MMAE: A Massive Multitask Audio Editing Benchmark

Anonymous authors
Paper under double-blind review

Abstract

We introduce **MMAE**, a **M**assive **M**ultitask **A**udio **E**diting benchmark, serving as the first comprehensive evaluation testbed designed for general-purpose instruction-based audio editing. Spurred by the shift toward intelligent creation, interactive editing has rapidly expanded from visual domains, pioneered by models like Nano-banana 2 for images and Gemini-Omni for video, into audio. However, the current evaluation infrastructure lags severely, remaining highly fragmented and restricted to specific subdomains or basic operations. Unlike existing benchmarks that are limited in scope, MMAE extends to a broad spectrum of real-world scenarios, encompassing 7 distinct audio modalities, including sound, speech, music, and their mixtures. Furthermore, we establish a comprehensive taxonomy spanning 6 levels of task complexity, from basic modifications to multi-hop reasoning and multi-round editing, 2 levels of granularity, and 8 distinct operation types. Meticulously curated through human-agent collaboration, MMAE comprises 2,000 high-fidelity samples paired with a pioneering rubric-based evaluation framework. By decomposing free-form tasks into 17,741 verifiable criteria, this robust rubric-based paradigm enables a precise, multi-dimensional assessment of both instruction following and context consistency. Our extensive evaluation of leading models reveals that current systems remain far from achieving reliable edits. Strikingly, the Exact Match Rate (EMR) consistently falls below 5% and plummets to an absolute 0% in complex, mixed-modality tasks, exposing critical bottlenecks in precise execution and structural robustness. We hope MMAE will serve as a catalyst for future advances in the intelligent creation community, providing a clear diagnostic roadmap and establishing a standardized, long-lasting evaluation paradigm for next-generation audio editing systems.

1 Introduction

Intelligent editing has witnessed remarkable breakthroughs in recent years, transitioning generative models from single-element editing to interactive manipulation. In the visual domain, image editing applications like Nano-banana 2 (Google DeepMind, 2026b) have successfully matured into practical production workflows, while recent advancements in video editing models like Gemini-Omni (Google DeepMind, 2026a) have vastly expanded the creative boundaries. Spurred by this shift toward interactive intelligence, the audio community has recently seen a surge of instruction-based audio editing models (Lan et al., 2025; Tao et al., 2025; Yan et al., 2025a;b; Chen et al., 2026b; Tian et al., 2026). By allowing users to alter speech, music, or sound effects via open-ended natural language instructions, these models represent the next-generation paradigm for intelligent audio generation and editing systems.

However, the current evaluation infrastructure for audio editing lags severely behind. To effectively assess these rapidly evolving systems, a next-generation evaluation framework must achieve breakthroughs in two critical dimensions: 1) data coverage and 2) evaluation paradigm. First, existing benchmarks (Peng et al., 2024; Yan et al., 2025a;b; Chen et al., 2026b) are highly fragmented, typically restricted to specific subdomains (e.g., speech-only or sound-only) or basic operations (e.g., addition, removal, replacement). A modern testbed must comprehensively cover speech, music, and sound effects, while featuring complex scenarios designed to stress-test a model's entire cognitive pipeline from nuanced perception and implicit reasoning to high-fidelity generation.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

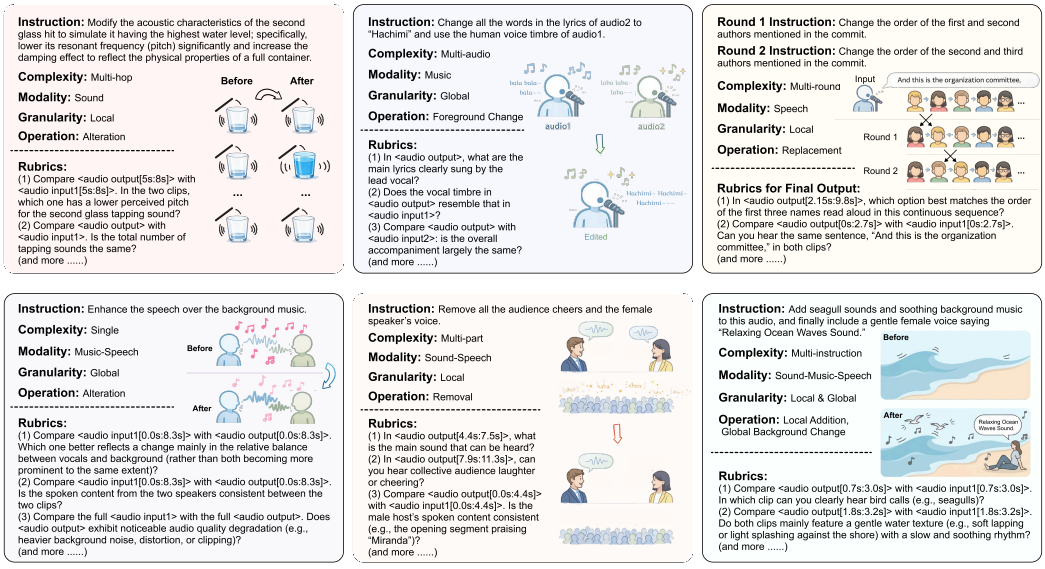


Figure 1: Examples from the MMAE benchmark, illustrating the overall taxonomy and the proposed rubric-based evaluation paradigm. These examples span diverse modalities, complexity, and operation types, with a subset of associated rubrics shown for clarity.

Second, traditional metrics have proven inadequate for open-ended instructional editing. To establish a robust and reliable evaluation paradigm, rubric-based frameworks have recently been validated in text reinforcement learning reward (Gunjal et al., 2025), audio reasoning (Ma et al., 2026a), and image editing (Bai et al., 2026), which successfully decompose free-form, multifaceted tasks into structured and verifiable criteria. Extending this proven paradigm presents a highly effective solution for evaluating complex, instruction-based audio editing.

To bridge this critical gap, we introduce **MMAE**, a **M**assive **M**ultitask **A**udio **E**diting benchmark, serving as the first comprehensive evaluation benchmark designed for general-purpose instruction-based audio editing. As illustrated in Figure 1, MMAE encompasses challenging, creative, and practical editing tasks, pairing each audio-instruction instance with a tailored set of rubrics to precisely quantify the editing outcome. Specifically, we establish a comprehensive taxonomy spanning 7 types of audio modalities (e.g., sound, speech, music, and their mixtures), 6 levels of task complexity ranging from simple modifications to complex demands (e.g., multi-hop reasoning and context-aware multi-round editing), 2 levels of granularity covering both local and global edits, and 8 distinct operation categories. Besides, the novel rubric-based paradigm enables a fine-grained, multi-dimensional assessment of editing correctness and generation quality, maximizing both reliability and interpretability to effectively diagnose model capabilities. Finally, the data curation pipeline of MMAE is rigorously designed, integrating heuristic audio collection, human-agent collaborative annotation, multi-stage refinement, and strict manual quality control. This extensive effort yields a high-fidelity benchmark comprising 2,000 diverse samples and 17,741 meticulously crafted rubrics.

Based on the MMAE benchmark, our evaluation of the 5 latest audio editing models reveals that while current systems possess basic capabilities, they remain far from achieving reliable and flawless edits. Strikingly, the Exact Match Rate (EMR) consistently falls below 5% across all models, even plummeting to an absolute 0% in complex, mixed-modality scenarios. Our detailed analysis highlights several critical limitations driving this failure: models fundamentally struggle to balance precise instruction execution with the strict preservation of unrelated acoustic contexts; they exhibit a severe lack of structural robustness as task complexity and cross-domain synchronization demands increase; and they reveal a clear decoupling between average metric competency and flawless execution. Furthermore, incorporating external agentic planners yields no consistent improvement, and we found critical bottlenecks on both the understanding and generation sides.

Our contributions are summarized as: **(1)** We introduce MMAE, the first universal benchmark for evaluating instruction-based audio editing capabilities, aiming to establish a standardized evaluation

paradigm for next-generation audio editing systems. **(2)** We construct a comprehensive, scalable evaluation suite with a systematic taxonomy, high-quality annotations, and well-defined rubric-based metrics, enabling rigorous assessment of audio editing performance.**(3)** We benchmark a wide range of leading audio editing models, exposing critical bottlenecks in current models or systems. Our empirical insights provide a clear diagnostic roadmap to guide the development of more advanced and robust next-generation systems.

