MMAG: A MULTI-CONTROL MIXED AUDIO GENERATION BENCHMARK

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent audio generation systems have progressed from single-modality synthesis to generating complex acoustic scenes containing speech, music, and sound effects. Therefore, evaluating these models requires assessing multiple interacting capabilities, including semantic fidelity, speaker consistency, and temporal control, yet existing benchmarks focus on isolated domains or coarse-grained descriptions. To address this gap, we introduce the Multi-control Mixed Audio Generation (MMAG) benchmark. MMAG contains approximately 4,000 manually verified audio clips with rich annotations covering speech content, speaker identity, music attributes, sound events, and temporal relationships, together with dedicated subsets for voice cloning and timestamp-conditioned generation. We further propose a systematic evaluation protocol that measures acoustic fidelity, speech quality, semantic alignment, and temporal accuracy. Benchmarking representative agentic orchestrators, unified audio-visual generation models, and native mixed-audio generators reveals substantial performance trade-offs across these capabilities, with no existing model performing consistently well. Our results highlight the remaining challenges of controllable mixed audio generation and establish MMAG as a comprehensive benchmark for future research. [1]

1 INTRODUCTION

Research in audio generation has historically progressed along three parallel tracks, targeting speech, music, and sound effects separately. Each track has developed its own specialized model architectures and datasets, with little cross-domain interaction (Hung et al., 2026; Zhou et al., 2026; Copet et al., 2023).

With the development of generative models, several works have started to integrate different audio domains. These efforts fall into three categories: agentic orchestrators that compose scenes via expert models, unified audio-visual generators that produce synchronized video-audio, and native mixed-audio generators that unify speech, music, and effects in a single end-to-end system.

The emergence of mixed audio generation models poses new challenges for existing benchmarks. The first challenge lies in limited audio content and textual annotations. Most existing benchmarks are designed for single-domain tasks: Text-to-Speech (TTS) test sets such as LibriSpeech-PC (Meister et al., 2023) contain only clean speech, and Text-to-Music (TTM) benchmarks like Music-Bench (Melechovsky et al., 2024) consist solely of instrumental music, neither covering cross-domain content. In contrast, Text-to-Audio (TTA) test sets such as AudioCaps (Kim et al., 2019) do include speech and sound elements but suffer from coarse captions that omit speech transcriptions and speaker details. MECAT (Niu et al., 2026) produces richer captions for mixed scenes, yet it lacks temporal ordering of sound events for generation, which would require a more comprehensive annotation pipeline.

The second challenge is the absence of fine-grained control mechanisms in existing benchmarks, as users in practice demand not only high-quality audio but also precise controllability over specific attributes. One example is zero-shot voice cloning in TTS (Zhou et al., 2026), where the model mimics a speaker's timbre from a short voice prompt. This necessitates that the benchmark provide suitable prompt samples. However, selecting such segments from real-world audio is challenging due to complex acoustic scenes. Another example is temporal control in TTA, which requires the model

[1] The project page is available: `https://anonymous3387a8c947.github.io/MMAG/`

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.

Table 1: Comparison with existing audio-generation benchmarks. Domain denotes the covered audio categories, while Condition denotes the available annotations or control signals.

Benchmark	Size	Domain			Condition						
		Speech	Sound	Music	Transcript	Speaker attr.	Sound event	Music info	Temporal order	Voice prompt	Timestamp
LibriSpeech-PC	1.2k	✓	✗	✗	✓	✗	✗	✗	✗	✓	✗
MusicBench	0.4k	✗	✗	✓	✗	✗	✗	✓	✗	✗	✗
AudioCaps	4.8k	✓	✓	✗	✗	✗	✓	✗	✓	✗	✗
MECAT	20k	✓	✓	✓	✓	✓	✓	✓	✗	✗	✗
MMAG (Ours)	**4k**	✓	✓	✓	✓	✓	✓	✓	✓	✓**0.7k**	✓**1.8k**

to generate sound events at precisely specified timestamps (Xie et al., 2025). This requires obtaining precise timestamps for each element during the annotation process.

To address these challenges, we introduce MMAG, a benchmark for Multi-control Mixed Audio Generation. We construct three evaluation settings that differ in conditioning signals: (1) text-only control on the main set, (2) extra speaker identity control via a short voice prompt, and (3) temporal control via timestamp-detailed captions. We select cross-domain audio samples from test sets of AudioCaps, VGGSound, and MECAT. Extending the multi-expert annotation pipeline of MECAT with a timestamp-labeling branch, we construct detailed captions including speech transcriptions, speaker attributes (age, gender, dialect, and emotion), sound event descriptions, musical information (genre, instrument and mood), and temporal ordering. After manual inspection, the main set, voice cloning subset, and timestamp subset comprise around 3,980, 690, and 1,800 audio-text pairs respectively, with the latter two providing reference voices and precise temporal boundaries respectively. Table 1 compares MMAG with existing benchmarks across coverage of domains and conditions.

Using MMAG, we systematically evaluate representative models from these emerging system families. To provide a holistic assessment of mixed audio generation, we establish an evaluation protocol across different dimensions: acoustic fidelity, speech quality, semantic consistency, and temporal controllability. Our findings reveal that existing models struggle to achieve balanced performance across different metrics. Moreover, introducing voice prompts leads to a noticeable performance drop, and most models exhibit little capability for fine-grained temporal control.

Our contributions are summarized as follows:

1. We introduce **MMAG**, the first benchmark for *compositional evaluation* of mixed audio generation. MMAG comprises a manually verified benchmark with rich annotations spanning speech, music, sound effects, and temporal relationships, together with dedicated subsets for prompt-based voice cloning and timestamp-conditioned generation.
2. We establish a unified evaluation framework that jointly measures acoustic fidelity, speech quality, semantic consistency and temporal control, providing a holistic assessment of mixed audio generation beyond conventional single-metric evaluation.
3. We benchmark representative agentic orchestrators, unified audio-visual models, and native mixed-audio generators under this unified protocol, revealing that current systems exhibit substantial trade-offs across generation quality and controllability, with no single model consistently excelling across all evaluation dimensions.

2 Related Work

Benchmarks for audio generation Existing benchmarks for audio generation include AudioCaps (Kim et al., 2019) and VGGSound (Chen et al., 2020), which provide large-scale real-world audio with coarse annotations. MECAT (Niu et al., 2026) is built from ACAV100M (Lee et al., 2021), which also serves as a valuable audio source and offers richer annotations through expert-model and LLM combination. However, none of these resources are designed for compositional evaluation of mixed audio generation.

Evaluation of audio generation Existing metrics for audio generation have been developed along separate lines, each targeting a specific quality dimension. For distributional metrics, Fréchet Audio

Distance (FAD) (Kilgour et al., 2018a), Fréchet Distance (FD), Kullback-Leibler (KL) divergence, and Inception Score (IS) (Salimans et al., 2016) are widely used to quantify how closely the generated audio matches the real distribution. Alongside these metrics, no-reference predictors such as Audiobox Aesthetics (Tjandra et al., 2025) provide perceptual quality assessments across multiple axes. For speech metrics, Word Error Rate (WER), speaker similarity (SPK-SIM), and UTMOS[2] address transcription accuracy, identity preservation, and naturalness respectively. For semantic metrics, Contrastive Language-Audio Pretraining (CLAP) score (Wu et al., 2023) provides global text-audio alignment. AnyAudio-Judge (Li et al., 2026) achieves fine-grained semantic evaluation by decomposing captions into atomic rubrics, as demonstrated on hard-negative mixed-audio pairs. While these metrics cover complementary aspects of quality, they remain scattered across isolated benchmarks, with no unified protocol for evaluating controllable mixed audio generation.

Audio generation systems Recent audio generation systems can be grouped into three paradigms.

Agentic orchestrators, represented in our evaluation by AuDirector (Ren et al., 2026), employ LLMs as central agents to invoke expert models including TangoFlux (Hung et al., 2026), IndexTTS2 (Zhou et al., 2026), and MusicGen (Copet et al., 2023). AuDirector further incorporates critic models such as CLAP and MiMo-Audio (Xiaomi, 2025) to iteratively evaluate and refine generated audio quality. This training-free design enables flexible adaptation to diverse conditioning inputs, including voice references or timestamps.

Unified audio-visual generation models accept diverse input modalities and synthesize synchronized audio-visual content. Representative models in our evaluation include MOVA (SII-OpenMOSS Team et al., 2026), Ovi (Low et al., 2025), UniAVGen (Zhang et al., 2025), LTX-2 (HaCohen et al., 2026), and JavisDiT++ (Liu et al., 2026), which can synthesize synchronized video alongside general audio covering speech, music, and sound effects.

Native mixed-audio generators, represented by Dasheng AudioGen (Mei et al., 2026), synthesize mixed audio scenes from text. Ming-Omni-TTS (Inclusion AI et al., 2025), a speech-centric system derived from TTS, extends its scope to joint audio-music generation.

These increasingly capable systems motivate a benchmark capable of evaluating multiple control dimensions simultaneously.

3 Constructing MMAG

MMAG is designed to enable *compositional evaluation* of mixed audio generation, where speech, music, and sound effects must be generated coherently within a single acoustic scene. Rather than introducing another collection of audio-text pairs, our construction process is guided by three principles: (1) preserving realistic mixed-audio scenarios, (2) providing rich semantic and temporal annotations for fine-grained control, and (3) enabling multi-condition control through dedicated subsets. Figure 1 illustrates the overall construction process.

3.1 Data Sources and Filtering

Candidate clips are collected from test sets of AudioCaps, VGGSound, and MECAT, which provide complementary acoustic coverage for diverse speech, music, and environmental sound combinations while remaining comparable to widely used evaluation resources. We apply CED (Dinkel et al., 2024) to detect the presence of these modalities and retain only clips that contain cross-domain content. For clips containing speech, we further apply expert models to annotate speaker number and language, retaining only single-speaker English clips for reliable speech annotation. We additionally filter out samples with poor recording quality or audible distortion using production quality metric (Tjandra et al., 2025) to ensure audio fidelity.

3.2 Annotation Pipeline

Fine-grained expert annotation. For speech, expert models infer speaker gender, age, accent, emotion, and transcription. For music, we annotate instrumentation, genre, and mood. For sound

[2]UTMOS-v2 (Baba et al., 2024).

