Testing Audio Captioning Metrics with Controlled Semantic Perturbations

Assel Yermekova [1] **Vadim Popov** [1] **Tasnima Sadekova** [1] **Georgii Aparin** [1]

Abstract

There has been rapid progress in audio captioning models and evaluation metrics in recent years. However, existing captioning metrics are often evaluated only through aggregate benchmark scores, with limited analysis of their robustness to semantic variations, paraphrases, omissions, and hallucinated details. In this work, we introduce StressCaps, a diagnostic challenge set designed to systematically evaluate captioning metrics under controlled semantic perturbations. The benchmark contains both meaning-preserving transformations and semantically corrupted captions. Using StressCaps, we evaluate a broad set of commonly used audio captioning metrics and analyze their strengths and limitations across different perturbation categories. Our experiments show that many metrics remain overly sensitive to surface-level textual changes despite preserving semantic meaning, while semantic similarity metrics such as FENSE and SBERT demonstrate stronger robustness to paraphrasing but remain vulnerable to unsupported or hallucinated additions. These findings highlight significant limitations of current automatic caption evaluation methods and motivate the development of more semantically reliable metrics for long-form and open-ended caption generation. The dataset and StressCaps generation pipeline are available at https://github.com/Allessyer/stresscaps.git.

1. Introduction

Audio captioning is the task of generating natural language descriptions of the acoustic content of an audio recording, covering events, sources, and scenes. The task was first formulated by Drossos et al. (2017), who introduced a model for generating textual descriptions from environmental sound recordings and established audio captioning as a distinct problem at the intersection of audio understanding and natural language generation. Since then, audio captioning has been studied as a core capability of modern audio understanding systems, with applications in assistive technologies for the hearing impaired, content-based multimedia retrieval and indexing, surveillance and acoustic monitoring, robotic perception, and, more recently, audio-grounded reasoning in large audio-language models (Mei et al., 2022; Xu et al., 2024; Goel et al., 2025).

The development of datasets such as AudioCaps (Kim et al., 2019) and Clotho (Drossos et al., 2019) enabled rapid progress in supervised audio caption generation and established standardized evaluation settings for the task. Alongside model development, the community adopted automatic evaluation metrics inherited from related text generation tasks, including BLEU (Papineni et al., 2001), ROUGE (Lin, 2004), METEOR (Denkowski & Lavie, 2014), CIDEr (Vedantam et al., 2015), and SPICE (Anderson et al., 2016), while later works introduced audio-specific evaluation approaches such as CLAP-based similarity metrics (Elizalde et al., 2023) and MACE (Dixit et al., 2024).

Despite the increasing number of proposed evaluation metrics, their robustness and semantic reliability remain insufficiently understood. Most existing audio captioning works continue to report aggregate scores using a small subset of traditional metrics (Goel et al., 2025; Wang et al., 2025) with limited analysis of metric behavior under controlled semantic variations. However, different metrics capture fundamentally different properties of generated captions, including lexical overlap, semantic similarity, fluency, and audio-text alignment, which can lead to inconsistent evaluations across paraphrases, omissions, or hallucinated details.

This problem becomes increasingly important with the recent transition from short audio captions to long-form and highly detailed descriptions enabled by multimodal large language models. Recent systems such as Omni-Captioner (Ma et al., 2026) and Qwen3-Omni-Captioner (Xu et al., 2025) demonstrate the ability to generate substantially longer and more expressive captions containing richer semantic and temporal information. As caption length and complexity increase, it becomes crucial to systematically

[1]Huawei Noah's Ark Lab. Correspondence to: Assel Yermekova <allessyer@gmail.com>.

Proceedings of the 43^{rd} International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026.

analyze the strengths and limitations of existing caption evaluation metrics for both short and long-form captions, which remain largely unexplored.

To address this gap, we introduce **StressCaps**, a diagnostic challenge set designed to systematically evaluate captioning metrics under controlled semantic perturbations. StressCaps contains both meaning-preserving transformations and semantically corrupted captions. Using this benchmark, we conduct a comprehensive analysis of commonly used caption evaluation metrics and identify their major strengths and failure modes across different perturbation categories. Our contributions are summarized as follows:

- We provide a systematic analysis of widely used caption evaluation metrics under controlled semantic perturbations.
- We identify the strengths and limitations of existing metrics across meaning-preserving and semantically corrupted caption transformations.
- We release StressCaps together with a perturbation generation pipeline for future evaluation research.