2 Related Work

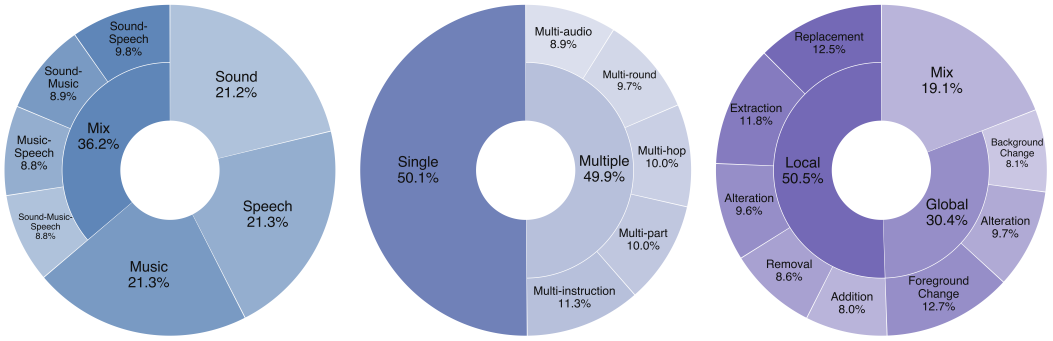
2.1 Audio Editing Models

Audio editing has been studied in relatively constrained settings in earlier works (Jiang et al., 2023; Wang et al., 2025; Ellis et al., 2025; Xu et al., 2024; Manor & Michaeli, 2024; Huang et al., 2024; Liang et al., 2024; Wang et al., 2023; Jia et al., 2025; Liang et al., 2025; Peng et al., 2024; Zheng et al., 2025), typically limited to specific modalities or predefined operation types. More recently, a growing body of research (Lan et al., 2025; Tao et al., 2025; Yan et al., 2025a;b; Chen et al., 2026b; Tian et al., 2026) has emerged along several dimensions, such as natural language instruction-guided editing, cross-modal and multi-domain generalization, and open-ended compositional editing, reflecting a broader trend toward versatile and unified audio editing systems.

Early exploration into general audio (specifically, sound effects) editing was pioneered by AUDIT (Wang et al., 2023), which introduced a latent diffusion model trained on synthetic triplets to perform text-guided addition, removal, and replacement of sound events. AudioEditor and AudioMorphix (Jia et al., 2025; Liang et al., 2025) subsequently demonstrated that comparable editing capabilities could be achieved in a training-free method. Moving beyond basic event manipulation, MMEdit (Tao et al., 2025) expanded the task scope to more diverse operations by leveraging an audio language model for joint source-instruction understanding. Alternatively, SmartDJ (Lan et al., 2025) introduced a hierarchical two-stage framework, where an audio language model decomposes high-level declarative instructions into atomic steps, which are then executed sequentially by a diffusion-based editor. In the speech domain, VoiceCraft (Peng et al., 2024) proposed a neural codec language model with a novel token rearrangement procedure for zero-shot speech content editing. CosyEdit (Chen et al., 2026a) demonstrated that speech editing can be unlocked from pretrained zero-shot TTS models via lightweight post-training with only a little amount of data. Recent advances further push the boundaries of expressiveness and integration: Step-Audio-EditX (Yan et al., 2025b) introduces nuanced control over emotion and paralinguistic attributes, while Ming-UniAudio (Yan et al., 2025a) unifies speech understanding, generation, and free-form editing within a single framework via a continuous VAE tokenizer. Several recent systems pursue broader unification across diverse audio domains. Audio-Omni (Tian et al., 2026) combines a frozen multimodal LLM with a diffusion transformer to handle generation and editing across sound, music, and speech simultaneously. AudioChat (Chen et al., 2026b) achieves unified understanding, generation, and editing through a novel transfusion forcing objective. Furthermore, InstructAV2AV (Zheng et al., 2026) and SpongeBob (Liang et al., 2026) extend the scope to joint audio-visual editing with synchronized cross-modal generation.

2.2 Audio Editing Evaluations

Despite the rapid emergence of audio editing models, a comprehensive and dedicated evaluation benchmark for this field remains absent. Current evaluation efforts are highly fragmented and strictly domain-specific. In the speech domain, existing test sets like RealEdit (Peng et al., 2024) focus on localized manipulation (insertion, deletion, and substitution) evaluated via word error rate (WER) and speaker similarity. While newer benchmarks such as Ming-Freeform-Audio-Edit (Yan et al., 2025a) and Step-Audio-Edit-Benchmark (Yan et al., 2025b) expand into semantic, acoustic, and expressive attributes, they remain confined to a restricted set of task archetypes that lacks broader operational diversity. For general audio, evaluation provided by MMEdit (Tao et al., 2025) and Audio-Omni (Tian et al., 2026) is also limited to a narrow set of basic operations, whereas StoryGen-Eval (Chen et al., 2026b) targets multi-source storytelling using FLAM-based (Wu et al., 2025) metrics. Crucially, these evaluation sets are each restricted to a narrow subset of editing tasks within a single domain, and typically rely on traditional signal-level metrics (e.g., FAD, LSD, CLAP similarity) or generic MOS ratings, failing to explicitly assess editing correctness.



(a) Modality distribution (b) Complexity distribution (c) Operation distribution

Figure 2: Distribution of the MMAE benchmark across three taxonomy dimensions.

To bridge this critical gap, MMAE serves as the first comprehensive benchmark designed for universal audio editing evaluation across all audio modalities, including sound, music, speech, and their mix. MMAE establishes a systematic taxonomy spanning modality, complexity, and operation dimensions. Moving beyond synthetic data, it features diverse, real-world audio samples paired with human-annotated natural language instructions. Furthermore, instead of relying on coarse, insufficiently robust metrics, MMAE employs fine-grained, rubric-based evaluation where each sample is assessed through multiple atomic, objective criteria targeting both instruction following and content consistency. This design enables precise, interpretable, and universal measurement of editing quality, providing a robust diagnostic tool for next-generation audio editing models.

3 MMAE

3.1 Overview

MMAE is designed to evaluate next-generation, instruction-based intelligent audio editing systems. Moving beyond simple execution, it demands that models seamlessly integrate three core capabilities: **perception** (to understand the source audio context), **reasoning** (to interpret complex, implicit user intent), and **generation** (to execute high-fidelity edits).

MMAE comprises *2k samples* paired with over *17k fine-grained rubrics*, covering challenging, creative, and practical editing tasks that require models to handle complex transformations across a wide range of editing scenarios. Each sample features an open-ended natural language instruction spanning diverse audio modalities, task complexity, and editing operations. The benchmark is constructed through a rigorous pipeline, where initial annotations are generated via human-agent collaboration and subsequently refined, validated, and quality-controlled through expert review. This workflow ensures both diversity and high quality, making MMAE a reliable benchmark for advancing the field.

3.2 Taxonomy

We design a parallel taxonomy to systematically characterize audio editing tasks from three orthogonal dimensions: **Modality**, **Complexity**, and **Operation**. These dimensions are defined in a compositional manner, enabling flexible combinations to cover a wide spectrum of multitask audio editing scenarios. Figure 2 presents the overall distribution of MMAE.

Modality. The modality dimension captures the types of audio involved in the editing task. Considering the diverse and often mixed-modality nature of real-world audio, we focus on seven categories: *sound*, *music*, *speech*, and their combinations, including *sound-music*, *sound-speech*, *music-speech*, and *sound-music-speech*.

Complexity. The complexity dimension characterizes the structural and cognitive difficulty of the task, which clearly reflects the cases of realistic usage scenarios. Tasks are stratified into six levels: **(i)** ***Single***: basic editing tasks involving a single operation on a single element; **(ii)** ***Multi-part***: tasks with

Table 1: Key statistics of the MMAE benchmark.

Statistics	Number
Total Samples	2,000
Total Rubrics	17,741
Avg. Operations / Sample	1.22
Avg. Audio Duration / Sample	14.46 sec
Avg. Instruction Length	14.00 words

Statistics	Number
Avg. Rubrics / Sample	8.87
Avg. IF Rubrics / Sample	3.58
Avg. Consistency Rubrics / Sample	5.29
Avg. Choices / Rubric	3.53
Avg. Rubric Question Length	25.45 words

a single instruction that involves multiple elements; **(iii)** ***Multi-instruction***: samples with commands composed of multiple independent single instructions; **(iv)** ***Multi-audio***: tasks involving multiple audio sources to accomplish; **(v)** ***Multi-round***: iterative editing across multiple turns, where later edits depend on earlier ones; **(vi)** ***Multi-hop***: tasks requiring multi-hop reasoning or intermediate inference to determine the expected output.

Operation. The operation dimension describes the types of editing actions, organized by granularity into *local* and *global* edits. **(i)** ***Local edits*** focus on specific segments or elements within the audio, including *addition*, *removal*, *replacement*, *extraction*, and *alteration* (e.g., modifying localized attributes such as timbre or pitch). **(ii)** ***Global edits*** *Global* edits operate on the entire audio or its overall characteristics, including *background change*, *foreground change*, and *global alteration*. Depending on the instruction, each sample may involve either a single operation or an arbitrary composition of multiple operations.

3.3 Evaluation Paradigm

Evaluating instruction-based audio editing requires a paradigm capable of benchmarking both editing correctness and generation quality. Existing metrics fall short when confronted with complex, open-ended multi-modal editing. To address this, we introduce a structured evaluation framework anchored in an **instance-level, rubric-based paradigm**. By decomposing multifaceted editing tasks into localized, verifiable checkpoints, this approach provides an objective and interpretable mechanism to diagnose model performance.

Evaluation Dimensions. To rigorously assess the multi-faceted nature of audio editing, we evaluate models along two core, complementary dimensions: **(i)** ***Instruction Following*** measures editing execution accuracy, specifically, whether the model precisely performs the modifications requested by the natural language instruction. **(ii)** ***Consistency*** evaluates context preservation, ensuring that all acoustic elements irrelevant to the editing command remain strictly unaltered. Together, these dimensions capture the fundamental trade-off in audio editing: executing precise modifications or global transformations while maintaining the holistic fidelity of the original audio.