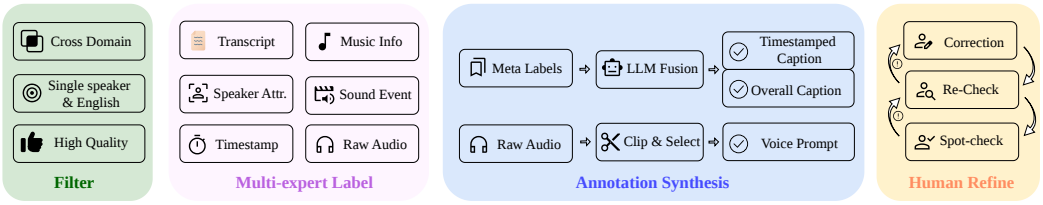


Figure 1: Overview of the MMAG annotation pipeline. Expert models extract fine-grained attributes from audio clips. An LLM aggregates these annotations into coherent overall and timestamped captions, while speech segments are further processed to construct voice prompts for voice cloning evaluation. Human inspection is performed to ensure annotation quality.

events, we identify foreground and background sound classes along with ambient acoustic context. In addition, we introduce a timestamping branch that provides precise temporal boundaries for all annotated elements across different domains.

Overall Caption. An LLM (DeepSeek-AI, 2025) aggregates annotations from multiple expert models and synthesizes them into a single coherent caption. The prompt instructs the model to faithfully retain expert-derived facts, reconcile contradictions, and avoid unsupported inferences. The LLM further leverages timestamp information to describe the temporal progression of audio elements, using phrases such as "at the beginning," "subsequently," or "throughout." This temporal structuring is particularly important in TTA tasks, where constructing a coherent narrative of the acoustic scene is essential for generation quality.

Timestamped Caption. In addition to the overall description, we generate a separate timestamped caption for each clip, primarily intended for fine-grained temporal control evaluation. This caption shares the same core semantic content as the overall one, but differs in precise timestamps for speech segments and foreground acoustic events. At the event level, for sustained events such as background music, we retain only coarse temporal descriptions such as "throughout the audio." At the clip level, we discard the timestamped caption entirely for any clip whose duration is almost entirely occupied by speech or background audio, since such samples offer limited scope for fine-grained temporal conditioning.

Voice Prompt. For audio samples from AudioCaps and VGGSound, we retrieve the original raw audio from YouTube. The longest speech segment in each target clip is first identified as the speaker reference. Separately, we apply voice activity detection (VAD) (Gao et al., 2023) to detect all speech segments in each raw recording. Only speech segments outside the target clip are retained to ensure no temporal overlap. These segments are then cropped or concatenated to a suitable duration, generating a set of candidate prompts. Finally, a speaker verification model (Wang et al., 2023) is used to compare each candidate with the reference, and the top five candidates are retained for manual inspection.

Human quality control. To correct hallucinations introduced by model-based annotation, we perform manual inspection to ensure annotation quality. This inspection focuses on two aspects: (1) the factual completeness and accuracy of the generated captions, and (2) the selection of final voice prompts from the candidate segments. The manual inspection consists of three rounds: initial correction, re-check, and final spot-check. The final spot-check pass rate reaches approximately 90%. Given the inherent ambiguity in certain audio categories and temporal boundaries, the pass rate provides reasonable assurance of annotation quality.

3.3 Benchmark Subsets

The final benchmark comprises three subsets: a main set for comprehensive evaluation, and two additional subsets targeting specific capabilities. An illustrative example is shown in Figure 2. The source-dataset composition and detailed distribution statistics of each subset are provided in Appendix C.

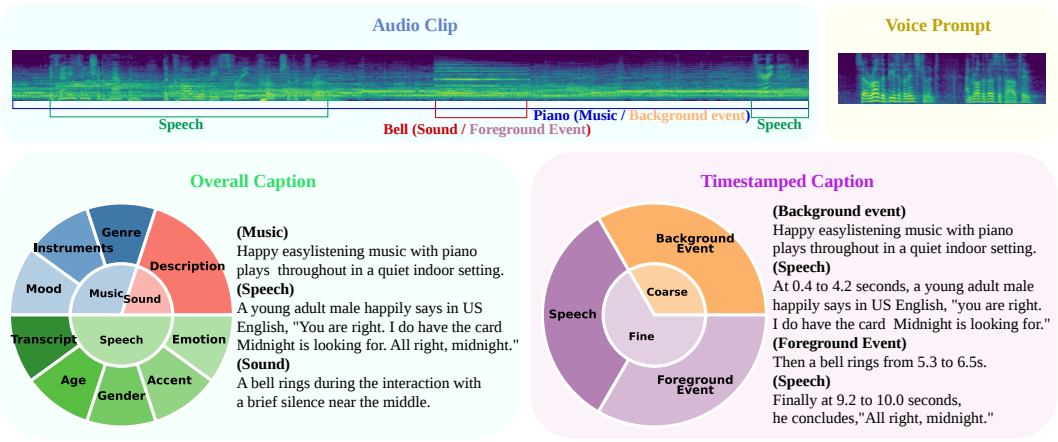


Figure 2: Illustration of an MMAG sample and its conditioning signals. The audio clip serves as the evaluation target, while the non-overlapping voice prompt provides the reference speaker identity. The overall caption describes speech, sound, music, and their relative temporal order, whereas the timestamped caption provides precise timestamps for speech segments and foreground events.

Main set. The main set contains 3,974 annotated audio clips with manually verified overall captions. It serves as the primary evaluation set for acoustic fidelity, speech quality, and semantic consistency.

Voice Cloning subset. The voice cloning subset contains 691 clips, each paired with a voice prompt and a detailed overall caption. Prompts are 2–5 seconds, matching the prompt length range in LibriSpeech-PC for zero-shot voice cloning. This subset enables separate evaluation of speaker similarity in addition to the main-set metrics.

Timestamp subset. The timestamp subset contains approximately 1,828 clips selected from the main set, each annotated with fine-grained temporal boundaries for speech segments and foreground acoustic events. This subset is specifically designed to evaluate temporal control in generation.

4 Evaluation Protocol

4.1 Evaluated Models

We evaluate models with their official checkpoints: unified audio-visual generation models (LTX-2.3, Ovi 1.1, MOVA-720p, JavisDiT++, UniAVGen), native mixed-audio generators (Dasheng AudioGen, Ming-Omni-TTS), and the agentic orchestrator AuDirector. Dasheng AudioGen supports structured input for fine-grained control with improved transcript stability. We evaluate both the unstructured and structured versions, denoted as Dasheng-AudioGen-Base and Dasheng-AudioGen-Fine. For Ming-Omni-TTS, we evaluate two variants: 0.5B and 16.8B-A3B (denoted as 16.8B for brevity).

We evaluate all models on the main set using their respective supported input formats. Models requiring a reference image, including MOVA and UniAVGen, receive a blank image for uniform conditioning. JavisDiT++ does not accept transcript input, so we evaluate it without transcripts. Ming-Omni-TTS requires structured text input, so we use an LLM to preprocess the caption into the required format.

On the voice cloning subset, we evaluate UniAVGen, Ming-Omni-TTS, and AuDirector. We adapt AuDirector to use the provided prompt audio as input instead of retrieving a speaker embedding from its internal timbre library.

On the timestamp subset, we evaluate LTX-2, Ovi, MOVA, UniAVGen, Dasheng AudioGen, and AuDirector, all of which support natural language input. JavisDiT++ is excluded from this subset due to its lack of transcript support, while the subset contains extensive speech timestamps. Ming-Omni-TTS is also omitted, as its structured input format is incompatible with timestamp control. AuDirector is also extended with timestamp-based processing and synthesis capabilities for this subset.

Detailed inference configurations and input-processing prompts for all evaluated models are provided in Appendices E and F, respectively.

4.2 Automatic Metrics

Acoustic Fidelity. We evaluate acoustic fidelity from two complementary perspectives: distributional similarity and perceptual quality. For distributional similarity, we report Fréchet Distance (FD) in the VGGish feature space (Kilgour et al., 2018b), along with KL divergence and Inception Score (IS) computed using a pretrained audio classifier on AudioSet (Gemmeke et al., 2017) labels. For completeness, we additionally provide FD scores computed with PANNs CNN14 (Kong et al., 2020) features in Appendix D, as the two FD variants exhibit similar trends. For perceptual quality, we include Audiobox Aesthetics (Tjandra et al., 2025), a suite of no-reference predictors. We report production quality (PQ) and production complexity (PC) as primary metrics, since audio quality and content complexity are core dimensions for mixed audio tasks. Content enjoyment (CE) and content usefulness (CU) are deferred to the appendix D, as they are more subjective and application-dependent.

Speech Quality. For clips containing speech, WER is computed from Whisper-Large-v3 (Radford et al., 2023) transcriptions and capped at 1.0 for robustness. For the voice cloning subset, we extract speaker embeddings using WeSpeaker-ResNet-221 (Wang et al., 2023) and compute SPK-SIM. UTMOS_{v2} (Baba et al., 2024) estimates speech naturalness.

Semantic Consistency. CLAP score (Wu et al., 2023) assesses semantic consistency at the global level by measuring the cosine similarity between the input caption and the generated audio in a contrastive embedding space. To further assess semantic consistency at a fine-grained level, we follow the official AnyAudio-Judge implementation (Li et al., 2026). Qwen3-30B-A3B-Instruct-2507 is used (Yang et al., 2025) to decompose each caption into a set of atomic rubric items, each targeting a single dimension such as speaker age or music genre. AnyAudio-Judge-30B then scores the generated audio against each rubric independently. For each rubric, the model produces a compliance probability derived from the softmax-normalized logits of the "yes" vs. "no" response. The final sample score is obtained by averaging all rubric-level scores, denoted as AAJ.

Temporal Control. For the timestamp subset, we compute a frame-based F1 (Frame-F1) to evaluate temporal alignment. The ground-truth timestamps for speech segments and foreground acoustic events are extracted from timestamped captions using Qwen3-30B-A3B-Instruct-2507. For predictions, we apply WhisperX (Bain et al., 2023) to extract speech timestamps and PE-A-Frame-Large (Vyas et al., 2025) to extract foreground event timestamps from the generated audio. We discretize each clip into 40-ms frames and label a frame as positive when at least half of its duration overlaps a corresponding segment. Speech is treated as a single activity class, whereas sound events additionally require category matching. Precision, recall, and Frame-F1 are computed for each clip as

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \qquad R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \qquad \text{Frame-F1} = \frac{2PR}{P + R}.$$

We macro-average the clip-level scores and apply the same procedure separately to speech and foreground sound segments, reporting $\mathrm{Speech}_{\mathrm{F1}}$ and $\mathrm{Sound}_{\mathrm{F1}}$.