2. Related Works

Audio captioning research has primarily relied on benchmark datasets containing short audio clips paired with human-written captions. AudioCaps (Kim et al., 2019) introduced over 46k audio-caption pairs derived from AudioSet and became one of the most widely adopted benchmarks for audio caption generation. Clotho (Drossos et al., 2019) further expanded the task by providing multiple diverse captions per audio clip and emphasizing richer acoustic scene descriptions. These datasets established the standard evaluation setting for short-form audio captioning and remain dominant benchmarks in current literature, with surveys (Mei et al., 2022; Xu et al., 2024) documenting their central role in driving model and metric development.

Automatic evaluation of caption generation has historically relied on metrics originally developed for machine translation and summarization, including BLEU (Papineni et al., 2001), ROUGE (Lin, 2004), and METEOR (Denkowski & Lavie, 2014). Later metrics such as CIDEr (Vedantam et al., 2015), SPICE (Anderson et al., 2016), and SPIDEr (Liu et al., 2017) were designed to better capture semantic similarity and consensus-based evaluation for captioning tasks, with extensions such as SPIDEr-max (Labbé et al., 2022) adapting the metric to settings where systems produce multiple caption candidates. More recent approaches introduced embedding-based semantic similarity metrics, including BERTScore (Zhang et al., 2020) and FENSE (Zhou et al., 2022), which improve robustness to paraphrasing and linguistic variation. Audio-grounded evaluation methods such as CLAP-based similarity (Elizalde et al., 2023) and MACE (Dixit et al., 2024) further attempt to measure audio-text alignment without relying solely on ground-truth captions. For long-form caption evaluation, LLM-as-a-judge approaches (Zheng et al., 2023) have recently become increasingly popular, with large audio-language models (Goel et al., 2025; Chu et al., 2024) using LLMs to score open-ended responses on benchmarks like AIR-Bench (Yang et al., 2024), although the reliability and biases of such judges in the audio domain remain insufficiently understood.

Recent multimodal large language models have shifted audio captioning from short sentence generation toward long-form and highly detailed descriptions. Qwen3-Omni-Captioner (Xu et al., 2025) demonstrated paragraph-length audio caption generation without explicit prompting, while Omni-Captioner (Ma et al., 2026) systematically analyzed the trade-off between descriptive detail and hallucination. These models mark a qualitative shift in what an "audio caption" denotes. Rather than a short sentence describing salient acoustic events, contemporary captions now jointly cover a wide range of audio and speech understanding tasks within a single paragraph-length output. A modern caption may simultaneously perform sound event detection, verbatim speech transcription, cross-lingual translation, speaker emotion recognition, speaker gender and age estimation, music and ambient scene description, and acoustic environment characterization. Thus, what was previously a fragmented landscape of task-specific systems such as ASR, speaker analysis, emotion recognition, music tagging, etc., now turns into a single descriptive output that jointly represents multiple audio understanding tasks, posing new challenges for evaluation. Traditional audio captioning metrics, designed for short event-level descriptions, are ill-suited to assess captions that interleave factual content with subjective acoustic description.

Several recent works have attempted to analyze caption evaluation metrics beyond aggregate benchmark scores. FENSE (Zhou et al., 2022) examined whether image-caption metrics transfer to the audio domain and showed that embedding-based metrics correlate more strongly with human judgments than n-gram overlap metrics, but its analysis is centered on overall correlation rather than category-specific failure modes. MACE (Dixit et al., 2024) introduced a reference-free, audio-grounded metric and compared it against existing metrics under hallucination-style perturbations at the level of aggregate rankings. Beyond audio, BERTScore (Zhang et al., 2020) reported robustness analyses for text generation metrics under paraphrase and adversarial transformations, providing methodological inspiration for our setup.

