automatic feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 5967–5994. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. 2023. Analyzing and mitigating object hallucination in large vision-language models. *CoRR*, abs/2310.00754.

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 *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022*, pages 981–985. IEEE.

A Factor Construction

Annotation Construction This section provides a detailed demonstration of the annotation process. We engaged specialized annotators to manually annotate structured data for the given audio, including all sound events occurring in the audio along with their respective sources and attributes, as illustrated in Figure 8. We further categorize attributes into nine major classes. Annotators can use the attribute labels provided in Figure 9 as a reference for annotation. On average, each sound event is annotated with at least four attributes.

Factor extraction via LLM The prompts are input into an LLM extractor, as shown in Table 10, to generate structured data in JSON format. The format of data is then checked and refined via an automated program.

B Model Details

Encoder The text encoder and audio encoder are trained on the audio-text retrieval task, using the parameters provided in (Mei et al., 2023). For the matching of each pair of entities in graph, GloVe (Pennington et al., 2014) is used to transform phrases into word vectors for similarity calculation.

Attribute subcategory score We conduct greedy graph matching on anchor attributes, then semantically map them to the closest attribute labels in the attribute tag library, categorizing them into corresponding subcategories, thereby obtaining scores in that subcategory.

C Experiment

Ablation Study and Analysis We analyzed cross-modal similarity (cm) and refined annotation (anno) through ablation study as Table 11. Poorer performance in HC and MM splits is evident for cm only, for it lacks of fine-grained interaction and only distinguishes prominent differences (e.g. in HI and HM). Particularly notable in the HC split is that "w/o. anno" outperforms X-ACE. As for the two candidates in HC are reference captions of a given audio in the test data, and the annotation based on references contains all features extracted from these two sentences. Consequently, a self-reference issue arises in the evaluation process, leading to a slight decrease in fairness and correlation with human judgement. This issue does not impede our use of X-ACE to evaluate sentences generated by ALMs.

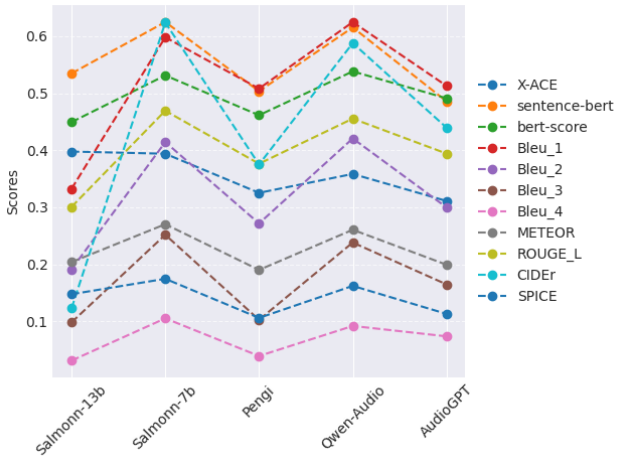


Figure 6: Evaluation results of Metrics for ALMs.

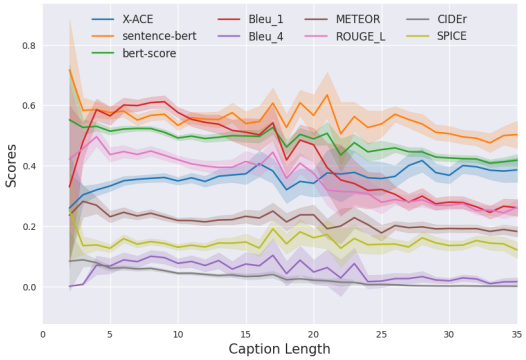


Figure 7: Relationship between different metric scores and caption length

Temporal Relation Reasoning Setting We picked sound event pairs and generated 178 samples, each formatted as shown in Table 12.

D Evaluate ALMs on Metrics

Overall Evaluations As shown in Figure 6, Salmonn and Qwen-Audio exhibit superior performance in AAC, while scores for Pengi and AudioGPT show great fluctuations. X-ACE curve aligns with the trend of SPICE, for they both encompass multiple factors. However, SPICE focuses on visible entities rather than audio characteristics. Notably, X-ACE is the only one to discern superiority of Salmonn-13b over 7b. For the 13b model expresses more diversely, which X-ACE can identify and match, resulting in higher scores.