5 Results and Analysis

We report results on the main set, voice cloning subset, and timestamp subset in Tables 2, 3, and 4. For all tables, arrows (↑ / ↓) indicate the direction of better performance, and best results per column are bolded and underlined.

5.1 Main Set Results

Table 2 compares all evaluated models on the main set. The most salient observation is that no model dominates across different metrics simultaneously.

Table 2: Results on the main set.

Category	Model	Acoustic					Speech		Semantic	
		FD↓	KL↓	IS↑	PC↑	PQ↑	WER↓	UTMOS_{v2}↑	CLAP↑	AAJ↑
Agentic	AuDirector	5.19	1.60	2.62	3.69	6.71	0.08	2.68	**0.34**	0.76
Audio-Visual	LTX-2	2.89	1.31	2.90	3.04	7.22	0.04	2.90	**0.34**	0.76
	Ovi	3.91	1.50	2.06	2.60	6.05	0.13	2.25	0.30	0.74
	MOVA	4.38	1.28	2.59	2.98	6.74	0.15	2.69	0.31	0.78
	JavisDiT++	5.58	1.47	**3.49**	**3.84**	5.98	–	1.64	0.30	0.59
	UniAVGen	10.19	2.28	1.44	2.35	6.75	0.10	3.00	0.25	0.64
Mixed-Audio	Dasheng-AudioGen-Base	**1.78**	**0.87**	3.17	3.55	6.67	0.11	2.77	**0.34**	**0.80**
	Dasheng-AudioGen-Fine	1.79	0.89	3.11	3.64	6.79	0.11	2.76	0.33	**0.80**
	Ming-Omni-TTS-0.5B	8.72	2.80	1.56	1.60	**7.35**	**0.03**	**3.43**	0.26	0.65
	Ming-Omni-TTS-16.8B	7.98	2.46	1.65	1.79	7.14	**0.03**	3.33	0.27	0.67

Dasheng AudioGen obtains the strongest distributional audio scores. Dasheng-AudioGen-Base achieves the best FD (1.78) and KL (0.87) among the evaluated systems, while Dasheng-AudioGen-Fine delivers nearly identical FD/KL performance with slightly higher perceptual quality scores. This result confirms Dasheng AudioGen's superior performance in terms of acoustic fidelity among the evaluated models. Its high CLAP and AAJ scores further demonstrate strong semantic alignment and fine-grained consistency with the input captions. However, its speech metrics are not dominant: both variants obtain WER = 0.11, lagging behind the leading models on speech intelligibility. Ming-Omni-TTS shows the opposite trend. Both the 0.5B and 16.8B variants achieve the best WER (0.03) on the main set, and the 0.5B model obtains the highest UTMOS_{v2} and PQ among all models. This confirms the advantage of speech-specialized generation for intelligibility and perceived speech naturalness. However, Ming-Omni-TTS shows a systematic disadvantage on distribution indicators, PC and semantic metrics, ranking well below the average. As a speech-centric system, it handles only a limited set of label-formatted sound events. However, the diversity and richness of non-speech content in MMAG exceed its capability. Consequently, it frequently omits music and sound effects, which degrades acoustic and semantic scores while inflating speech-related metrics.

Among audio-visual generation models, LTX-2 delivers the most balanced performance on the main set. It achieves strong FD, competitive KL and IS, a tied-best CLAP score, and nearly the lowest WER. Its PQ is also high, second only to Ming-Omni-TTS-0.5B. This suggests that unified audio-video models can transfer useful scene-generation abilities to audio-only evaluation even when visual input is held constant. MOVA and Ovi show moderate performance across most metrics, while UniAVGen underperforms acoustic fidelity despite achieving reasonable UTMOS_{v2}, largely because it omits most music and sound events in its generations. JavisDiT++ achieves the highest IS and PC, implying diverse and acoustically complex outputs, but its speech is largely unintelligible, resulting in limited performance on speech quality metrics. At fine-grained semantic-consistency level, AAJ scores align with this trend. MOVA, LTX-2, and Ovi achieve similar scores (0.74-0.78). UniAVGen (0.64) and JavisDiT++ (0.59) score considerably lower, reflecting their respective weaknesses in sound/music generation and speech intelligibility.

The agentic orchestrator AuDirector ties for the best CLAP, along with competitive PC and AAJ scores, yet it lags behind the leading systems on other metrics. Although AuDirector can generate nearly all constituent audio elements, the final synthesis suffers from poor volume equalization and abrupt transitions between acoustic events. This results in perceptually disjointed outputs that differ from real acoustic distributions and consequently degrade both distribution and speech metrics.

5.2 Voice Cloning Subset Results

Table 3 compares each model with and without voice-prompt conditioning on the same 691 samples.

Introducing the additional voice-prompt condition degrades several metrics. For UniAVGen and both Ming-Omni-TTS variants, WER increases while PQ and UTMOS_{v2} decrease. AuDirector maintains the same WER of 0.10 because its dedicated TTS module handles the voice prompt separately, but its PQ and UTMOS_{v2} also decrease. At the same time, voice prompting provides more precise speaker-identity condition. Accordingly, SPK-SIM improves for every model, while acoustic metrics other than PQ and semantic metrics generally improve.

Table 3: Controlled comparison on the same 691-sample voice cloning subset. *w/o* and *w/* denote generation without and with voice-prompt conditioning, respectively. A superscript ⋆ marks the better result within each model pair; ties at the reported precision are unmarked. Ming: Ming-Omni-TTS.

Category	Model	Prompt	Acoustic					Speech			Semantic	
			FD↓	KL↓	IS↑	PC↑	PQ↑	WER↓	SPK-SIM↑	UTMOS_{v2}↑	CLAP↑	AAJ↑
Agentic	AuDirector	w/o	6.36	2.32	2.69⋆	3.44	6.03⋆	0.10	0.53	2.68⋆	0.34	0.75
		w/	**3.80**⋆	**2.11**⋆	**2.61**	**3.54**⋆	5.56	0.10	0.72⋆	2.32	**0.35**⋆	**0.75**
Audio-Visual	UniAVGen	w/o	13.52	3.10	1.33	2.67	6.68⋆	0.13⋆	0.53	3.00⋆	0.21	0.64
		w/	6.70⋆	2.58⋆	1.58⋆	3.28⋆	5.84	0.28	0.67⋆	2.40	0.26⋆	0.70⋆
Mixed-Audio	Ming-0.5B	w/o	15.14	3.69	1.24	1.47	7.49⋆	0.04⋆	0.55	3.43⋆	0.21	0.63
		w/	8.72⋆	3.32⋆	1.26⋆	1.81⋆	**6.54**	0.06	0.76⋆	**2.85**	0.26⋆	0.66⋆
	Ming-16.8B	w/o	14.39	3.45	1.31	1.60	7.23⋆	0.03⋆	0.54	3.33⋆	0.23	0.64
		w/	7.39⋆	2.99⋆	1.38⋆	2.12⋆	6.43	**0.04**	**0.78**⋆	2.79	0.27⋆	0.68⋆

These results indicate that speaker voice conditioning introduces a distinct control challenge: the model must not only generate the requested acoustic content, but also preserve speaker identity from a short prompt and maintain intelligibility. SPK-SIM confirms that models extract useful identity information from the prompt: Ming-Omni-TTS-16.8B achieves the highest SPK-SIM of 0.78, followed by Ming-Omni-TTS-0.5B at 0.76 and AuDirector at 0.72. However, these high similarity scores do not guarantee uniformly better audio quality. The model with the best SPK-SIM does not achieve the best performance on acoustic and semantic indicators. This highlights a trade-off between speaker identity preservation and full-scene generation quality.

Overall, AuDirector achieves strong performance on semantic metrics and PC, highlighting the advantage in content planning and semantic preservation. Its strong SPK-SIM also demonstrates good generalizability to voice cloning. However, its most notable weakness lies in audio quality, as indicated by the lowest UTMOS_{v2} and PQ among all models. UniAVGen exhibits the most pronounced degradation on this subset, suggesting that balancing and integrating multimodal information remains a challenge for unified audio-visual generation models. Ming-Omni-TTS continues to show strong speech capability alongside poor sound event generation, mirroring its main-set performance.

5.3 Timestamp Subset Results

Table 4: Results on the timestamp subset. DSAG: Dasheng-AudioGen.

Category	Model	Acoustic					Speech		Semantic		Temporal	
		FD↓	KL↓	IS↑	PC↑	PQ↑	WER↓	UTMOS_{v2}↑	CLAP↑	AAJ↑	$\text{Speech}_{\text{F1}}$↑	Sound_{F1}↑
Agentic	AuDirector	6.59	2.01	3.08	**4.16**	6.19	0.15	2.31	**0.35**	0.78	**0.72**	**0.57**
Audio-Visual	LTX-2	2.38	1.46	3.35	3.35	**6.97**	**0.05**	2.74	0.33	0.77	0.64	0.44
	Ovi	3.99	1.67	2.23	2.74	6.05	0.19	2.12	0.27	0.78	0.65	0.39
	MOVA	4.36	1.82	2.91	2.85	6.69	0.14	2.47	0.29	0.77	0.63	0.42
	UniAVGen	11.18	2.52	1.39	2.66	6.56	0.15	**2.90**	0.23	0.67	0.63	0.29
Mixed-Audio	DSAG-Base	1.96	1.10	3.67	3.47	3.48	0.43	2.68	0.31	0.78	0.60	0.41
	DSAG-Fine	**1.77**	**1.09**	**3.76**	3.49	6.64	0.13	2.65	0.31	**0.80**	0.62	0.42

Table 4 evaluates fine-grained temporal control. Current models struggle to follow timestamped instructions for both speech and acoustic events. Even the best-performing system, AuDirector, reaches $\text{Speech}_{\text{F1}} = 0.72$ and $\text{Sound}_{\text{F1}} = 0.57$, leaving a substantial gap to reliable temporal-controllable synthesis. Most unified audio-visual models and Dasheng AudioGen show near-random timestamp controllability, while trends in the other metrics mirror the main set.