Rubric-Based Metrics. To enable fine-grained and objective evaluation, we adopt a rubric-based evaluation paradigm. For each audio editing sample, we design a set of comprehensive, atomic, and mutually independent evaluation criteria (rubrics), which assess model performance from multiple perspectives. Each rubric is formulated as a multiple-choice question, where one correct option corresponds to successful editing behavior, and the other incorrect options indicate failure. An external judger, instantiated as a high-performance audio language model, is responsible for selecting the appropriate option based on the editing output. And each rubric is assigned a binary score (1 or 0) based on whether the option selected by the judger matches the correct option. Our rubric design is governed by four foundational principles: **(i)** ***Completeness***: The evaluation criteria should cover all relevant aspects of the editing task to avoid missing important factors. **(ii)** ***Atomicity***: Each rubric should focus on a single, indivisible property (e.g., duration, timbre, background sound, spoken content, etc.) that falls well within the perceptual and evaluative capabilities of the judge model. **(iii)** ***Orthogonality***: Different rubrics should be independent, such that the outcome of one criterion does not directly imply another. **(iv)** ***Objectivity***: Rubrics should be defined based on observable and verifiable properties, minimizing ambiguity and subjective judgment, thereby enabling the judger to function as a reliable measurement instrument.

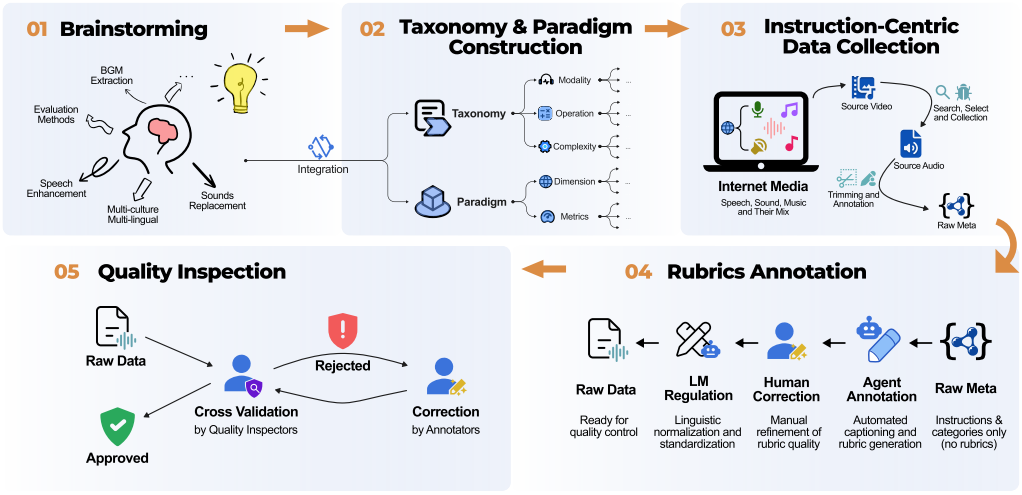


Figure 3: A comprehensive 5-stage **data curation pipeline** of our MMAE benchmark.

Statistics. As detailed in Table 1, MMAE comprises 2,000 samples and 17,741 rubrics, averaging 8.87 rubrics per sample (3.58 for Instruction Following and 5.29 for Consistency). On average, each sample spans 14.46 seconds and contains 1.22 editing operations, guided by a 14-word instruction. Furthermore, the evaluation questions average 25.45 words with 3.53 choices per rubric, ensuring a highly discriminative setup for comprehensive assessment.

3.4 Data Curation Pipeline

As illustrated in Figure 3, MMAE is constructed through a systematic five-stage pipeline designed to ensure both diversity and high-quality of the benchmark:

1) Brainstorming. We organize multiple rounds of brainstorming sessions with expert annotators to collect diverse audio editing ideas. Given the open-ended nature of audio editing, this stage focuses on gathering a wide range of intuitive, creative, and practical editing scenarios, covering different modalities and levels of complexity.

2) Taxonomy & Paradigm Construction. Based on the brainstorming outcomes, we consolidate the core design of the benchmark, including the task taxonomy and the evaluation paradigm. In particular, we establish the rubric-based evaluation framework and the orthogonal taxonomy spanning modality, complexity, and operation. This stage provides a systematic foundation for subsequent data construction.

3) Instruction-Centric Data Collection. Annotators manually search and collect audio data from online videos, including retrieving raw audio and trimming it into input clips. For each instance, annotators write instructions and label relevant metadata, including audio modality, task complexity, operation types, and keywords. To ensure balanced coverage across different task types, we adopt a dynamic balancing strategy along the three core taxonomy dimensions, resulting in a diverse and well-distributed dataset.

4) Rubrics Annotation. We employ a human-agent collaborative workflow to efficiently construct evaluation rubrics. To mitigate description hallucinations and maximize caption precision, we leverage the Omni-Detective (Ma et al., 2026b) agentic pipeline to extract detailed audio captions from the raw clips, which has been proven effective in the Qwen-Omni series (Xu et al., 2025; Team, 2026). An LLM is then utilized to generate initial rubric drafts from detailed captions, user instructions, and metadata. Subsequently, human annotators refine these rubrics by adding, removing, or refining individual rubric items. Finally, an LLM is used for post-processing, including normalization and standardization of rubric expressions. This human-agent collaborative workflow significantly improves both annotation efficiency and effectiveness.

5) Quality Inspection. Strict quality control is enforced through a dedicated cross-review protocol. Each data item is independently reviewed by blind inspectors who have no prior exposure. Samples that fail quality checks are iteratively revised until they pass the acceptance criteria, while irreparable cases are discarded to guarantee a high-fidelity final dataset.

4 Experimental Setup

Benchmarking Candidates We evaluate five recent audio editing models on the MMAE benchmark: Step-Audio-EditX (Yan et al., 2025b), Ming-UniAudio (Yan et al., 2025a), MMEdit (Tao et al., 2025), Audio-Omni (Tian et al., 2026), and SmartDJ (Lan et al., 2025). All models are end-to-end systems, except that SmartDJ is additionally tested with an external planner (Gemini 2.0 Flash (Google DeepMind, 2024)) that decomposes complex instructions into sequential atomic edits. We denote these two settings as SmartDJ w/o planner and SmartDJ w/ planner, respectively. Due to input length constraints, MMEdit, Audio-Omni, and SmartDJ are evaluated only on samples with input duration $\leq$10 seconds (801 out of 2,000 samples). Step-Audio-EditX and Ming-UniAudio are evaluated on the full set. In addition, we include two reference baselines to contextualize model performance: **(i) *Identity*:** directly returns the input audio without any modification, representing an upper bound on consistency but a lower bound on instruction following. **(ii) *Noise***: outputs pure Gaussian noise of matching duration, representing a baseline where no meaningful content is preserved.

Evaluation Details Qwen3-Omni (Xu et al., 2025) is the external MLLM judger. For each rubric, the judger is provided with the relevant audio(s) and prompted to perform explicit perception and reasoning before selecting an answer from the given options. To ensure evaluation stability, each rubric is queried three times independently; a binary score of 1 is awarded via a majority vote (at least 2/3 alignment with the ground truth) and 0 otherwise. Option positions are randomly shuffled per query to mitigate positional bias. For each sample, rubric scores are averaged within their respective axes to yield the *Instruction Following Rate* (IFR) and *Consistency Rate* (CR), respectively. We additionally report the *Exact Match Rate* (EMR), defined as the proportion of samples where all rubrics are answered correctly, which serves as a stringent metric quantifying the ratio of perfectly executed edits.

5 Experimental Results

5.1 Main Results

The main evaluation results are presented in Table 2. Across all evaluated models, the EMR consistently remains below 5%, with several models even hitting absolute zero (0%) in complex mixed-modality settings, confirming that MMAE poses substantial challenges for current audio editing systems. In the full evaluation set, Step-Audio-EditX establishes the top baseline yet still only achieves a modest 44.86% IFR and 58.88% CR, while Ming-UniAudio drops significantly lower to 29.82% IFR and 52.71% CR. Among the models restricted to the $\leq$10s subset, Audio-Omni leads with 50.73% IFR and 56.93% CR, followed by MMEdit and the SmartDJ variants.

When breaking down performance by task complexity (Table 2a), a universal degradation is observed across all models transitioning from single to multiple categories. Cross-modal analysis (Table 2b and 2c) further highlights distinct domain biases. Step-Audio-EditX and Ming-UniAudio perform relatively better on speech, while Audio-Omni shows better results on sound and music editing tasks. For mixed-modality tasks, all models exhibit notably lower scores, with Sound-Music-Speech being the hardest category.

Overall, these results indicate that current models possess basic audio editing capabilities, yet consistently fail to achieve flawless edits. They either miss certain intended modifications (low IFR), inadvertently alter content that should be preserved (low CR), and hardly demonstrate the ability to perform perfect editing (low EMR).

5.2 Observation & Discussion

We highlight several key findings below.