Several audio captioning metrics fall outside the scope of our evaluation. CB-Score, SPICE+ (Gontier et al., 2023),

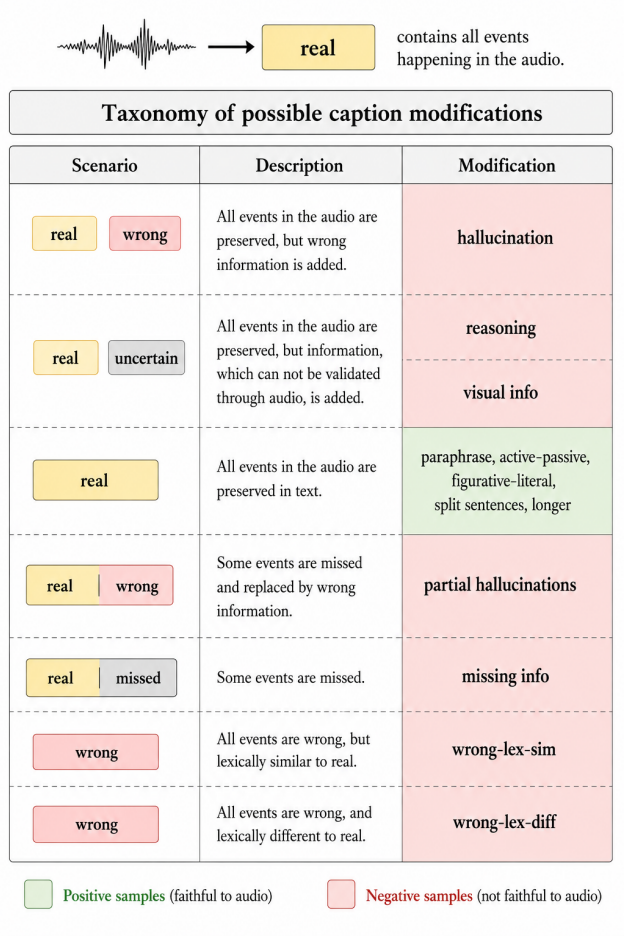

Figure 1. Taxonomy of caption modifications used in StressCaps. Modifications are organized into two top-level branches: *meaning-preserving* transformations, which alter the surface form while retaining the original semantics, and *semantically corrupted* transformations, which inject content errors such as omissions, hallucinations, or attribute substitutions.

SBF (Mahfuz et al., 2023), and s2v (Bhosale et al., 2023) do not provide publicly available implementations, preventing reproducible comparison. Two recent works are particularly close to ours in motivation and merit explicit comparison. BRACE (Guo et al., 2025) similarly evaluates audio captioning metrics, but its design produces a single aggregate score per metric over a heterogeneous set of caption pairs. While useful for choosing between metrics, this aggregation obscures the specific failure modes of each metric: it does not reveal which linguistic phenomena cause a metric to misalign with human judgment, and, consequently, offers limited guidance for diagnosing or improving existing metrics. Our work is complementary in this regard, decomposing metric behavior along a typology of controlled caption transformations. Omni-Cloze (Ma et al., 2026) takes a different evaluation approach altogether, framing caption assessment as a cloze-style multiple-choice task scored by a large language model judge. While methodologically interesting, this design introduces a dependency on proprietary or compute-intensive LLMs to produce reliable judgments. As our study targets metrics that can be deployed freely with open-source models, we exclude LLM-based evaluation from our scope, though we view it as complementary future work.

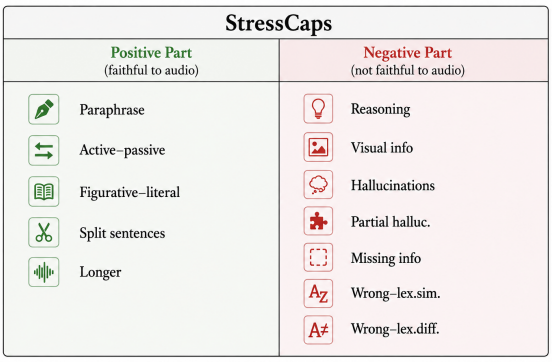

Figure 2. Overall structure of the StressCaps benchmark. Each original AudioCaps reference is paired with multiple perturbed variants spanning all modification categories, enabling per-category diagnostic analysis of caption evaluation metrics.

3. Benchmark Construction and Experimental Analysis

3.1. Taxonomy of Semantic Perturbations

The taxonomy of caption perturbations used in StressCaps is illustrated in Figure 1. The benchmark consists of two complementary subsets: a *positive* subset containing meaning-preserving transformations and a *negative* subset containing semantically corrupted captions.