Among existing approaches, AuDirector achieves the highest temporal scores, benefiting from its specific adaptation for timestamp input and time-aware synthesis. However, this temporal controllability comes at a substantial fidelity cost, as AuDirector has greater degradation on FD and UTMOS_{v2} relative to its main-set performance than other models. In addition, the event detectors used for evaluation often fail to locate precise temporal boundaries in AuDirector's outputs, lowering both

Table 5: Caption-confusion proportion (%) on the main and timestamp sets. DSAG: Dasheng-AudioGen; Ming: Ming-Omni-TTS.

Agentic / Audio-Visual			Mixed-Audio		
Model	**Main**	**Timestamp**	**Model**	**Main**	**Timestamp**
AuDirector	**1.88**	**5.43**	DSAG-Base	18.21	56.48
LTX-2	5.12	7.41	DSAG-Fine	21.88	36.04
Ovi	13.92	19.70	Ming-0.5B	3.74	–
MOVA	4.59	7.00	Ming-16.8B	2.80	–
UniAVGen	3.75	6.63			

$\text{Speech}_{\text{F1}}$ and Sound_{F1}. This indicates that more sophisticated synthesis strategies are needed to reduce the fidelity cost of temporal conditioning.

Beyond AuDirector, most other models also exhibit performance degradation across acoustic and speech metrics on the timestamp subset. The decrease in indicators can be attributed to two factors. First, the timestamp input may fall outside the training distribution of existing models. Second, the timestamp subset itself contains more complex and acoustically diverse audio content.

Beyond model performance, the automatic metrics exhibit a limitation on this subset. Specifically, AAJ scores exhibit minimal dispersion across models. The rubrics include timestamp-aware items, but the judge model exhibits limited discriminative ability on them, leading to undifferentiated scores. This limitation points toward the need for more temporally aware judge models in future work.

5.4 Error Patterns

We observe that some models exhibit caption confusion: they tend to produce spoken content drawn from the caption but outside the target transcript, even with correct input formatting. To quantify this phenomenon, we design a diagnostic metric **Proportion** in Appendix D, defined as the contribution of such extraneous content to the final WER. A higher Proportion indicates more severe leakage, independent of the absolute WER level.

The results in Table 5 show that this issue is particularly pronounced in Ovi and Dasheng AudioGen. On the main set, roughly 10%-22% of word errors are attributable to caption confusion. On the timestamp subset, this problem is further amplified. Dasheng-AudioGen-Base exhibits a Proportion of 56.48%, indicating a strong correlation between increased WER and caption confusion. After adopting structured input, Dasheng-AudioGen-Fine shows substantial reductions in both WER and Proportion, though its Proportion remains higher than that of other models.

Compared to the main set, all models show increased Proportion on the timestamp subset. This suggests that the introduction of timestamps raises the difficulty of recognizing and generating accurate transcripts, highlighting the need for more diverse training data. Dasheng-AudioGen-Fine benefits from LLM-based input processing, which significantly reduces the confusion, and AuDirector also attains a strong Proportion score via LLM-driven input processing and expert model invocation. Leveraging LLMs to normalize raw inputs into a form better suited for models represents an effective strategy to improve model robustness.

6 Conclusion

To address the lack of comprehensive evaluation for mixed audio generation, we propose MMAG. It comprises a main set of approximately 4,000 clips, a timestamp subset of roughly 1,800 clips, and a voice cloning subset of roughly 690 clips. MMAG is curated through a multi-stage pipeline that integrates expert models, LLM-based caption synthesis, and human quality control. We have established a comprehensive evaluation framework and systematically benchmarked existing models. Our findings reveal that no existing model achieves balanced performance across acoustic fidelity, speech quality, semantic consistency, and temporal control. Instead, we observe systematic trade-offs between generation quality and controllability, calling for more robust and controllable generation.

AI USE STATEMENT

We used models from the Qwen3 family, Gemini, and DeepSeek-R1 to assist with data annotation and structured information extraction. GPT-5.6 Sol was used to assist with the implementation of the Frame-F1 metric and language polishing of the paper. All AI-assisted annotations, code, and text were reviewed and verified by the authors, who take full responsibility for the final content.

ETHICS STATEMENT

We have read and adhere to the ICLR Code of Ethics. We hired external annotators for manual quality control of generated captions and voice prompt selection. All annotators were compensated at a rate above local minimum wage. Annotators received detailed instructions and were informed of their right to withdraw at any time. No personally identifiable information was collected from annotators, and the annotation task did not involve sensitive content. All audio data used in this benchmark are sourced from existing publicly available datasets (AudioCaps, VGGSound, MECAT) and are used for research purposes in accordance with their original licenses. We do not foresee direct negative societal impacts from this work; the benchmark is intended to facilitate transparent and reproducible evaluation of audio generation models. No external sponsorship or conflict of interest influenced the design or conclusions of this work.

REPRODUCIBILITY STATEMENT

We follow the ICLR reproducibility guidelines. All benchmark data have been publicly released. Most evaluation metrics are established metrics commonly used in text-to-audio and text-to-speech research, while the implementation of Frame-F1 is fully specified in the paper. The evaluated models largely follow their official implementations. Appendix E provides the checkpoints, inference configurations, and output-processing details, and Appendix F reports the prompts used for model-specific input processing. We will release the complete evaluation pipeline to facilitate reproducible and category-wise evaluation.

REFERENCES

Kaito Baba, Wataru Nakata, Yuki Saito, and Hiroshi Saruwatari. The t05 system for the voicemos challenge 2024: Transfer learning from deep image classifier to naturalness mos prediction of high-quality synthetic speech. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pp. 818–824. IEEE, 2024.

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. WhisperX: Time-accurate speech transcription of long-form audio. *arXiv preprint arXiv:2303.00747*, 2023.

Felix Burkhardt, Johannes Wagner, Hagen Wierstorf, Florian Eyben, and Björn Schuller. Speech-based age and gender prediction with transformers. In *Speech Communication; 15th ITG Conference*, pp. 46–50. VDE, 2023.

Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. VGGSound: A large-scale audio-visual dataset. In *Proceedings of ICASSP*, 2020.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. Simple and controllable music generation. *Advances in neural information processing systems*, 36:47704–47720, 2023.

DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Heinrich Dinkel, Yongqing Wang, Zhiyong Yan, Junbo Zhang, and Yujun Wang. Ced: Consistent ensemble distillation for audio tagging. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 291–295. IEEE, 2024.

Zhifu Gao, Zerui Li, Jiaming Wang, Haoneng Luo, Xian Shi, Mengzhe Chen, Yabin Li, Lingyun Zuo, Zhihao Du, Zhangyu Xiao, et al. Funasr: A fundamental end-to-end speech recognition toolkit. *arXiv preprint arXiv:2305.11013*, 2023.

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

Sreyan Ghosh, Arushi Goel, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang-gil Lee, Chao-Han Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, et al. Audio flamingo 3: Advancing audio intelligence with fully open large audio language models. *Advances in Neural Information Processing Systems*, 38:41819–41886, 2026.

Yoav HaCohen, Benny Brazowski, Nisan Chiprut, Yaki Bitterman, Andrew Kvochko, and Avishai Berkowitz. LTX-2: Efficient joint audio-visual foundation model. *arXiv preprint arXiv:2601.03233*, 2026.

Chia-Yu Hung, Navonil Majumder, Zhifeng Kong, Ambuj Mehrish, Amir Zadeh, Chuan Li, Rafael Valle, Bryan Catanzaro, and Soujanya Poria. Tangoflux: Super fast and faithful text to audio generation with flow matching and clap-ranked preference optimization. In *International Conference on Learning Representations*, volume 2026, pp. 150793–150816, 2026.

Inclusion AI, Biao Gong, Cheng Zou, Chuanyang Zheng, Chunluan Zhou, and Canxiang Yan. Ming-Omni: A unified multimodal model for perception and generation. *arXiv preprint arXiv:2506.09344*, 2025.

Jaeyong Kang and Dorien Herremans. Towards unified music emotion recognition across dimensional and categorical models. *arXiv preprint arXiv:2502.03979*, 2025.

Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. Fréchet audio distance: A metric for evaluating music enhancement algorithms. *arXiv preprint arXiv:1812.08466*, 2018a.

Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. Fr\'echet audio distance: A metric for evaluating music enhancement algorithms. *arXiv preprint arXiv:1812.08466*, 2018b.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. AudioCaps: Generating captions for audios in the wild. In *Proceedings of NAACL-HLT*, 2019.

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

Sangho Lee, Jiwan Chung, Youngjae Yu, Gunhee Kim, Thomas Breuel, Gal Chechik, and Yale Song. Acav100m: Automatic curation of large-scale datasets for audio-visual video representation learning. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10254–10264. IEEE, 2021.

Haitao Li, Tian Tan, Yuguang Yang, Shan Yang, and Xie Chen. Anyaudio-judge: A dynamic rubric-based benchmark and evaluator for audio instruction following. *arXiv preprint arXiv:2606.03116*, 2026.

Yizhi Li, Ruibin Yuan, Ge Zhang, Yinghao Ma, Xingran Chen, Hanzhi Yin, Chenghao Xiao, Chenghua Lin, Anton Ragni, Emmanouil Benetos, et al. Mert: Acoustic music understanding model with large-scale self-supervised training. In *International Conference on Learning Representations*, volume 2024, pp. 12181–12204, 2024.

Kai Liu, Yanhao Zheng, Kai Wang, Shengqiong Wu, Rongjunchen Zhang, Jiebo Luo, Dimitrios Hatzinakos, Ziwei Liu, Hao Scofield Fei, and Tat-Seng Chua. Javisdit++: Unified modeling and optimization for joint audio-video generation. In *International Conference on Learning Representations*, volume 2026, pp. 150592–150618, 2026.

Chetwin Low, Weimin Wang, and Calder Katyal. Ovi: Twin backbone cross-modal fusion for audio-video generation. *arXiv preprint arXiv:2510.01284*, 2025.

Ziyang Ma, Zhisheng Zheng, Jiaxin Ye, Jinchao Li, Zhifu Gao, Shiliang Zhang, and Xie Chen. emotion2vec: Self-supervised pre-training for speech emotion representation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 15747–15760, 2024.

Jiahao Mei, Heinrich Dinkel, Yadong Niu, Xingwei Sun, Gang Li, Yifan Liao, Jiahao Zhou, Junbo Zhang, Jian Luan, and Mengyue Wu. Dasheng AudioGen: A unified model for generating coherent audio scenes from text. *arXiv preprint arXiv:2605.27838*, 2026.