Table 2: Main results on the MMAE benchmark. (a) Performance grouped by complexity, reporting scores for *single* and *multiple* categories, along with the overall score. (b)(c) Performance breakdown across different modalities, with scores reported within each category. IFR = Instruction Following Rate, CR = Consistency Rate, EMR = Exact Match Rate. The best results are presented in **bold**. "Identity" denotes directly returning the input without modification, while "Noise" denotes generating pure noise. Results under these settings are reported as baselines for reference. *MMEdit, Audio-Omni, and SmartDJ are either limited to inputs of at most 10 seconds or trained solely on data with durations $\leq$ 10 seconds. Accordingly, we evaluate these models only on samples with duration $\leq$ 10 seconds (801 samples).

(a) Results by complexity.

Model	Single			Multiple			Overall		
	IFR	CR	EMR	IFR	CR	EMR	IFR	CR	EMR
Identity	23.87	96.15	2.89	30.90	92.09	6.32	27.37	94.13	4.60
Noise	35.03	15.23	0.00	29.12	16.14	0.00	32.08	15.68	0.00
Step-Audio-EditX	**46.64**	**59.06**	3.99	**43.06**	**58.69**	**2.11**	**44.86**	**58.88**	3.05
Ming-UniAudio	31.74	53.83	**4.59**	27.90	51.57	1.81	29.82	52.71	**3.20**
MMEdit*	48.39	54.27	4.86	36.94	39.86	1.90	43.12	47.64	3.50
Audio-Omni*	**58.43**	**64.57**	**6.25**	**41.70**	**47.94**	**3.52**	**50.73**	**56.93**	**4.99**
SmartDJ* w/o planer	42.52	63.91	5.56	33.12	45.43	**3.52**	38.20	55.41	4.62
SmartDJ* w/ planer	47.54	55.09	3.47	36.06	40.38	2.71	42.26	48.33	3.12

(b) Results on single modality.

Model	Sound			Music			Speech		
	IFR	CR	EMR	IFR	CR	EMR	IFR	CR	EMR
Identity	30.40	94.44	6.37	25.59	94.42	4.93	31.24	92.59	3.99
Noise	40.78	20.25	0.00	31.45	13.89	0.00	20.90	14.72	0.00
Step-Audio-EditX	**46.51**	**51.38**	**3.07**	**42.75**	**47.84**	**1.88**	**43.52**	**77.27**	4.69
Ming-UniAudio	28.77	47.88	2.12	28.97	34.71	0.94	34.13	76.01	**7.04**
MMEdit*	53.07	54.22	4.88	41.85	47.19	3.94	30.52	35.40	0.99
Audio-Omni*	**56.58**	56.02	**7.72**	**56.97**	52.55	5.51	**43.14**	**68.29**	1.98
SmartDJ* w/o planer	48.12	**64.45**	6.50	38.50	**56.88**	**7.09**	28.03	56.22	**2.97**
SmartDJ* w/ planer	51.74	53.19	4.88	46.17	46.06	2.36	32.17	43.00	0.99

(c) Results on mixed modality.

Model	Sound-Music			Sound-Speech			Music-Speech			Sound-Music-Speech		
	IFR	CR	EMR	IFR	CR	EMR	IFR	CR	EMR	IFR	CR	EMR
Identity	21.16	94.79	2.23	28.11	96.15	7.69	25.82	95.71	2.86	22.08	91.92	1.71
Noise	34.09	18.14	0.00	39.03	12.42	0.00	32.15	11.88	0.00	29.97	16.18	0.00
Step-Audio-EditX	**41.30**	**44.61**	0.00	**48.79**	**66.61**	5.13	**48.66**	**65.98**	**4.00**	**44.73**	**58.47**	1.71
Ming-UniAudio	27.57	34.94	0.00	29.23	65.84	**6.15**	29.84	57.37	2.29	26.93	51.02	**2.86**
MMEdit*	49.29	48.04	4.92	**46.15**	53.25	5.08	44.04	**60.24**	**3.85**	36.84	46.25	1.85
Audio-Omni*	**57.32**	43.41	3.28	43.68	**54.47**	**8.47**	**46.19**	52.47	**3.85**	**42.37**	**50.97**	1.85
SmartDJ* w/o planer	42.40	55.77	1.64	33.99	44.28	3.39	33.27	39.05	1.92	34.82	35.34	**3.70**
SmartDJ* w/ planer	48.72	**63.67**	**8.20**	38.30	39.00	5.08	37.05	46.09	0.00	29.61	46.57	0.00

1) Higher complexity and mixed modalities degrade performance. As shown in Table 2a, all models exhibit a clear performance drop from single to multiple complexity tasks. For example, Audio-Omni's IFR decreases from 58.43% to 41.70%, and its CR drops from 64.57% to 47.94%. Similarly, comparing Table 2b and Table 2c, mixed-modality tasks are generally harder: the sound-music-speech category consistently yields the lowest scores across models, while speech editing in isolation tends to produce the highest CR (e.g., 77.27% for Step-Audio-EditX). These findings indicate that current models lack structural robustness for editing tasks that require complex reasoning and cross-domain synchronization, highlighting the critical need to bridge the gap between reactive single-operation edits and universal, mixed-modality audio manipulation.

2) IFR and CR present a fundamental trade-off. The reference baselines illustrate this clearly. The Identity baseline achieves near-perfect CR (94.13%) but poor IFR (27.37%). Its non-trivial IFR partly comes from extraction tasks or multi-round scenarios where the final output happens to match the original. The Noise baseline attains an IFR of 32.08% but an extremely low CR of 15.68%, as random noise can accidentally satisfy certain deletion-verification rubrics. Comparing these baselines to the evaluated models further highlights the inadequacy of current systems: effective audio editing requires simultaneously making precise modifications and preserving unrelated content, a balance that remains difficult. This also verifies our design of reporting IFR and CR separately rather than as a single composite score since a combined metric would allow models to trivially inflate consistency by simply not editing. The EMR metric complements this by measuring the proportion of samples where all dimensions are fully satisfied.

3) Average competency and flawless execution show divergence. Interestingly, we discover a clear decoupling between average dimension scores (IFR & CR) and perfect editing rates (EMR). Intuitively, a model with higher average sub-metrics should yield a superior exact match rate. However, our empirical results challenge this assumption: Step-Audio-EditX substantially outperforms Ming-UniAudio in both average IFR (44.86% vs. 29.82%) and CR (58.88% vs. 52.71%), yet its EMR is unexpectedly lower (3.05% vs. 3.20%). This discrepancy conceptually mirrors the "mean-seeking" versus "mode-seeking" behaviors in generative modeling. Step-Audio-EditX acts as a mean-seeking generalist editor: it achieves broad competence by partially satisfying instructions across samples, but it frequently makes minor errors that ruin the perfect edit, leading to a low EMR. Conversely, Ming-UniAudio exhibits as a mode-seeking specialist model: it fails entirely on a large portion of the data, but when it does succeed, it hits the exact target, resulting in a higher EMR. This highlights that optimizing for average metric improvements does not linearly guarantee holistic reliability for true audio editing intelligence.

4) Agent-guided planning shows limited improvement. Comparing SmartDJ with and without its planner, we observe no consistent improvement. The planner variant achieves higher IFR (42.26% vs. 38.20%) but lower CR (48.33% vs. 55.41%), and does not outperform SmartDJ w/o planner on overall EMR. Further error analysis reveals that this underperformance stems from bottlenecks in both understanding and generation. On the understanding side, the external planner still struggles with precise multimodal perception, often misinterpreting complex audio contexts. On the generation side, the base model cannot reliably execute atomic operations. Consequently, while decomposing tasks marginally improves instruction adherence (higher IFR), forcing a fragile base model through cascaded, iterative generation steps inevitably accumulates artifacts and severely degrades audio consistency (lower CR). This suggests that future research should focus on building up the base model's editing fidelity before relying purely on symbolic high-level planners.

6 Conclusion

In this work, we present **MMAE**, the first comprehensive benchmark for evaluating instruction-guided audio editing across sound, music, speech, and mixtures of them. Motivated by the lack of a unified and rigorous evaluation framework in this rapidly evolving field, MMAE introduces a systematic taxonomy spanning modality, complexity, and operation dimensions, paired with a rubric-based evaluation paradigm that enables fine-grained, objective, and scalable assessment of editing quality along both instruction following and content consistency. Our evaluation of representative audio editing models reveals that current systems, despite possessing basic editing capabilities, remain far from achieving reliable and precise edits. Overall performance is low across all evaluated models, with exact match rates below 5%, and significant degradation is observed on complex multi-operation tasks and mixed-modality scenarios. These findings underscore the substantial gap between current capabilities and the demands of real-world audio editing applications. MMAE highlights key directions in this field including improving atomic editing fidelity, developing models with universal modality support, and advancing robust agent-guided systems for compositional editing. We hope MMAE serves as an effective, challenging yet inspiring benchmark and resource for the community to track progress, identify bottlenecks, and guide future research in audio editing.

AI Use Statement

Generative AI tools were used in the benchmark construction and evaluation workflow described in this paper. The annotation pipeline used an audio-captioning agent to produce detailed captions, a language model to draft rubrics from captions, instructions, and metadata, and a language model to normalize rubric expressions. Human annotators refined the drafts, and independent inspectors reviewed the samples, as described in Section 3.4. Qwen3-Omni was used as the automated evaluation judge, and Gemini 2.0 Flash was used as the planner in the corresponding SmartDJ setting (Section 4). AI assistance was also used to migrate this manuscript to the ICLR template, reorganize its presentation, check formatting, and identify issues requiring author review. No new experimental measurements were produced during this manuscript conversion. The authors are responsible for the final manuscript, research claims, data, and artifacts, including AI-assisted material.