Association between assessment and caption length It can be observed in Figure 7 that for descriptions with short lengths, such as phrases of only two words, Sentence-BERT, BERTScore, CIDEr, and SPICE surprisingly provide high scores.

Audio	Sound Event	Source	Attribute
Ylq9RvAA4mqY.wav	talking clanking sizzle	man metal food, oil	Off and on, Mid-size, Alongside, Gentle Several times, Softly, Die down, Nearby, Low, In the background Continuous, Noisy, Alongside, Steady, In the background
Y7P0N61TVOxE.wav	playing clanking talk	music glass people	Continuous, Mid-size, Clear, Euphonious Suddenly, once, Clear, loud, shortly, in the background, high-pitched lightly, coarse, Interrupted, in the background
Y2ABngPM3raQ.wav	talking croaking tapping sound	man frog	Off and on, Mid-size, Deep, Alongside, Gentle, Clear Suddenly, Noisy, repeatedly, Shrill, Die down, Alongside Several times, softly, Interrupted, Clear, Fast
YJhGp7HmRQxg.wav	chirping neighing clack	birds horse metal	in the background, Off and on, Faint, In the distance Once, Mid-size, Suddenly, Clear Once, softly, Muffled, In the distance
YPWjEfOkb6ro.wav	talk running	crowd of people water	in the background, Continuous, Noisy, Low, coarse Continuous, Loud, Speed up, Clear, harder

Table 8: Examples of manual annotation for *AudioCaps-F*.

Categories	Tag
Duration	Off and on / Intermittent / Occasional / Sparse / Briefly / Rhythmic Continuous / Successive / Frequently / Sustained / Repeatedly Suddenly / Rapidly / Quickly / Fast / Shortly / Slowly
Volume	Big / Loud / Noisy / Strong / Powerful / Heavy Small / Lightly / Softly / Quiet / Faint / Slight Mid-size
Pitch	Muffled / Low / Deep Sharp / High-pitched / Shrill / Searing / Piercing Mid-pitched
Location	On a XXX / Against XX / Outdoors / Indoors / In the background / In the foreground
Count	Several times / X times / Once / Twice
Distance	In the distance / Distant / Nearby / Alongside
Variation	Slow down / Die down / Interrupted / Turn off / Over Speed up / Harder / Begins / Turn on / Goes up and down
Tone	Formal / Casual / Serious / Excited / Gentle / Angry / Skeptical / Commanding
Other	Synthesized / Digital / Electronic / Coarse / Clear / steady

Table 9: Attribute subcategories and its corresponding tag (for reference to annotators)

Template	Please help me to extract 3 type of nodes in the audio caption: sound event (sound events be described by caption), source (who generates the sound event), attribute (attribute of sound event). Example1: caption = "A man speaking as monkeys scream and dogs bark followed by birds cawing in the distance and a loud explosion". You need to act as an audio caption extractor, to extract dict of which keys are sound events, value are attribute list toward sounds and source (if no, list should be replaced by None). The return demo is {"speaking":{"source":["man"],"attr":None}, "scream":{"source":["monkeys"],"attr":None},"bark":{"source":["dog"],"attr":None}, "cawing":{"source":["birds"],"attr":["in the distance"]},"explosion":"source":None,"attr":["loud"]}. You should pay attention on the sound event (not sound source),e.g. "speaking","bark","cawing","explosion" should be copied from the original caption. I'll show you the rest of the caption below, you give me a json return in the format.You do this work in json format!!! Not show me codes.
Input	Now I will give you caption: "A machine makes stitching sounds while people are talking in the background"
GPT3.5 Output	{ "stitching": { "source": ["machine"], "attr": null }, "talking": { "source": ["people"], "attr": ["in the background"] } }

Table 10: The prompts for factor extraction.