Aleksandr Meister, Matvei Novikov, Nikolay Karpov, Evelina Bakhturina, Vitaly Lavrukhin, and Boris Ginsburg. Librispeech-pc: Benchmark for evaluation of punctuation and capitalization capabilities of end-to-end asr models. In *2023 IEEE automatic speech recognition and understanding workshop (ASRU)*, pp. 1–7. IEEE, 2023.

Jan Melechovsky, Zixun Guo, Deepanway Ghosal, Navonil Majumder, Dorien Herremans, and Soujanya Poria. Mustango: Toward controllable text-to-music generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 8293–8316, 2024.

Yadong Niu, Tianzi Wang, Heinrich Dinkel, Xingwei Sun, Jiahao Zhou, Gang Li, Jizhong Liu, Xunying Liu, Junbo Zhang, and Jian Luan. MECAT: A multi-experts constructed benchmark for fine-grained audio understanding tasks. *arXiv preprint arXiv:2507.23511*, 2026.

Alexis Plaquet and Hervé Bredin. Powerset multi-class cross entropy loss for neural speaker diarization. *arXiv preprint arXiv:2310.13025*, 2023.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pp. 28492–28518. PMLR, 2023.

Mirco Ravanelli, Titouan Parcollet, Adel Moumen, Sylvain De Langen, Cem Subakan, Peter Plantinga, Yingzhi Wang, Pooneh Mousavi, Luca Della Libera, Artem Ploujnikov, et al. Open-source conversational ai with speechbrain 1.0. *Journal of Machine Learning Research*, 25(333):1–11, 2024.

Yiming Ren, Xuenan Xu, Ziyang Zhang, Wen Wu, Baoxiang Li, and Chao Zhang. Audirector: A self-reflective closed-loop framework for immersive audio storytelling. *arXiv preprint arXiv:2605.11866*, 2026.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training GANs. In *Advances in Neural Information Processing Systems*, 2016.

SII-OpenMOSS Team, Donghua Yu, Mingshu Chen, Qi Chen, Qi Luo, and Qianyi Wu. MOVA: Towards scalable and synchronized video-audio generation. *arXiv preprint arXiv:2602.08794*, 2026.

Andros Tjandra, Yi-Chiao Wu, Baishan Guo, John Hoffman, Brian Ellis, Apoorv Vyas, Bowen Shi, Sanyuan Chen, Matt Le, Nick Zacharov, Carleigh Wood, Ann Lee, and Wei-Ning Hsu. Meta Audiobox aesthetics: Unified automatic quality assessment for speech, music, and sound. *arXiv preprint arXiv:2502.05139*, 2025.

Apoorv Vyas, Heng-Jui Chang, Cheng-Fu Yang, Po-Yao Huang, Luya Gao, Julius Richter, Sanyuan Chen, Matt Le, Piotr Dollár, Christoph Feichtenhofer, et al. Pushing the frontier of audiovisual perception with large-scale multimodal correspondence learning. *arXiv preprint arXiv:2512.19687*, 2025.

Hongji Wang, Chengdong Liang, Shuai Wang, Zhengyang Chen, Binbin Zhang, Xu Xiang, Yanlei Deng, and Yanmin Qian. Wespeaker: A research and production oriented speaker embedding learning toolkit. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *Proceedings of ICASSP*, 2023.

LCT Xiaomi. Mimo-audio: Audio language models are few-shot learners. *CoRR, abs/2512.23808*, 2025.

Zeyu Xie, Xuenan Xu, Zhizheng Wu, and Mengyue Wu. Picoaudio: Enabling precise temporal controllability in text-to-audio generation. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Guozhen Zhang, Zixiang Zhou, Teng Hu, Ziqiao Peng, Youliang Zhang, Yi Chen, Yuan Zhou, Qinglin Lu, and Limin Wang. UniAVGen: Unified audio and video generation with asymmetric cross-modal interactions. *arXiv preprint arXiv:2511.03334*, 2025.

Siyi Zhou, Yiquan Zhou, Yi He, Xun Zhou, Jinchao Wang, Wei Deng, and Jingchen Shu. Indextts2: A breakthrough in emotionally expressive and duration-controlled auto-regressive zero-shot text-to-speech. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 35139–35148, 2026.

Juan Zuluaga-Gomez, Sara Ahmed, Danielius Visockas, and Cem Subakan. CommonAccent: Exploring Large Acoustic Pretrained Models for Accent Classification Based on Common Voice. In *Interspeech 2023*, pp. 5291–5295, 2023. doi: 10.21437/Interspeech.2023-2419.

A Implementation Details

To obtain high-quality annotations for each audio clip, we employ specialized expert models for each attribute dimension. Table 6 summarizes the complete set of models used in our annotation pipeline.

For speaker number estimation, we adopt a conservative strategy by leveraging two complementary models: Pyannote-SD 3.1 and Audioflamingo3. We take the minimum of the two predictions as the final speaker number, as preliminary experiments revealed that both models tend to overestimate the number of speakers.

Table 6: Attribute and Corresponding Models for Each Category

Category	Attribute	Model
speech	speaker number	Pyannote-SD 3.1 (Plaquet & Bredin, 2023) Audioflamingo3 (Ghosh et al., 2026)
	language	Speechbrain-ECAPA (Ravanelli et al., 2024)
	age&gender	Audeering-AGR (Burkhardt et al., 2023)
	accent	CommonAccent (Zuluaga-Gomez et al., 2023)
	emotion	Emotion2Vec (Ma et al., 2024)
	transcript	Whisper-Large-V3 (Radford et al., 2023)
	timestamp	WhisperX (Bain et al., 2023)
sound	event	CED (Dinkel et al., 2024) Gemini-2.5-pro (Comanici et al., 2025)
	timestamp	PE-A-Frame-Large (Vyas et al., 2025)
music	instrument&genre	Musical Descriptor (Li et al., 2024)
	mood	Music2Emo (Kang & Herremans, 2025)

B LLM Prompt for Caption

To generate the descriptive captions for our benchmark, we design the following instruction prompt for the LLM. The prompt enforces strict constraints on factual consistency, listening-only perspective, and structured hierarchical output. To improve information completeness, the model is required to first generate a content-focused caption for each category. The complete prompt is presented below.

```
You are an expert AI audio analyst. Your task is to synthesize captions using the given
    "audio information" in JSONL format.

========================================
I. GENERAL RULES
========================================
a) Carefully identify conflicting information between fields and avoid mentioning
    conflicting aspects in the final caption. Focus only on consistent and unopposed
    information. Do not invent details not present in the data.

b) Do not use parentheses to provide detailed explanation in any output (e.g., avoid "
    Middle age (30-40 years old)").

c) All answers must be created from the perspective of someone who ONLY LISTENED to the
    audio, without any technical/model references or quantitative metrics (e.g. MOS).

========================================
II. CAPTION DEVELOPMENT FRAMEWORK
========================================

A. Content-Focused Caption
----------------------------------------
Generate a detailed description that focuses on individual content dimensions. Not all
    categories below may be present; if a field is missing, leave it blank or mark as
    unavailable. Do not hallucinate.
```

```
1) Speech:
   - Incorporate transcription content and speaker attributes (age, gender, language,
     accent, emotion, etc.).
   - Infer speaker identity only when sufficient evidence exists.
   - The complete transcription must be included verbatim, enclosed in quotation marks,
     without omission or summarization.
   - Example: "A middle-aged male captain aged 40-50 calmly stated in standard English,
     'The plane is about to land.'"

2) Music:
   - Synthesize music descriptions based on the provided information (instrument, genre,
      moods).
   - Key, instrument, and genre are given as label-probability pairs. Integrate all
     relevant information to make informed judgments.
   - The existing music labels have limited credibility and should only serve as
     auxiliary references for resolving ambiguous cases.
   - Example: "Slow-tempo electronic music with dark melodic elements and synthesized
     bass."

3) Sound:
   - Generate a complete description of sound events based on the "event_segments" field.

   - Sound events exclude content from major categories such as speech or music (e.g., "
     man talking", "guitar").
   - Include coarse-grained temporal information, such as event sequences and
     approximate ranges, but avoid precise timestamps.
   - Example: "After a barking dog, the sound of a baby crying could be heard, with wind
      mixed in the background."

4) Environment:
   - Infer environmental context from the existing speech, sound, and music information,
      and provide a refined environment caption when evidence is sufficient.
   - Example: "Indoor/outdoor hybrid environment with significant engine interference."

B. Systematic Caption
----------------------------------------
Generate holistic captions that integrate all content dimensions.

1) Overall Caption (several sentences):
   - Include as much information as possible, covering speech, music, sound, environment
     , and other relevant dimensions.
   - If transcription exists, include the complete transcription verbatim, enclosed in
     quotation marks. The speech content may be summarized at the beginning, but the
     complete transcription must appear subsequently.
   - Example: "A 20-30 young male speaker delivers an emotionally charged English
     monologue expressing pride, 'I love my country deeply'. Accompanied by dark-toned
     electronic music with steady percussion. Vehicle engine noise persists throughout
     the recording."

2) Timestamped Caption (several sentences):
   - Include as much information as possible, covering speech, music, sound, environment
     , and other relevant dimensions.
   - If transcription exists, include the complete transcription verbatim, enclosed in
     quotation marks.
   - Provide specific timestamps for each sentence and foreground event by integrating
     transcription-segments and event_segments data. Identify all foreground events,
     merging those with similar categories or highly overlapping time intervals.
   - For background sound events, do not provide precise timestamps; instead, describe
     their occurrence using coarse-grained terms such as "start", "end", "middle", or "
     entire process".
   - Example: "In 0.2-1.5s, a 20-30 young female speaker delivers an emotionally charged
      English monologue expressing pride, 'I love my country deeply'. Then applause
     erupted between 2.1-5.9 seconds. In 7.5-9.4s, the speaker said, 'my speech is over,
```

```
    thank you all.' There is dark electronic music and stable percussion in the
    background."
```

C DATASET DETAILS

Table 7 summarizes the source-dataset composition of each MMAG evaluation split.

Table 7: Source-dataset composition of the MMAG evaluation splits.