Ethics Statement

MMAE uses audio clips collected from online videos and evaluates editing of speech, music, sound, and their mixtures. Such material can involve copyright, privacy, speaker identity, and consent considerations. Public accessibility of a source does not establish permission for redistribution or for every downstream use. Audio-editing systems can also be misused to alter speech or impersonate speakers. Appropriate permissions and safeguards are therefore important for use and distribution of benchmark material. The human review procedure in Section 3.4 addresses annotation quality; it does not by itself establish copyright clearance or participant consent.

Reproducibility Statement

Section 3 describes the benchmark taxonomy, rubric design, statistics, and curation pipeline. Section 4 specifies the evaluated systems, duration-based subset restriction, reference baselines, judge model, repeated-query procedure, option shuffling, and reported metrics. Appendix A provides example tasks and rubrics; Appendix B gives the sample metadata schema; and Appendix C provides the evaluation prompts. These descriptions document the reported procedure; additional implementation and run details should accompany a reproducible release.

References

Xuehai Bai, Yang Shi, Yi-Fan Zhang, Xuanyu Zhu, Yuran Wang, Yifan Dai, Xinyu Liu, Yiyan Ji, Xiaoling Gu, and Yuanxing Zhang. Edit-Compass & EditReward-Compass: A Unified Benchmark for Image Editing and Reward Modeling. *arXiv preprint arXiv:2605.13062*, 2026.

Junyang Chen, Yuhang Jia, Hui Wang, Jiaming Zhou, Yaxin Han, Mengying Feng, and Yong Qin. CosyEdit: Unlocking End-to-End Speech Editing Capability from Zero-Shot Text-to-Speech Models. *arXiv preprint arXiv:2601.05329*, 2026a.

William Chen, Prem Seetharaman, Rithesh Kumar, Oriol Nieto, Shinji Watanabe, Justin Salamon, and Zeyu Jin. AudioChat: Unified Audio Storytelling, Editing, and Understanding with Transfusion Forcing. *arXiv preprint arXiv:2602.17097*, 2026b.

Daniel PW Ellis, Eduardo Fonseca, Ron J Weiss, Kevin Wilson, Scott Wisdom, Hakan Erdogan, John R Hershey, Aren Jansen, R Channing Moore, and Manoj Plakal. Recomposer: Event-roll-guided generative audio editing. *arXiv preprint arXiv:2509.05256*, 2025.

Google DeepMind. Introducing Gemini 2.0: our new AI model for the agentic era, 2024. URL https://blog.google/innovation-and-ai/models-and-research/google-deepmind/google-gemini-ai-update-december-2024/.

Google DeepMind. Gemini Omni: Native multimodal generation and video model., 2026a. URL https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-omni/.

Google DeepMind. Nano Banana 2: Google's latest AI image generation model., 2026b. URL https://blog.google/innovation-and-ai/technology/ai/nano-banana-2/.

Anisha Gunjal, Anthony Wang, Elaine Lau, Vaskar Nath, Yunzhong He, Bing Liu, and Sean Hendryx. Rubrics as rewards: Reinforcement learning beyond verifiable domains. *arXiv preprint arXiv:2507.17746*, 2025.

Rongjie Huang, Ruofan Hu, Yongqi Wang, Zehan Wang, Xize Cheng, Ziyue Jiang, Zhenhui Ye, Dongchao Yang, Luping Liu, Peng Gao, et al. Instructspeech: Following speech editing instructions via large language models. In *Proc. ICML*, 2024.

Yuhang Jia, Yang Chen, Jinghua Zhao, Shiwan Zhao, Wenjia Zeng, Yong Chen, and Yong Qin. Audioeditor: A training-free diffusion-based audio editing framework. In *Proc. ICASSP*, 2025.

Ziyue Jiang, Qian Yang, Jialong Zuo, Zhenhui Ye, Rongjie Huang, Yi Ren, and Zhou Zhao. Fluentspeech: Stutter-oriented automatic speech editing with context-aware diffusion models. In *Proc. ACL*, 2023.

Zitong Lan, Yiduo Hao, and Mingmin Zhao. Guiding audio editing with audio language model. *Proc. NeurIPS*, 2025.

Jinhua Liang, Huan Zhang, Haohe Liu, Yin Cao, Qiuqiang Kong, Xubo Liu, Wenwu Wang, Mark D Plumbley, Huy Phan, and Emmanouil Benetos. WavCraft: Audio editing and generation with large language models. *arXiv preprint arXiv:2403.09527*, 2024.

Jinhua Liang, Yuanzhe Chen, Yi Yuan, Dongya Jia, Xiaobin Zhuang, Zhuo Chen, Yuping Wang, and Yuxuan Wang. AudioMorphix: Training-free audio editing with diffusion probabilistic models. *arXiv preprint arXiv:2505.16076*, 2025.

Sen Liang, Cong Wang, Fengbin Guan, Zhentao Yu, Yiting Lu, Yuanzhi Wang, Yuan Zhou, Xin Li, and Zhibo Chen. SpongeBob: Sync-Aware Harmonious Audio-Visual Generative Editing. *arXiv preprint arXiv:2605.25193*, 2026.

Ziyang Ma, Ruiyang Xu, Yinghao Ma, Chao-Han Huck Yang, Bohan Li, Jaeyeon Kim, Jin Xu, Jinyu Li, Carlos Busso, Kai Yu, et al. The interspeech 2026 audio reasoning challenge: Evaluating reasoning process quality for audio reasoning models and agents. *arXiv preprint arXiv:2602.14224*, 2026a.

Ziyang Ma, Ruiyang Xu, Zhenghao Xing, Yunfei Chu, Yuxuan Wang, Jinzheng He, Jin Xu, Pheng-Ann Heng, Kai Yu, Junyang Lin, et al. Omni-captioner: Data pipeline, models, and benchmark for omni detailed perception. *Proc. ICLR*, 2026b.

Hila Manor and Tomer Michaeli. Zero-shot unsupervised and text-based audio editing using DDPM inversion. *Proc. ICML*, 2024.

Puyuan Peng, Po-Yao Huang, Shang-Wen Li, Abdelrahman Mohamed, and David Harwath. Voicecraft: Zero-shot speech editing and text-to-speech in the wild. In *Proc. ACL*, 2024.

Ye Tao, Wen Wu, Chao Zhang, Mengyue Wu, Shuai Wang, and Xuenan Xu. MMEDIT: A Unified Framework for Multi-Type Audio Editing via Audio Language Model. *arXiv preprint arXiv:2512.20339*, 2025.

Qwen Team. Qwen3.5-omni technical report. *arXiv preprint arXiv:2604.15804*, 2026.

Zeyue Tian, Binxin Yang, Zhaoyang Liu, Jiexuan Zhang, Ruibin Yuan, Hubery Yin, Qifeng Chen, Chen Li, Jing Lv, Wei Xue, et al. Audio-Omni: Extending Multi-modal Understanding to Versatile Audio Generation and Editing. *Proc. SIGGRAPH*, 2026.

Helin Wang, Meng Yu, Jiarui Hai, Chen Chen, Yuchen Hu, Rilin Chen, Najim Dehak, and Dong Yu. Ssr-speech: Towards stable, safe and robust zero-shot text-based speech editing and synthesis. In *Proc. ICASSP*, 2025.

Yuancheng Wang, Zeqian Ju, Xu Tan, Lei He, Zhizheng Wu, Jiang Bian, et al. Audit: Audio editing by following instructions with latent diffusion models. *Proc. NeurIPS*, 2023.

Yusong Wu, Christos Tsirigotis, Ke Chen, Cheng-Zhi Anna Huang, Aaron Courville, Oriol Nieto, Prem Seetharaman, and Justin Salamon. Flam: Frame-wise language-audio modeling. *Proc. ICML*, 2025.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025.

Manjie Xu, Chenxing Li, Dan Su, Wei Liang, Dong Yu, et al. Prompt-guided precise audio editing with diffusion models. *Proc. ICML*, 2024.

Canxiang Yan, Chunxiang Jin, Dawei Huang, Haibing Yu, Han Peng, Hui Zhan, Jie Gao, Jing Peng, Jingdong Chen, Jun Zhou, et al. Ming-UniAudio: Speech LLM for Joint Understanding, Generation and Editing with Unified Representation. *arXiv preprint arXiv:2511.05516*, 2025a.

Chao Yan, Boyong Wu, Peng Yang, Pengfei Tan, Guoqiang Hu, Li Xie, Yuxin Zhang, Fei Tian, Xuerui Yang, Xiangyu Zhang, et al. Step-Audio-EditX Technical Report. *arXiv preprint arXiv:2511.03601*, 2025b.

Haojie Zheng, Yixin Yang, Siqi Yang, Shuchen Weng, and Boxin Shi. Instructav2av: Instruction-guided audio-video joint editing. *arXiv preprint arXiv:2605.18467*, 2026.