Positive perturbations preserve the semantic content of the original caption while modifying its linguistic form. These transformations include active/passive voice conversion, sentence restructuring, figurative-to-literal rewriting, paraphrasing, and elaborative rewriting. In contrast, negative perturbations intentionally corrupt semantic information while maintaining fluent and realistic language. These modifications include unsupported reasoning, visual information insertion, hallucinated events, missing information, and semantically incorrect captions with either high or low lexical overlap.

This taxonomy evaluates whether captioning metrics can distinguish semantic equivalence from semantic corruption beyond surface-level lexical similarity. Positive perturbations primarily test robustness to linguistic variability, while negative perturbations evaluate sensitivity to hallucinations, omissions, and semantic inconsistencies.

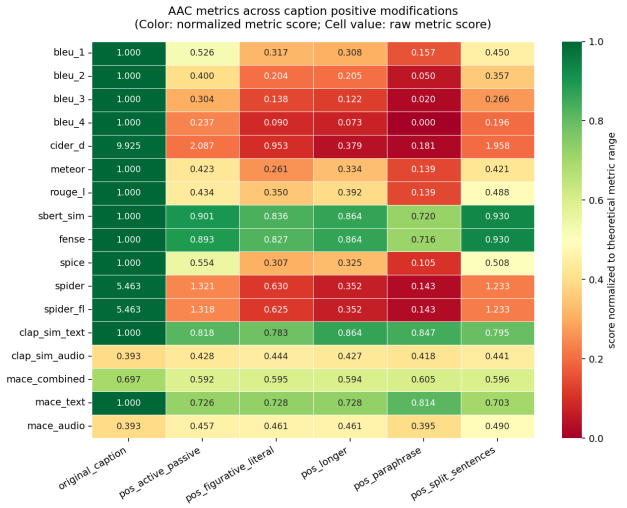

Figure 3. Behavior of caption evaluation metrics on *meaning-preserving* perturbations in StressCaps. Each row corresponds to a metric and each column to a modification category. Cell values report the average score assigned to perturbed variants relative to the original reference; values close to the reference score indicate robustness to surface-level variation, while large drops indicate over-sensitivity to lexical or syntactic changes.

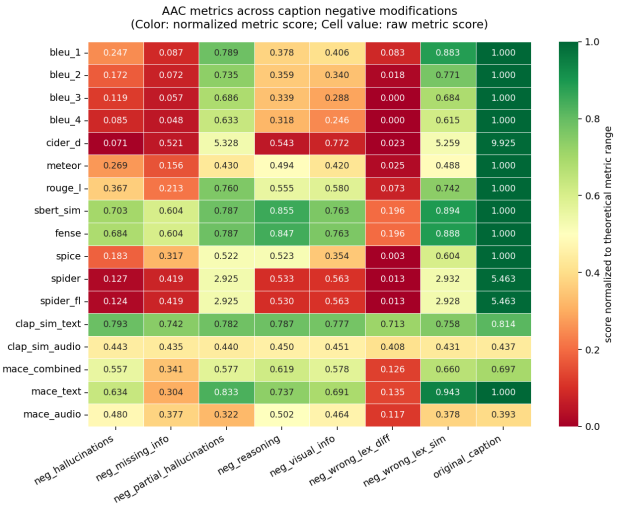

Figure 4. Behavior of caption evaluation metrics on *semantically corrupted* perturbations in StressCaps. Each row corresponds to a metric and each column to a modification category. Lower scores indicate that a metric more successfully penalizes semantic corruption, while scores close to the original reference indicate failure to detect the introduced errors.

3.2. Dataset Construction Pipeline

We constructed StressCaps using 100 audio-caption pairs from the AudioCaps test split. For each caption, we generated controlled perturbations using Qwen3-VL-235B-Instruct. The final benchmark contains 1,200 generated caption modifications (12 per original caption) spanning 2 perturbation categories: 5 meaning-preserving and 7 meaning-corrupted transformations. The generation process was iterative, with manual inspection and prompt refinement to ensure that each perturbation matched its intended category. The final prompts are released in the repository.

3.3. Metrics Evaluation on the Benchmark

Evaluation setup. We use StressCaps to evaluate a representative set of caption evaluation metrics covering the main families in current use: n-gram overlap metrics (BLEU, ROUGE-L, METEOR, CIDEr-D), scene-graph matching (SPICE, SPIDEr, SPIDEr-FL), embedding-based semantic similarity metrics (SBERT, FENSE), and audio-grounded measures (CLAP-based similarity, MACE). For every reference caption in StressCaps, each metric is computed against the original reference and against each perturbed variant produced under the taxonomy described in Section 3. To make scores comparable across metrics with different absolute ranges, we report the per-category mean score relative to the score the metric assigns to the unperturbed reference. This yields, for each (metric, category) pair, a single value indicating how strongly the metric reacts to that type of modification.