Source dataset	Main	Voice cloning	Timestamp
AudioCaps	114	72	69
VGGSound	1,022	619	695
MECAT	2,838	–	1,064
Total	**3,974**	**691**	**1,828**

The differences in source counts primarily reflect the sizes and content distributions of the original test sets rather than deliberate balancing by source dataset. AudioCaps has a comparatively small test set and therefore contributes fewer eligible examples after filtering. In contrast, MECAT contains a larger proportion of mixed-audio recordings that satisfy the MMAG inclusion criteria, resulting in its greater contribution to the main and timestamp sets. All data are utilized exclusively as audio, with their textual annotations excluded.

We perform a comprehensive statistical analysis of the benchmark to characterize its composition and coverage. Figure 3 presents the label distributions on the main set across multiple dimensions, including gender, age group, emotion type, and accent category. For the timestamp subset, we report the proportion of samples with precise temporal annotations, covering both speech segments and foreground acoustic events.

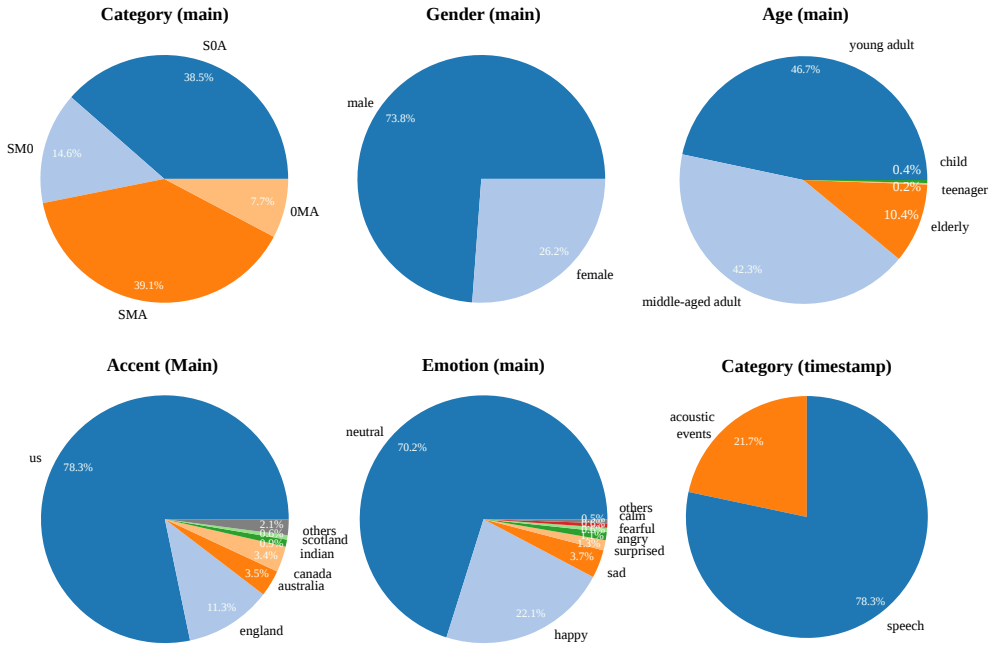


Figure 3: Label distribution of MMAG

To examine musical diversity, we extract instrument information from the generated captions using an LLM. Figure 4 presents the per-clip instrument count distribution, revealing the varying instrumental

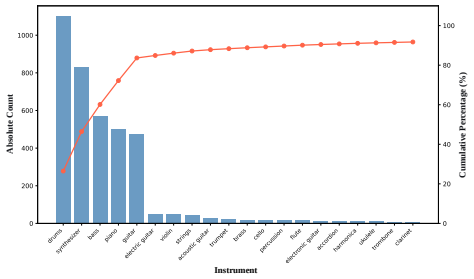
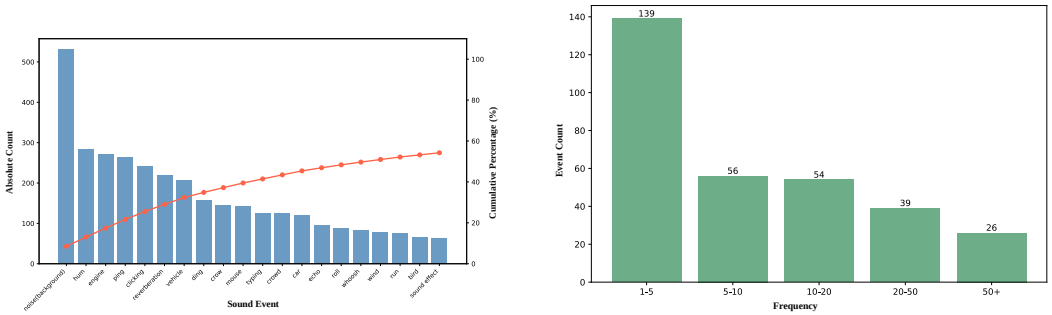


Figure 4: Instrument distribution of MMAG



(a) Sound event distribution

(b) Sound event frequency

Figure 5: Sound-event statistics of MMAG: (a) event-category distribution and (b) event-frequency distribution.

complexity across samples. For sound events, we extract event types from the captions via the same LLM, yielding roughly 320 distinct categories after excluding speech- and music-related events. In comparison, after removing 'speech,' 'music,' and their descendants from the AudioSet ontology, roughly 360 categories remain. The comparable category coverage between MMAG and AudioSet demonstrates that our benchmark achieves broad coverage of diverse sound events.

However, as shown in Figure 5, the distribution of sound events in MMAG is not perfectly balanced. This skew largely reflects the natural prevalence of certain event types in real-world recordings, inherited from the underlying data sources. While this imbalance presents challenges for evaluating performance on rare categories, it also more faithfully mirrors the long-tail distribution characteristic of real acoustic environments.

For the voice cloning subset, we compute the SPK-SIM between each voice prompt and its corresponding target audio sample, achieving an average score of approximately 0.82. By comparison, the reference score computed on the LibriSpeech-PC dataset is approximately 0.86.

The slightly lower performance on our subset can be attributed to the presence of background acoustic content in some of the extracted voice prompts, which inevitably interferes with speaker embedding extraction and reduces the similarity score. However, such background-accompanied prompts are inherently reasonable in the context of mixed audio generation, where the model is expected to extract and preserve the target speaker's timbre not only from perfectly clean speech, but also from realistic, non-isolated audio segments. From the SPK-SIM evaluation perspective, WeSpeaker remains generally robust to background speech and environmental noise. While not perfect in complex acoustic scenes, the metric still provides a meaningful measure of speaker similarity.

D Additional Results

In addition to the primary metrics, we report supplementary results for three auxiliary metrics (FD_{CNN14}, CE, and CU) in the supplementary Tables 8 for each of the three benchmark subsets.

Table 8: Comparison across three benchmark subsets. Best results within each subset are highlighted in **bold** and underlined.

Category	Model	Aux. Metric			WER Any.			
		FD_{CNN14}↓	CE↑	CU↑	Hit Rate (%)↓	WER_{caps} (%)↓	WER (%)↓	Pro. (%)↓
Main Set								
Agentic	AuDirector	27.12	5.25	6.18	2.40	0.15	8.00	**1.88**
Audio-Visual	LTX-2	17.09	5.88	6.70	1.64	0.19	3.71	5.12
	Ovi	22.47	5.31	6.66	12.35	1.87	13.43	13.92
	MOVA	18.56	5.62	6.21	8.69	0.69	15.03	4.59
	JavisDiT++	21.05	4.11	4.85	–	–	–	–
	UniAVGen	46.80	5.34	5.96	5.03	0.36	9.61	3.75
Mixed-Audio	DSAG-Base	**10.46**	5.47	6.05	8.28	2.03	11.15	18.21
	DSAG-Fine	11.27	5.55	6.20	8.98	2.44	11.15	21.88
	Ming-0.5B	82.31	**5.90**	**6.97**	1.39	0.11	2.94	3.74
	Ming-16.8B	64.81	5.83	6.85	**1.12**	**0.07**	**2.50**	2.80
Voice Cloning Set								
Agentic	AuDirector	**40.33**	4.48	5.10	2.19	0.23	10.00	**2.30**
Audio-Visual	UniAVGen	50.08	4.60	4.95	12.41	1.59	27.50	5.78
Mixed-Audio	Ming-0.5B	117.27	**5.45**	**6.26**	1.90	**0.16**	6.08	2.63
	Ming-16.8B	92.57	5.41	6.10	**1.01**	0.26	**4.48**	5.80
Timestamp Set								
Agentic	AuDirector	34.71	4.55	5.57	7.57	0.82	15.10	**5.43**
Audio-Visual	LTX-2	19.10	**5.56**	**6.41**	**2.97**	**0.40**	**5.40**	7.41
	Ovi	24.63	5.08	5.53	20.82	3.68	18.68	19.70
	MOVA	27.52	5.31	6.14	7.08	1.01	14.42	7.00
	UniAVGen	50.56	5.14	5.63	9.15	1.00	15.09	6.63
Mixed-Audio	DSAG-Base	**13.43**	5.22	5.88	70.26	24.02	42.53	56.48
	DSAG-Fine	13.83	5.23	6.03	15.71	4.62	12.82	36.04

FD_{CNN14} computes the Fréchet distance between feature embeddings extracted from a pre-trained PANNs CNN14 model, measuring the distributional similarity between generated and reference audio. CE and CU are aesthetic predictors from the Audiobox suite. These metrics either partially overlap with the primary metrics in the main paper or are not central to the core evaluation focus. We therefore include their full results in the appendix for completeness.

WER Any. metrics: To assess the quality of predicted transcriptions, we design a specialized analysis targeting a specific error type: cases where the model generates content from the non-transcription parts of the reference caption (e.g., environmental descriptions, music annotations, or sound event descriptions) rather than the spoken transcript only. We define three metrics for this error type:

- **Hit Rate (%)**: The percentage of samples containing at least one such error.
- **WER_{caps}**: The average WER computed exclusively on samples that contain the aforementioned error type.
- **Proportion**: The ratio of WER_{caps} to the overall WER across all samples, i.e., Proportion = WER_{caps}/WER. A higher Proportion indicates that this error type contributes more substantially to the overall transcription error, signaling a more severe issue.

These metrics enable diagnostic analysis of whether a model tends to mistakenly incorporate non-speech textual content into its transcription output, providing a direct measure of its ability to generate only the spoken content as instructed.