Zhisheng Zheng, Puyuan Peng, Anuj Diwan, Cong Phuoc Huynh, Xiaohang Sun, Zhu Liu, Vimal Bhat, and David Harwath. VoiceCraft-X: Unifying Multilingual, Voice-Cloning Speech Synthesis and Speech Editing. In *Proc. EMNLP*, 2025.

A Demo Examples

We present representative samples from the MMAE benchmark to illustrate the diversity of tasks and the granularity of rubric-based evaluation.

Case 1

multi-audio | *music* | *Global – Foreground change*

Instruction: Change all the words in the lyrics of audio2 to "Hachimi" and use the human voice timbre of audio1.

#	Category	Rubric
1	IF	**Q:** In <audio output>, what are the words mainly sung clearly by the lead vocalist? ✔ Repeatedly sung onomatopoeic word "Hachimi" ✘ Mainly German lyrics (such as "Zuerst lag ich in einem Ei") ✘ Repeatedly sung Japanese "はちみつ " ✘ Cannot hear the specific words clearly (like humming / no lyrics) ✘ Other lyrics, or lyrics mixing multiple languages ✘ None of the above
2	IF	**Q:** Does the timbre of the human voice in <audio output> sound like <audio input1>? ✔ Yes ✘ No ✘ No human voice heard ✘ None of the above
3	Con.	**Q:** Compare <audio output> and <audio input2>: is the overall accompaniment basically the same? ✔ Yes, the overall accompaniment is basically the same ✘ No, the accompaniment is clearly changed / replaced / has elements greatly added or removed ✘ No accompaniment ✘ None of the above
4	Con.	**Q:** Compare <audio output> and <audio input2>. Are the melody and rhythm of the vocals basically the same? ✔ Yes, basically consistent ✘ No ✘ No vocals ✘ None of the above
5	Con.	**Q:** Compare <audio output> and <audio input2>: Does <audio output> exhibit obvious audio quality degradation (such as newly added background noise/electrical hum, obvious distortion clipping, metallic compression artifacts, or pumping that becomes suddenly louder and quieter)? ✔ No, no obvious audio quality degradation is heard ✘ Yes, obvious audio quality degradation can be heard ✘ None of the above

---

Case 2

multi-round | *speech* | *Local – Replacement*

Round 1 Instruction: Change the order of the first and second authors mentioned in the commit.

Round 2 Instruction: Change the order of the second and third authors mentioned in the commit.

#	Category	Rubric
1	IF	**Q:** In <audio output[0s:]>, in this continuous reading aloud of names, which of the following best matches the order of the first three names read? ✔ Hongyi Li → Shinji Watanabe → Abdulrahim Mohammed ✘ Abdulrahim Mohammed → Hongyi Li → Shinji Watanabe ✘ Hongyi Li → Abdulrahim Mohammed → Shinji Watanabe ✘ Shinji Watanabe → Hongyi Li → Abdulrahim Mohammed ✘ None of the above
2	IF	**Q:** Compare <audio output[2.15s:4.9s]> and <audio input1[4.35s:7.6s]>: Is the name that is read aloud first in the <audio output[2.15s:4.9s]> segment the same name content as "Me Hongyi Li," in the <audio input1[4.35s:7.6s]> segment? ✔ Yes, both are "Me Hongyi Li" ✘ No, it is not this name or it is unclear ✘ None of the above
3	IF	**Q:** Compare <audio output[4.4s:7.9s]> and <audio input1[6.6s:9.8s]>: Is the name read immediately afterward in the <audio output[4.4s:7.9s]> segment the same name content as "Shinji Watanabe," in the <audio input1[6.6s:9.8s]> segment? ✔ Yes, both are "Shinji Watanabe" ✘ No, it is not that name or it is unclear ✘ None of the above
4	IF	**Q:** Compare <audio output[6.8s:10.3s]> and <audio input1[2.15s:5.35s]>: Is the name subsequently read aloud in the <audio output[6.8s:10.3s]> segment the same name content as "Abdulrahim Mohammed," in the <audio input1[2.15s:5.35s]> segment? ✔ Yes, both are "Abdulrahim Mohammed" ✘ No, it is not that name or is unclear ✘ None of the above
5	IF	**Q:** In <audio output[2.15s:9.8s]>, which of the following groups of names can all be clearly heard (order not required), and each appears at least once? ✔ Abdulrahim Mohammed, Hongyi Li, Shinji Watanabe ✘ Tara Sainath, Karen Livescu, Shang-Wen Li ✘ Yvonne Dambhar, Emmanuel Dupoux, Tara Sainath ✘ Only one or two of them can be heard, and they cannot be identified as a group ✘ None of the above
6	Con.	**Q:** Compare <audio output[0s:2.7s]> and <audio input1[0s:2.7s]>: Can the same sentence "And this is the organization committee," be heard in both segments, and are the speaker's timbre and tone consistent? ✔ Yes, the content and speaker characteristics are consistent ✘ No, the content is missing/different or the speaker characteristics are clearly different ✘ None of the above
7	Con.	**Q:** Compare <audio output[1.7s:3.15s]> and <audio input1[1.7s:3.15s]>: Can a softly spoken filler word "uh," (around approximately 00:02.2) be heard in both segments, accompanied by a similar brief pause? ✔ Yes, both segments have a similar "uh," and pause ✘ No, the "uh," is clearly absent or very different in <audio output[1.7s:3.15s]> or <audio input1[1.7s:3.15s]> ✘ None of the above

#	Category	Rubric

8	Con.	**Q:** Compare <audio output[2.15s:3.45s]> and <audio input1[2.15s:3.45s]>: Is the word "include," present in both segments and similar in pronunciation/rhythm? ✔ Yes, include is consistently present and pronounced similarly ✘ No, include is missing/changed into another word/has obvious differences ✘ None of the above
9	Con.	**Q:** Compare <audio output[8.8s:11.9s]> and <audio input1[8.8s:11.9s]>: Are the names read in the two segments both "Tara Sainath," (including the feeling of a comma-like pause at the end)? ✔ Yes, both are "Tara Sainath," ✘ No, the names are different or unclear ✘ None of the above
10	Con.	**Q:** Compare <audio output[10.9s:13.7s]> and <audio input1[10.9s:13.7s]>: Are the names read in the two segments both "Karen Livescu,"? ✔ Yes, both are "Karen Livescu," ✘ No, the names are different or unclear ✘ None of the above
11	Con.	**Q:** Compare <audio output[12.7s:15.1s]> and <audio input1[12.7s:15.1s]>: are the names read aloud in the two segments both "Shang-Wen Li,"? ✔ Yes, both are "Shang-Wen Li," ✘ No, the names are different or unclear ✘ None of the above
12	Con.	**Q:** Compare <audio output[14.1s:15.8s]> and <audio input1[14.1s:15.8s]>: In both segments, can "Yvonne Dambhar," be heard, immediately followed by a soft "uh," and then entering the final name? ✔ Yes, the structure and content are consistent (Yvonne Dambhar → uh → final name) ✘ No, some part is missing or the order/content is clearly different ✘ None of the above
13	Con.	**Q:** Compare the entire <audio output> and <audio input1>: Is the overall audio quality of <audio output> not significantly degraded (e.g., no newly introduced distortion, obvious noise floor, pops, clipping, overly strong compression / metallic artifacts)? ✔ Yes, no obvious newly introduced audio-quality degradation is heard ✘ No, obvious newly introduced audio-quality degradation can be heard ✘ None of the above

---

Case 3

multi-hop | *sound* | *Local – Removal*

Instruction: Remove barks from younger dogs.

#	Category	Rubric
1	IF	**Q:** In <audio output>, can clear dog barking / dog yelping transient sounds (short, sharp "woof / whine" type) be heard? ✔ Can be heard ✘ Cannot be heard ✘ None of the above
2	Con.	**Q:** In <audio output>, can multiple short dog barks (possibly slightly overlapping) be heard? ✔ Multiple dog barks / barking can be heard ✘ No dog barking can be heard; overall it is close to silence ✘ None of the above

#	Category	Rubric
3	IF	**Q:** In <audio output>, is it possible to hear one or two short dog barks / dog vocalizations at the end? ✔ Can be heard; there is still dog barking / dog vocalization at the end ✘ Cannot be heard; there is no dog barking / dog vocalization at the end ✘ None of the above
4	Con.	**Q:** Compare the overall background of <audio output> and <audio input1> (regardless of whether there is dog barking). Do both present near-complete silence, with no added background noise/hum/ambient sound? ✔ Yes, the backgrounds of both are similarly close to silent and have no added sounds ✘ No, there is a clear difference in background noise between <audio output> and <audio input1>, or added ambient sound appears ✘ None of the above
5	Con.	**Q:** Compare <audio output> and <audio input1>. Does <audio output> contain any new non-canine sounds that are not present in <audio input1> (such as cat meowing, bird calls, human voices, music, or mechanical sounds)? ✔ No, no new non-canine sounds appeared ✘ Yes, at least one new non-canine sound appeared ✘ None of the above
6	Con.	**Q:** Compare the overall audio quality of <audio output> and <audio input1>. Does <audio output> show noticeable degradation (such as distortion, clipping, over-compression/pumping, obvious quantization noise, or high-frequency graininess)? ✔ No, <audio output> does not show noticeable audio quality degradation ✘ Yes, <audio output> shows noticeable audio quality degradation ✘ None of the above
7	IF	**Q:** Does <audio output> contain only one type of dog barking sound? ✔ Yes ✘ No, it contains at least 2 types of dog barking sounds. ✘ None of the above
8	IF	**Q:** Which type of dog bark in <audio input1> is the dog barking sound in <audio output> consistent with? ✔ The sharp, thin, clear bark of a young dog. ✘ The deep, melodious bark of an older dog. ✘ None of the above

---

Case 4

multi-part | sound-music-speech | Global – Foreground change

Instruction: Change the accented Chinese dialogue in this clip to standard Mandarin pronunciation.