As shown in Figure 3, traditional lexical-overlap metrics such as BLEU, CIDEr, METEOR, ROUGE-L, SPICE, SPIDEr, and SPIDEr-FL substantially degrade under meaning-preserving perturbations despite the semantic content remaining unchanged. In contrast, semantic similarity metrics such as FENSE and SBERT demonstrate stronger robustness to paraphrasing, sentence restructuring, and stylistic variations. However, Figure 4 shows that these semantic metrics are considerably less sensitive to unsupported or hallucinated information, often assigning relatively high scores to semantically corrupted captions. Meanwhile, lexical-overlap metrics are more sensitive to such perturbations. These results highlight a fundamental trade-off between robustness to linguistic variability and sensitivity to semantic corruption in current caption evaluation metrics.

4. Conclusion

In this work, we introduced StressCaps, a diagnostic challenge set for evaluating captioning metrics under controlled semantic perturbations. Our analysis demonstrates that widely used caption evaluation metrics exhibit substantially different behavior across meaning-preserving and semantically corrupted transformations. While semantic similarity metrics are more robust to linguistic variability, they often fail to penalize hallucinated or unsupported information, whereas lexical-overlap metrics remain highly sensitive to surface-level wording changes. These findings highlight important limitations of current automatic evaluation methods and motivate future research toward more semantically reliable caption evaluation metrics.

5. Impact Statement

This work aims to improve the reliability of automatic evaluation for audio captioning systems. More reliable evaluation can facilitate the development of trustworthy audio-language models for accessibility, multimedia retrieval, and human-AI interaction. As StressCaps is an evaluation benchmark rather than a deployed model, we do not anticipate significant direct societal risks beyond those associated with existing audio captioning datasets.

References

Anderson, P., Fernando, B., Johnson, M., and Gould, S. Spice: Semantic propositional image caption evaluation. In Leibe, B., Matas, J., Sebe, N., and Welling, M. (eds.), *Computer Vision – ECCV 2016*, pp. 382–398, Cham, 2016. Springer International Publishing. ISBN 978-3-319-46454-1. URL https://panderson.me/images/SPICE.pdf.

Bhosale, S., Chakraborty, R., and Kopparapu, S. K. A novel metric for evaluating audio caption similarity. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2023. doi: 10.1109/ICASSP49357.2023.10096526.

Chu, Y., Xu, J., Yang, Q., Wei, H., Wei, X., Guo, Z., Leng, Y., Lv, Y., He, J., Lin, J., Zhou, C., and Zhou, J. Qwen2-audio technical report, 2024. URL https://arxiv.org/abs/2407.10759.

Denkowski, M. and Lavie, A. Meteor Universal: Language Specific Translation Evaluation for Any Target Language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pp. 376–380, Baltimore, Maryland, USA, 2014. Association for Computational Linguistics. doi: 10.3115/v1/W14-3348. URL http://aclweb.org/anthology/W14-3348.

Dixit, S., Deshmukh, S., and Raj, B. Mace: Leveraging audio for evaluating audio captioning systems, 2024. URL https://arxiv.org/abs/2411.00321.

Drossos, K., Adavanne, S., and Virtanen, T. Automated audio captioning with recurrent neural networks. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 374–378. IEEE, 2017.

Drossos, K., Lipping, S., and Virtanen, T. Clotho: An audio captioning dataset. *CoRR*, abs/1910.09387, 2019. URL http://arxiv.org/abs/1910.09387.

Elizalde, B., Deshmukh, S., and Wang, H. Natural language supervision for general-purpose audio representations, 2023. URL https://arxiv.org/abs/2309.05767.

Goel, A., Ghosh, S., Kim, J., Kumar, S., Kong, Z., gil Lee, S., Yang, C.-H. H., Duraiswami, R., Manocha, D., Valle, R., and Catanzaro, B. Audio flamingo 3: Advancing audio intelligence with fully open large audio language models, 2025. URL https://arxiv.org/abs/2507.08128.

Gontier, F., Serizel, R., and Cerisara, C. 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)*, pp. 1–5, 2023. doi: 10.1109/ICASSP49357.2023.10097021.