X-ACE shows a slight increase followed by a plateau in scores as the caption length increases, with slight fluctuations. On the contrary, most other metrics tend to decrease, as they do not encourage detailed descriptions to minimize hallucinations or mismatches. BLEU_1 exhibits a rapid rise

	AudioCaps				
	HC	HI	HM	MM	Total
cm	57.2	99.2	**94.1**	67.4	75.2
X-ACE	65.7	**99.6**	93.7	**76.8**	**81.8**
w/o. cm	63.7	93.5	90.0	73.1	78.0
w/o. anno	**67.7**	**99.6**	91.6	75.0	80.8
w/o. anno&cm	66.7	90.2	84.8	74.2	77.6

Table 11: Ablation studies on removal of modules

$X \xrightarrow{before} Y$	A river stream flows before a bell ring.
$X \xrightarrow{after} Y$	A river stream flows after a bell ring.
$Y \xrightarrow{after} X$	A bell rings after a river stream flows.

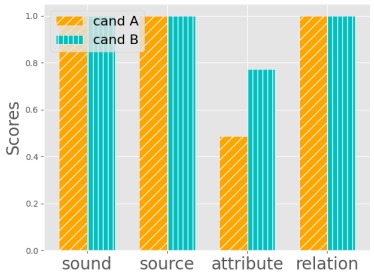
Table 12: The sample format used for temporal reasoning, with the first row as the reference and the last two rows as two candidates.

followed by a decline, the greatest fluctuation indicates that it is the most influenced by caption length. The scores of all metrics fluctuate greatly in the range of caption length from 15 to 25, where the metric assessments show little correlation with caption length.

Case Studies Case studies are conducted as shown in Fig. 8, where (a) displays sample comprising the ref (correctly described reference for the given audio), along with candidate A and B for evaluation, with blue highlighting the true mismatches. Analyzing the first sample, candidate A interchanged the attributes "distant" and "nearby" associated with two sounds as a bad caption. Candidate B only altered the sentence structure and employed synonym substitutions without changing the content as a good caption. As shown in (b), our A-ACE indicates a significant decrease in the attribute score for candidate A, while B shows a slight decline, demonstrating the sensitivity of our method in detecting attribute errors. Furthermore, in (c), comparing all metrics, the bold sections corresponding to X-ACE and A-ACE reveal results consistent with subjective evaluation, indicating that B outperforms A. It is worth noting that since the current case study utilizes only one sample set without establishing a sample library, the CIDEr score remains at 0 throughout.

Ref	Wind blows with some nearby rustling and distant passing traffic.
A	Wind blows with some distant rustling and nearby passing traffic. ✗
B	As some nearby is rustling and passing traffic is passing distantly, wind blows. ✓

(a) Caption Samples



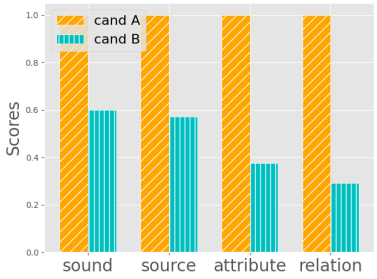
(b) X-ACE Assessment

	X-ACE	A-ACE	BLeu_1	BLeu_4	METEOR	ROUGE_L	CIDEr	SPICE	S-bert	BERTScore
A	**0.872**	**0.488**	0.999	0.375	0.478	0.8	0.0	0.667	0.991	0.925
B	**0.947**	**0.772**	0.615	6.57e-9	0.411	0.534	0.0	0.4	0.926	0.734

(c) Evaluation on Candidate A/B by different Metrics

Ref	An animal makes squeaking noises with buzzing background sounds, and a dog barks.
A	Squeaking noises are made by an animal and buzzing background sounds occur, while a dog is barking. ✓
B	An animal makes squeaking noises with crying background sounds, and a dog gasps. ✗

(a) Caption Samples



(b) X-ACE Assessment

	X-ACE	E-ACE	BLeu_1	BLeu_4	METEOR	ROUGE_L	CIDEr	SPICE	S-bert	BERTScore
A	**1.0**	**1.0**	0.588	3.06e-5	0.356	0.478	0.0	0.533	**0.931**	0.781
B	**0.460**	**0.60**	0.846	0.670	0.462	0.846	0.0	0.714	**0.842**	0.922

(c) Evaluation on Candidate A/B by different Metrics

Figure 8: Case study: evaluation of audio captions for two sample sets. Blue font in (a) indicates discrepancies between reference and candidate, (b) Demonstrates scoring of candidates A and B using X-ACE factors, (c) Displays evaluation of candidate pairs with various metrics, where bold signifies scores aligning with human perception.