#	Category	Rubric
1	IF	**Q:** Compare <audio output[0.0s:1.7s]> and <audio input1[0.0s:1.7s]>. In which segment is the male voice's Chinese pronunciation closer to "Standard Mandarin" (fewer regional accent features)? ✔ <audio output[0.0s:1.7s]> is closer to Standard Mandarin ✘ <audio input1[0.0s:1.7s]> is closer to Standard Mandarin ✘ The two are comparable in degree of Mandarin, with no obvious difference audible ✘ None of the above

#	Category	Rubric
2	IF	**Q:** Compare <audio output[11.8s:16.0s]> and <audio input1[11.8s:16.0s]>. In which segment is the male voice's Chinese pronunciation closer to "Standard Mandarin" (with fewer regional accent characteristics)? ✔ <audio output[11.8s:16.0s]> is closer to Standard Mandarin ✘ <audio input1[11.8s:16.0s]> is closer to Standard Mandarin ✘ No obvious difference can be heard between the two ✘ None of the above
3	Con.	**Q:** In <audio output[0.0s:1.4s]>, which of the following does this line sound more like? ✔ "我要验牌" ✘ Not this sentence (the words have clearly changed or the content is inconsistent) ✘ Unclear / covered by other sounds, making it impossible to judge ✘ None of the above
4	Con.	**Q:** In <audio output[11.8s:14.5s]>, which of the following does this line sound more like? ✔ "牌没有问题" ✘ Not this line (the words have clearly changed or the content is inconsistent) ✘ Unclear / covered by other sounds, making it impossible to judge ✘ None of the above
5	IF	**Q:** Compare <audio output[0.0s:1.4s]> and <audio input1[0.0s:1.4s]>. Do the two sound like the same adult male voice speaking (timbre and identity characteristics basically consistent), rather than having changed to another speaker? ✔ They are the same male voice (timbre/identity basically consistent) ✘ No, it sounds like a different person (gender/age/timbre clearly different) ✘ Unable to determine ✘ None of the above
6	IF	**Q:** Compare <audio output[11.8s:14.5s]> and <audio input1[11.8s:14.5s]>. Do they sound like the same adult male voice speaking (with basically consistent timbre and identity characteristics), rather than having switched to another speaker? ✔ The same male voice (timbre/identity basically consistent) ✘ No, it sounds like it switched to another person (gender/age/timbre clearly different) ✘ Unable to determine ✘ None of the above
7	Con.	**Q:** Compare <audio output[0.0s:1.4s]> and <audio input1[0.0s:1.4s]>. Are their speaking rate and pause positions basically consistent (disregarding accent differences, only considering whether rhythm/duration/start and end points are aligned)? ✔ Basically consistent (rhythm and start/end points are about the same) ✘ Inconsistent (clearly faster/slower, lengthened/shortened, or alignment points changed) ✘ Cannot determine ✘ None of the above
8	Con.	**Q:** Compare <audio output[11.8s:14.5s]> and <audio input1[11.8s:14.5s]>. Are their speaking rate and pause positions basically consistent (disregarding accent differences, only considering whether rhythm/duration/start and end points are aligned)? ✔ Basically consistent (rhythm and start/end points are about the same) ✘ Inconsistent (clearly faster/slower, lengthened/shortened, or alignment points changed) ✘ Unable to determine ✘ None of the above

#	Category	Rubric
9	Con.	**Q:** Compare <audio output[0.4s:12.5s]> and <audio input1[0.4s:12.5s]>. Are the close-up plastic/cellophane rubbing sounds (timing of occurrence, changes in density, and overall intensity trend) basically the same? ✔ Basically the same (sounds like the same segment of rubbing sound) ✘ Not the same (replaced / clearly distorted / different intensity trend) ✘ The rubbing sound in <audio output[0.4s:12.5s]> is clearly missing, or a clearly different new ambient sound appears ✘ None of the above
10	Con.	**Q:** Compare <audio output[14.0s:16.0s]> and <audio input1[14.0s:16.0s]>. Is the quiet tail segment at the end (residual reverberation and slight background noise) basically consistent, with no additional newly added sound effects or human voices? ✔ Basically consistent and no newly added elements ✘ Inconsistent: newly added sound effects/human voices appear, or the background noise/tail sound is noticeably altered ✘ Unable to determine ✘ None of the above
11	Con.	**Q:** Compare the entire <audio output> with the entire <audio input1>. Does the overall audio quality of <audio output> show degradation (such as obvious distortion/clipping, watery sounds caused by excessive noise reduction, obvious compression pumping, an abnormal increase in noise, or a significant decrease in clarity)? ✔ No obvious degradation heard ✘ Obvious degradation heard ✘ Unable to determine ✘ None of the above

---

Case 5

multi-instruction | *music-speech* | *Global – Background change* + *Global – Alteration*

Instruction: Change the background music to a guitar with the exactly same melody, while making the vocals deeper and more resonant, without changing the spoken content.

#	Category	Rubric
1	IF	**Q:** Compare the main accompaniment melody carrier in <audio input1[0.0s:10.0s]> and <audio output[0.0s:10.0s]>. Which one more clearly exhibits the characteristics of "guitar plucking / strumming / acoustic guitar or electric guitar timbre"? ✔ <audio output[0.0s:10.0s]> is more clearly dominated by guitar timbre ✘ <audio input1[0.0s:10.0s]> is more clearly dominated by guitar timbre ✘ Neither is clearly dominated by guitar timbre ✘ None of the above
2	IF	**Q:** At the melodic level, compare <audio input1[0.0s:10.0s]> and <audio output[0.0s:10.0s]>: do the "pitch direction and rhythmic contour" of the main melody/motif of the two sound basically consistent (like the same melody presented by different instruments)? ✔ Basically consistent ✘ Clearly inconsistent (the melody/motif has clearly changed) ✘ The melody in this segment of <audio output[0.0s:10.0s]> or <audio input1[0.0s:10.0s]> is unclear, making it difficult to compare ✘ None of the above

#	Category	Rubric
3	IF	**Q:** Compare the perceived pitch of the male narration voice in <audio input1[0.0s:10.0s]> and <audio output[0.0s:10.0s]>. Which one is more "deep" (overall pitch is lower)? ✔ <audio output[0.0s:10.0s]> is deeper ✘ <audio input1[0.0s:10.0s]> is deeper ✘ The difference between the two is not obvious ✘ None of the above
4	IF	**Q:** In <audio output[0.7s:4.0s]>, can the narration's spoken content still be recognized as "最高の伪装とは、自らを欺くこと。"? ✔ Yes, it can be recognized as this sentence ✘ No, it can be recognized as different content/sentence ✘ The human voice is too unclear to recognize ✘ None of the above
5	IF	**Q:** In <audio output[5.5s:9.3s]>, can the narrator's spoken content still be recognized as "俺は教練部の書記官、アルハイゼンだ。"? ✔ Yes, it can be recognized as this sentence ✘ No, it can be recognized as different content / a different sentence ✘ The voice is too unclear to recognize ✘ None of the above
6	Con.	**Q:** Compare the ending treatment of <audio input1> and <audio output>: do both present a structure where "near the end, all elements suddenly stop simultaneously, and the ending is very clean" (rather than an obvious fade-out or a very long trailing tail sound)? ✔ Yes, the ending structures of both are consistent ✘ No, the ending structures of <audio output> and <audio input1> are clearly different ✘ None of the above
7	Con.	**Q:** Compare <audio input1[3.5s:5.6s]> and <audio output[3.5s:5.6s]>: does the transition from the buildup into a denser, stronger beat occur at a similar time position (without being obviously much earlier/later)? ✔ Yes, the time position of the transition point is roughly consistent ✘ No, the transition point in <audio output[3.5s:5.6s]> is clearly much earlier/later or missing ✘ None of the above
8	Con.	**Q:** Compare <audio input1[0.0s:1.7s]> and <audio output[0.0s:1.7s]>: Do both beginnings present a relatively sparse build-up and gradually introduce the beat/layers (rather than being fully arranged from the start or having almost no content for a long time)? ✔ Yes, the overall build-up approach at the beginning is consistent ✘ No, the build-up approach at the beginning of <audio output[0.0s:1.7s]> has clearly changed ✘ None of the above
9	Con.	**Q:** Compare <audio input1[0.7s:3.5s]> and <audio output[0.7s:3.5s]>: Are the timing of the narration's appearance and its duration roughly consistent (not obviously advanced/delayed, truncated, or repeated)? ✔ Yes, the timing and duration are roughly consistent ✘ No, the narration timing in <audio output[0.7s:3.5s]> is obviously inconsistent (such as misaligned/truncated/repeated) ✘ None of the above

#	Category	Rubric

10	Con.	**Q:** Compare <audio input1[4.5s:10.0s]> and <audio output[4.5s:10.0s]>: Is the speed of the four-beat beat-hitting (overall perceived speed) kept consistent (without obvious speeding up or slowing down)? ✔ Yes, the perceived speed is consistent ✘ No, the perceived speed of <audio output[4.5s:10.0s]> is clearly different ✘ None of the above
11	Con.	**Q:** Compare <audio input1> and <audio output>: apart from differences in instrument timbre and vocal timbre, does <audio output> contain significant newly added content not present in <audio input1> (such as added vocal layers, abrupt environmental noise, or additional sound-effect passages)? ✔ No obvious newly added content heard ✘ Obvious newly added content heard ✘ None of the above
12	Con.	**Q:** Compare the overall listening impression of <audio input1> and <audio output>: does <audio output> exhibit obvious audio quality degradation (such as heavier background noise / hum, distortion or clipping, obvious compression pumping, harsh pops, or high-frequency graininess), causing a decrease in clarity? ✔ No obvious audio quality degradation ✘ Obvious audio quality degradation ✘ None of the above

---

Case 6

single | *sound-speech* | *Local – Extraction*

Instruction: Isolate and extract all sounds produced by dogs, such as barking or whining, while suppressing human speech and other environmental background noises.