Guo, T., Chen, H., Liang, H., Qiang, M., Zeng, B., Sun, L., Cui, B., and Zhang, W. Brace: A benchmark for robust audio caption quality evaluation, 2025. URL https://arxiv.org/abs/2512.10403.

Kim, C. D., Kim, B., Lee, H., and Kim, G. Audiocaps: Generating captions for audios in the wild. In *NAACL-HLT*, 2019.

Labbé, E., Pellegrini, T., and Pinquier, J. Is my automatic audio captioning system so bad? spider-max: A metric to consider several caption candidates. In *Proceedings of the 7th Detection and Classification of Acoustic Scenes and Events 2022 Workshop (DCASE2022)*, Nancy, France, November 2022. URL https://dcase.community/documents/workshop2022/proceedings/DCASE2022Workshop_Labbe_46.pdf.

Lin, C.-Y. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013.

Liu, S., Zhu, Z., Ye, N., Guadarrama, S., and Murphy, K. Improved image captioning via policy gradient optimization of spider. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pp. 873–881. IEEE Computer Society, 2017. doi: 10.1109/ICCV.2017.100. URL https://doi.org/10.1109/ICCV.2017.100.

Ma, Z., Xu, R., Xing, Z., Chu, Y., Wang, Y., He, J., Xu, J., Heng, P.-A., Yu, K., Lin, J., Chng, E. S., and Chen, X. Omni-captioner: Data pipeline, models, and benchmark for omni detailed perception, 2026. URL https://arxiv.org/abs/2510.12720.

Mahfuz, R., Guo, Y., Sridhar, A. K., and Visser, E. Detecting false alarms and misses in audio captions, 2023. URL https://arxiv.org/abs/2309.03326.

Mei, X., Liu, X., Plumbley, M. D., and Wang, W. Automated audio captioning: An overview of recent progress and new challenges. *EURASIP Journal on Audio, Speech, and Music Processing*, 2022(1):26, 2022.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, pp. 311, Philadelphia, Pennsylvania, 2001. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL http://portal.acm.org/citation.cfm?doid=1073083.1073135.

Vedantam, R., Zitnick, C. L., and Parikh, D. Cider: Consensus-based image description evaluation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4566–4575, 2015. doi: 10.1109/CVPR.2015.7299087. URL https://www.cv-foundation.org/openaccess/content_cvpr_2015/papers/Vedantam_CIDEr_Consensus-Based_Image_2015_CVPR_paper.pdf.

Wang, B., Zou, X., Lin, G., Sun, S., Liu, Z., Zhang, W., Liu, Z., Aw, A., and Chen, N. F. Audiobench: A universal benchmark for audio large language models, 2025. URL https://arxiv.org/abs/2406.16020.

Xu, J., Guo, Z., Hu, H., Chu, Y., Wang, X., He, J., Wang, Y., Shi, X., He, T., Zhu, X., Lv, Y., Wang, Y., Guo, D., Wang, H., Ma, L., Zhang, P., Zhang, X., Hao, H., Guo, Z., Yang, B., Zhang, B., Ma, Z., Wei, X., Bai, S., Chen, K., Liu, X., Wang, P., Yang, M., Liu, D., Ren, X., Zheng, B., Men, R., Zhou, F., Yu, B., Yang, J., Yu, L., Zhou, J., and Lin, J. Qwen3-omni technical report, 2025. URL https://arxiv.org/abs/2509.17765.

Xu, X., Xie, Z., Wu, M., and Yu, K. Beyond the status quo: A contemporary survey of advances and challenges in audio captioning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:95–112, 2024.

Yang, Q., Xu, J., Liu, W., Chu, Y., Jiang, Z., Zhou, X., Leng, Y., Lv, Y., Zhao, Z., Zhou, C., and Zhou, J. Air-bench: Benchmarking large audio-language models via generative comprehension, 2024. URL https://arxiv.org/abs/2402.07729.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SkeHuCVFDr.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., and Stoica, I. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Zhou, Z., Zhang, Z., Xu, X., Xie, Z., Wu, M., and Zhu, K. Q. Can audio captions be evaluated with image caption metrics? In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 981–985, 2022. doi: 10.1109/ICASSP43922.2022.9746427. URL https://ieeexplore.ieee.org/abstract/document/9746427.