E Evaluated Model Details

We evaluate all systems using publicly released checkpoints and their official inference implementations. To ensure a uniform evaluation budget across systems and avoid best-of-N selection, we generate one output per input–model setting using the fixed seed reported below. Given the substantial computational cost of full-benchmark generation, we do not perform repeated sampling across seeds. If inference fails for technical reasons (e.g., a runtime error or an empty, corrupted, or otherwise undecodable media file), we retry the same input at most five times and retain the first valid output. We never regenerate an output based on its perceptual quality or evaluation scores. For audio-visual models with video output, we extract the audio track from the generated video using FFmpeg before evaluation. Unless stated otherwise, all inference parameters, preprocessing procedures, and negative prompts follow the official default implementations. For reproducibility, we report the Hugging Face repositories and principal inference parameters below. We use the following shorthand for model repositories:

E.1 Agentic Orchestrator

AuDirector. AuDirector uses `Qwen/Qwen3-Omni-30B-A3B-Instruct` as its orchestration model. Its expert modules are IndexTTS2 (`IndexTeam/IndexTTS-2`) for speech, MusicGen Stereo Medium (`facebook/musicgen-stereo-medium`) for music, and TangoFlux (`declare-lab/TangoFlux`) with `google/flan-t5-large` for sound generation. We use `dropbox-dash/faster-whisper-large-v3-turbo` in FP16 on GPU for transcript verification and `google/embeddinggemma-300m` for text retrieval. The iterative critic consists of `XiaomiMiMo/MiMo-Audio-7B-Instruct`, its `XiaomiMiMo/MiMo-Audio-Tokenizer`, and `laion/clap-htsat-fused`. On the voice-cloning subset, the benchmark voice prompt is supplied directly to IndexTTS2 instead of retrieving a speaker from AuDirector's internal voice library. For timestamp-conditioned evaluation, we extend the planner to pass the requested temporal intervals to the expert synthesis and mixing stages. We use seed 142.

E.2 Native Mixed-Audio Generators

Dasheng AudioGen. We use `mispeech/Dasheng-AudioGen` with 25 sampling steps, classifier-free guidance scale 5.0, sway-sampling coefficient -1.0, and seed 42. Dasheng-AudioGen-Base receives the original overall caption. Dasheng-AudioGen-Fine uses the official agentic prompt-refiner template presented in Appendix G (*Prompt of Agentic Prompt Refiner*) of the original Dasheng AudioGen paper (Mei et al., 2026).

Ming-Omni-TTS. We evaluate both `inclusionAI/Ming-omni-tts-0.5B` and `inclusionAI/Ming-omni-tts-16.8B-A3B`. For both variants, we set the maximum number of decoding steps to 200, classifier-free guidance to 2.0, $\sigma = 0.25$, and temperature to 0, and use seed 1895. Captions are converted to the model's required structured format before inference. The supplied reference audio is used only in the voice-prompt condition.

E.3 Unified Audio-Visual Generators

LTX-2.3. We run the one-stage text-to-audio-video pipeline with the `Lightricks/LTX-2.3` checkpoint and `google/gemma-3-12b-it-qat-q4_0-unquantized` text encoder. Generation uses 242 frames at 24 fps, a first-stage resolution of 512×768, and 30 inference steps with seed 42. Audio CFG, STG, and rescaling strengths are set to 7.0, 1.0, and 0.7, respectively, with skip step 0. We retain the official negative prompt for visual artifacts, synchronization errors, muted or distorted audio, and unintended dialogue.

Ovi 1.1. We use the 10-second configuration from `chetwinlow1/Ovi`, selecting the `960x960_10s` model in text-to-video mode with the implementation's default seed 100. We generate at 704×1280 with 50 UniPC sampling steps, timestep shift 5.0, audio guidance 3.0, video guidance 4.0, and skip-layer-guidance layer 11. FP8 and CPU offloading are disabled. We use the released negative prompts for visual artifacts and robotic, muffled, echoing, or distorted audio.

MOVA. We use `OpenMOSS-Team/MOVA-720p` with 241 frames at 24 fps and output resolution 352×640. Inference uses 50 steps, CFG scale 5.0, sigma shift 5.0, seed 42, and CPU offloading. We retain the official artifact-suppression negative prompt.

JavisDiT++. We use `JavisVerse/JavisDiT-v1.0-jav` with seed 1024, 161 frames, 16 fps, 16-kHz audio, 480p resolution, and a 9:16 aspect ratio. Inference uses BF16, batch size 1, one generation loop, a five-frame conditioning window, and alignment interval 5. The official negative prompt is retained. Because the released interface does not support an explicit transcript field, JavisDiT++ is evaluated without transcript conditioning.

UniAVGen. We use `MCG-NJU/UniAVGen` with 50 inference steps, timestep shift 10.0, audio guidance 2.0, video guidance 3.0, skip-layer-guidance layer 11, MACFG proportion 0.5, and seed 2025. We use AVG mode for text-conditioned generation and RAVG mode for voice-prompt evaluation. The video negative prompt is empty, and a blank reference image is provided when required by the interface.

F Input-Processing Prompt Templates

This section provides the complete prompts used to transform MMAG captions into model-specific inputs. We start from the three-stage templates in the official AuDirector prompt release. We adapt the Stage-1 and Stage-3 prompts to MMAG while retaining the released Stage-2 prompt unchanged. The adapted final stage is further divided into two variants: Stage 3 for the main set and Stage 3T for the timestamp subset. Consequently, the main-set pipeline uses Listings 1, 2, and 3, whereas the timestamp pipeline replaces the final stage with Listing 4. Ming-Omni-TTS uses Listing 5 to convert each overall caption into its required structured input.

F.1 AuDirector Prompts

F.1.1 Adapted Stage 1: Speaker profile and transcript extraction

Listing 1: MMAG-adapted AuDirector Stage-1 prompt for speaker-profile and transcript extraction.

```
Role: Audio script planning
Task: Based on the free text description provided by the user (including single speaker
    information, text transcription, and other information), complete the character
    information of the speaker and generate a monologue JSON strictly according to the
    transcription content.
If the input does not contain valid speaker information or text transcription, the
    default character information will be output.
Input (input format): A free text description that maybe include the following
    information: the speaker's gender, age range, accent, emotional state, and the
    speaker's text transcription **(enclosed by backslash-double-quote \")**
Core Instructions:
Information extraction: Ignore audio events, music, and environmental descriptions in
    the input text. Only focus on speaker information and text transcription.
Character completion: Extract and complete the speaker's character information (tone,
    accent, emotion, pause, etc.) based on input, and infer tone, intonation, and
    personality traits.
Output Format:
Strictly use the JSONL format (one independent JSON object per line). It is strictly
    prohibited to include Markdown code block tags (such as JSON), comments, or
    redundant text.
Character Information:
Output a line of JSON in the following format:
{"name": "character name", "desc": "a young/middle-aged/elderly male/female of XX-XX
    years old, with specific accent description, emotional description, timbre
    characteristic description, intonation characteristic description, and personality
    image description. "}
- If the input is invalid (no speaker information), the **default** output is:
    - {"name": "Unknown", "speaker_age": "20-30", "speaker_sex": "Male", "desc": "A
    young male between the ages of 20-30, with no specific accent, neutral emotions, a
```

```
    clear and ordinary tone, a stable and non fluctuating tone, a vague personality
    image, and a lack of characteristics. "}

Desc field requirements:
-When there is speaker information (not default output), **'name' cannot be 'Unknown'**
-Must include age range (e.g. 20-30 years old)
-Must include gender (male/female)
-Must include accent (e.g. North American accent)
-Must contain emotions (such as happiness)
-Must include timbre characteristics (such as cool and steady, deep and mellow, hoarse
    and grainy, etc.)
-Must include intonation characteristics (such as low and slow, lazy and relaxed, crisp
    and fast speaking, etc.)
-Must include a personality and image description (such as: proud and cold, heavy on
    emotions and loyalty, full of hooliganism, etc.)
-The overall description should be vivid and suitable for sound interpretation
Example:
{"name": "Jenny", "speaker age": "20-30", "speaker sex": "female", "desc": "A young
    woman in her 20s and 30s, with a North American accent, calm emotions, a cool and
    steady tone, and a strong magnetic graininess. Her tone is low and slow, mixed with
     contempt and a desire for control, vividly depicting the image of a proud, cold,
    and powerful queen sister. "}
```

F.1.2 RETAINED STAGE 2: EMOTION LABELING

Listing 2: Released AuDirector Stage-2 prompt, retained unchanged for emotion-vector assignment.

```
Role: Audio Drama Emotion Labeler
Task: Process the monologue output from "1.prompt", append emotion vectors and pause
    durations.
Input:
Output from "1.prompt", containing character profiles and monologue:
Character Info: {"name": "Name", "speaker_age": "Age Range", "speaker_sex": "Male/Female
    ", "desc": "Description"}
monologue Info: {"speaker": "Name", "content": "Content"}
Processing Requirements:
Only process monologue lines containing speaker and text.
Extract character attributes from the Character Info for emotion inference (do not
    output the profile info).
Add emo_vector to every monologue line.
Output Format:
Strict JSONL format, with one JSON object per line. Output only the monologue parts.
Field Definitions:
speaker: Character name.
emo_vector: [Happy, Angry, Sad, Fear, Disgust, Depressed, Surprise, Neutral]. Range:
    0.0-1.0 (Keep 1 decimal place).
Rules:
Emotion Vector:
Casual/Daily: 0.0 - 0.3
Core Emotion: 0.4 - 0.7
Intense Emotion: 0.7 - 0.8 (Use 0.9+ sparingly)
Compound/Mixed emotions are allowed.

Example:
Input: {"speaker": "Jenny", "content": "Today's work must be completed on time."}
Output: {"speaker": "Jenny", "content": "Today's work must be completed on time.", "
    emo_vector": [0.0, 0.1, 0.0, 0.0, 0.0, 0.0, 0.0, 0.6]}
```