#	Category	Rubric
1	IF	**Q:** In <audio output>, can any human speech content be clearly heard (recognizable words or sentences)? ✔ No, basically no recognizable words or sentences can be heard ✘ Yes, human speech words or sentences can be recognized ✘ None of the above
2	IF	**Q:** In <audio output[4.0s:5.7s]>, is it possible to hear an obvious human shouting/speaking timbre (it is not required to make out the specific words)? ✔ Cannot hear obvious human shouting/speaking ✘ Can hear obvious human shouting/speaking ✘ None of the above
3	IF	**Q:** In <audio output[2.0s:15.0s]>, can obvious canine vocalizations such as dog barking / dog howling / whimpering still be heard? ✔ Yes, obvious canine vocalizations can be heard ✘ No, no obvious canine vocalizations can be heard ✘ None of the above
4	IF	**Q:** In <audio output[5.0s:10.0s]>, which of the following categories does the dominant sound best fit? ✔ Canine long howling / barking ✘ Human speech / shouting ✘ Mainly environmental noise or nearly silent ✘ None of the above

B Final Meta-Data Format

Each sample in the MMAE benchmark is stored as a JSON object. Below is a complete example illustrating the final released format:

Listing 1: Sample JSON annotation of MMAE.


```
{
    "id": "69e898163a050f39ac567501",
    "complexity": "single",
    "modality": "sound-speech",
    "granularity": [
        "local"
    ],
    "operations": [
        {
            "granularity": "local",
            "operation": "extraction"
        }
    ],
    "messages": [
        {
            "role": "user",
            "content": [
                {
                    "type": "text",
                    "text": "Isolate and extract all sounds produced by dogs, such as barking or whining, while suppressing human speech and other environmental background noises."
                },
                {
                    "type": "audio",
                    "audio_url": "wav/69e898163a050f39ac567501/audio1.wav"
                }
            ]
        }
    ],
    "tags": [
        [
            "Acoustic Event Detection / Sound Event Extraction",
            "Canine Vocalizations (Dog Barking / Whining)",
            "Non-Human Audio Signal",
            "Background Noise & Speech Suppression"
        ]
    ],
    "rubrics": [
        {
            "category": "Instruction Following",
            "question": "In <audio output>, can any human speech content be clearly heard (recognizable words or sentences)?",
            "right_choice": "No, basically no recognizable words or sentences can be heard",
            "wrong_choices": [
                "Yes, human speech words or sentences can be recognized",
                "None of the above"
            ]
        },
        {
            "category": "Instruction Following",
            "question": "In <audio output[4.0s:5.7s]>, is it possible to hear an obvious human shouting/speaking timbre (it is not required to make out the specific words)?",
```

```
            "right_choice": "Cannot hear obvious human shouting/speaking
    ",
            "wrong_choices": [
                "Can hear obvious human shouting/speaking",
                "None of the above"
            ]
        },
        {
            "category": "Instruction Following",
            "question": "In <audio output[2.0s:15.0s]>, can obvious
    canine vocalizations such as dog barking / dog howling / whimpering
    still be heard?",
            "right_choice": "Yes, obvious canine vocalizations can be
    heard",
            "wrong_choices": [
                "No, no obvious canine vocalizations can be heard",
                "None of the above"
            ]
        },
        {
            "category": "Instruction Following",
            "question": "In <audio output[5.0s:10.0s]>, which of the
    following categories does the dominant sound best fit?",
            "right_choice": "Canine long howling / barking",
            "wrong_choices": [
                "Human speech / shouting",
                "Mainly environmental noise or nearly silent",
                "None of the above"
            ]
        }
    ]
}
```

C Evaluation Prompt

Below is the exact prompt used to query the external MLLM judger (Qwen3-Omni) for rubric-based evaluation. The system prompt establishes the role and guidelines, while the user prompt provides task-specific instructions for answering each rubric question.

System Prompt

```
You are an audio analysis assistant.

Your task is to answer user questions strictly based on the factual
content of the provided audio clips. Carefully listen to and
analyze the audio before answering.

Audio reference notation:
- A bare reference like <audio output> or <audio input1> refers to
the full duration of that audio clip.
- A reference with [start:end] denotes a time slice of that clip
(e.g. <audio output[0.0s:3s]> means the segment from 0.0s to 3.0s).
- An omitted start means the beginning of the clip (e.g. <audio
input1[:2.0s]> means from the beginning up to 2.0s).
- An omitted end means the end of the clip (e.g. <audio
output[1.5s:]> means from 1.5s to the end).
- A negative value counts backward from the end (e.g. <audio
input2[-2.5s:]> means the last 2.5 seconds of that clip).
- The slicing semantics follow Python-style [start:end] conventions.

Guidelines:
- Base your answers only on information that is clearly present or
can be directly inferred from the audio.
- Do NOT make up details that are not supported by the audio.
- If the audio does not contain enough information to answer the
question, say so explicitly.
- When relevant, reference specific parts of the audio (e.g., events,
sounds, speech content, timing).
- Be concise, clear, and accurate.
- If the audio contains speech, you may transcribe or summarize
relevant portions to support your answer.
- If the audio is noisy, ambiguous, or unclear, acknowledge the
uncertainty.

Your goal is to provide reliable, evidence-based answers grounded
in the audio content only.
```

User Prompt

```
<audio {label1}>: {audio1}

[<audio {label2}>: {audio2}]

[...]

Based on the objective content of the uploaded audio, answer the
following multiple-choice question. First carefully perceive,
analyze, and reason about the audio content, then choose exactly
one option from the list. Return only JSON with keys ``reason''
and ``choice''. Put the reason first and the final choice last. The
choice value must be a single uppercase letter identifying the
option.

Question:
{question}

Choices:
{choices}
```

D Data Curation Platform

We conducted data annotation and quality inspection using a professional platform that supports structured editing, version control, and multi-stage review. All meta data and rubrics were annotated, reviewed and corrected through this system. A Snapshot of the platform interface is shown in Figure 4.

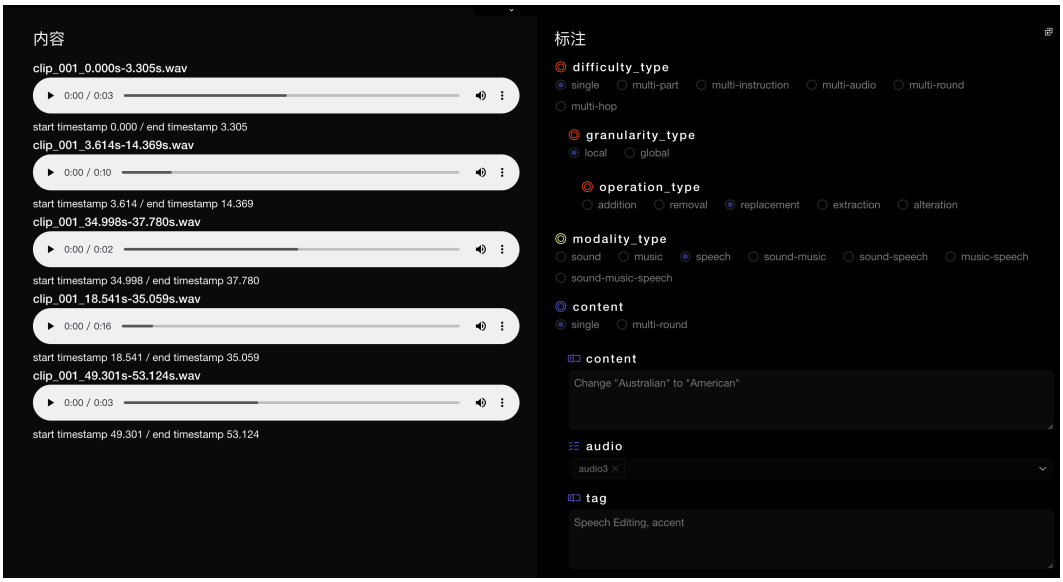


Figure 4: A Snapshot of the platform used for data annotation and quality inspection.