F.1.3 ADAPTED STAGE 3: MAIN-SET PRODUCTION PLANNING

Listing 3: MMAG-adapted AuDirector Stage-3 prompt used for the main set.

```
# Role
```

```
AI Audio Director. Convert text into audio production script (JSONL).

# Input
The input consists of two parts, concatenated in the following order:

1. **Transcription JSONL** (multiple lines, each line a JSON object):
   Each line contains:
   - `speaker`: speaker identifier (string)
   - `content`: the spoken text (string)
   - `timestamps`: character-level timestamps (list of objects, each with `char`, `start
     `, `end`)

2. **Audio Description Text** (single line of free text):
   A natural language description of audio events, music, and scene ambience.
The input is provided as a plain text file where the JSONL lines appear first, followed
     by exactly one line containing the audio description text.

# Processing Logic
For each item in the input list (tracked by `speech_seq` ID), and the audio description
     text, process according to the following logic:

1.  **Speech (Anchor)**:
    -   `audio_type`: "speech".
    -   `speech_seq`: "ID".
    -   `text`: `content` from input.
    -   Serves as the timeline anchor for all other sound effects.

2.  **BGM (Strictly Musical)**:
    -   **Definition**: Contains only instrumental performance, electronic synthesis, or
      melodic scoring.
    -   `placement`: Typically "overlap" (covers the current segment).
    -   `action`: "start" / "stop" (used for cross-segment control).
    -   `bgm_seq`: "ID" (same ID for start and stop pairing).
    -   `anchor_to_speech_seq`: Associates with the current `speech_seq` ID.
    -   Extracted from the audio description text.

3.  **SFX (Sound Effects)**:
    Uses the `placement` field to position sounds relative to speech:

    * **Case A: Before (Pre-positioning)**
        -   Scenario: Environmental setup or actions before speech (e.g., breathing,
     footsteps, waves entering).
        -   `anchor_to_speech_seq`: `speech_seq` ID.
        -   `placement`: "before"
        -   `offset`: 0 (system automatically moves forward).
        -   `vol`: -5 to -2 (volume range).
        -   `duration`: **Estimated from text input (typically 1.5s -- 4.0s)**. Long
     enough to establish scene or complete action, not too long.
        -   Extracted from the audio description text.

    * **Case B: After (Post-positioning)**
        -   Scenario: Pauses, reverb, independent events after speech (e.g., dog bark,
     hanging up phone).
        -   `anchor_to_speech_seq`: `speech_seq` ID.
        -   `placement`: "after"
        -   `offset`: 0 (system automatically delays).
        -   `vol`: -5 to -2 (volume range).
        -   `duration`: **Estimated from speech content (typically 1.0s -- 3.0s)**. For
     example, gunshot reverb is longer, phone hang is shorter.
        -   Extracted from the audio description text.

    * **Case C: Overlap (Simultaneous)**
        -   **Definition**: Sounds that occur during the speech. Divided into two types
     by `layer`:
```

```
        -   **Type 1: Global Ambience (Sustained environmental sound)**
            -   **Definition**: Continuous background sound that spans one or more
    sentences (rain, traffic, wind).
            -   **Logic**: Generate a record **only at the start** of the ambient sound.
            -   `placement`: "overlap"
            -   `vol`: -12 to -8 (volume range).
            -   `offset`: 0.0
            -   **`scope_speech_seqs`** (key field):
                *   **Definition**: An **integer list** containing all `speech_seq` IDs
    covered by this ambient sound.
                *   **Single-sentence ambience**: List contains only the current ID, e.g
    ., `[1]`.
                *   **Continuous ambience**: List contains all IDs from start to end.
                *   *Example*: Rain covers sentences 1 to 5 -> output `[1, 2, 3, 4, 5]`.
            -   Extracted from the audio description text.

        -   **Type 2: Sync Action (Point action)**
            -   Scenario: Table slap, cough, gunshot, door close (inserted in the middle
    ).
            -   `placement`: "overlap"
            -   `offset`: **Must** take the `start` time of the corresponding word from
    `timestamps`.
            -   `anchor_to_speech_seq`: `speech_seq` ID.
            -   `vol`: -5 to -2 (volume range).
            -   `duration`: **Short and physically appropriate (typically 0.3s -- 1.5s)
    **. For example, table slap (0.5s), door open (1.2s).
            -   Extracted from the audio description text.

# Output Schema
-   `clip_id`: Globally auto-incremented ID.
-   `audio_type`: "speech", "bgm", "sfx".
-   `speech_seq`: integer (exists only when `audio_type` = "speech").
-   `bgm_seq`: integer (exists only when `audio_type` = "bgm").
-   `placement`: "before", "after", "overlap" (null for `audio_type` = "speech").
-   `vol`: integer volume.
-   `desc`: English description. **SFX description must be specific** (e.g., "Heavy rain
    " instead of "Sad mood").
-   `duration`: float (seconds).
-   `anchor_to_speech_seq`: integer (the speech ID to which this clip is anchored).

# Critical Format Constraint
- Output **strict JSONL** (JSON Lines): **each line MUST contain a complete, valid JSON
    object.**
- **Do NOT split a single JSON object across multiple lines.**

# Example Scenario
Output:
{"clip_id": 1, "audio_type": "sfx", "desc": "Distant thunder rumble intro", "placement":
     "before", "vol": -3, "duration": 2.5, "anchor_to_speech_seq": 1}
{"clip_id": 2, "audio_type": "sfx", "desc": "Heavy rain ambience loop", "placement": "
    overlap", "vol": -10, "scope_speech_seqs": [1, 2, 3], "offset": 0.0}
{"clip_id": 3, "audio_type": "bgm", "bgm_seq": 1, "desc": "Sad piano melody", "action":
    "start", "placement": "overlap", "vol": -15, "anchor_to_speech_seq": 1}
{"clip_id": 4, "audio_type": "speech", "speech_seq": 1, "text": "The wind is rising and
    the sky is getting darker."}
{"clip_id": 5, "audio_type": "sfx", "desc": "Coughing", "placement": "overlap", "offset
    ": 1.2, "vol": -5, "duration": 0.8, "anchor_to_speech_seq": 2}
{"clip_id": 6, "audio_type": "speech", "speech_seq": 2, "text": "This terrible weather"}
{"clip_id": 7, "audio_type": "bgm", "bgm_seq": 1, "desc": "Sad piano melody", "action":
    "stop", "placement": "overlap", "vol": -15, "anchor_to_speech_seq": 3}
{"clip_id": 8, "audio_type": "speech", "speech_seq": 3, "text": "Let's go in."}
{"clip_id": 9, "audio_type": "sfx", "desc": "Door closing shut", "placement": "after", "
    vol": -2, "duration": 1.5, "anchor_to_speech_seq": 3}
```

F.1.4 ADAPTED STAGE 3T: TIMESTAMP-CONDITIONED PRODUCTION PLANNING

Listing 4: Timestamp-specific Stage-3T variant used for the MMAG timestamp subset.

```
Role: AI Audio Director
Task: Convert a natural language description of an audio scene into an engineering-grade
     JSONL audio production script. The description contains time-anchored events (
    speech, sound effects, music) and may include explicit timestamps or vague temporal
     references. Your output must be a single JSONL stream where each line is a valid
    JSON object representing one audio asset.

Input Format:
A single paragraph of natural language (in English or Chinese) that describes:
- Dialogue content (enclosed in double quotes, e.g., "And then you can add your contacts
    ...")
- Timestamps (e.g., "between roughly 1.29 and 4.48 seconds", "around the start", "near
    4.8 seconds", "throughout")
- Sound effects (e.g., clicking, footsteps, wind)
- Background ambiance or music
- Speaker attributes (gender, age, accent, etc.) -- used only for context, not output.

Example input:
"Around the start, a clicking sound occurs. Between roughly 1.29 and 4.48 seconds, the
    neutral teenage male speaker says in US English: \"And then you can add your
    contacts and I'm going to hit not now for now\", with another click heard during
    this period. Near 4.8 seconds, another click occurs. Between approximately 5.84 and
     9.26 seconds, the speaker states: \"And then from here, you can pick some
    interests that you're interested in so like\", while a fourth click appears around
    the middle of this segment. There is minimal background noise in a quiet indoor
    environment."

Assumptions:
- The entire audio scene lasts about 10 seconds unless otherwise indicated by the
    description.
- All times are in seconds, relative to the start of the audio (0.0).
- If explicit timestamps are given, use them exactly.
- If only vague references (e.g., "around the start", "throughout", "near the end") are
    given, infer reasonable start/end times based on the overall duration (10s) and the
     sequence of events. Distribute events evenly, ensuring no overlaps that violate
    physical plausibility (e.g., speech cannot overlap with itself).
- Each speech segment is treated as a separate speech asset.

Output Types (4 types, but only those present in the input should be emitted):
1. Speech -- dialogue content. Required for every quoted utterance.
2. BGM (Background Music) -- continuous music or score. Only if explicitly mentioned.
3. Foreground SFX -- discrete, precisely timed sound effects that are action-oriented (e.
    g., clicks, door slams, gunshots).
4. Background SFX -- continuous ambiance or environmental sounds (e.g., wind, rain,
    distant traffic).

Note: In this simplified output schema, we use absolute start/end times instead of
    speech_seq and anchor. All events are placed on the same global timeline.

Output Schema (each line is a JSON object):
{
  "clip_id": <integer, globally unique, starting from 1 and incrementing by 1 for every
    output line>,
  "audio_type": "<speech|sfx|bgm>",
  // For speech:
  "text": "<transcribed dialogue text>",
  // For sfx or bgm:
  "desc": "<English description of the sound, suitable for searching a sound library>",
  // For all types:
  "vol": <integer, decibel value in the recommended range>,
  "start": <float, start time in seconds>,
```

```
  "end": <float, end time in seconds>
}

Volume (vol) guidelines:
- Speech: -3 to 0 dB (clear, prominent)
- Foreground SFX: -15 to -5 dB (noticeable but not overpowering)
- Background SFX: -20 to -10 dB (ambient, supportive)
- BGM: -25 to -20 dB (underscore, non-intrusive)

Task Logic (step-by-step):

1. Parse the input description and identify all sound sources:
   - Extract every quoted string as a speech event. Note its start and end times if
     given; if not, infer from surrounding context.
   - Extract all sound effects (e.g., "clicking", "footsteps", "wind"). Classify them as
      Foreground SFX if they are discrete and tied to specific moments; otherwise as
     Background SFX if they are continuous or atmospheric.
   - Extract any mention of music or melody as BGM.

2. For each identified event, determine its start and end times:
   - If explicit timestamps are provided (e.g., "between 1.29 and 4.48 seconds"), use
     those values directly.
   - If a phrase like "around the start" appears, assign a start near 0.0 (e.g., 0.0 to
     0.5) and an end based on duration (e.g., 0.5 or until the next event).
   - If "throughout" or "background" is mentioned, set start = 0.0 and end = 10.0 (or
     the inferred total duration).
   - If a click is said to occur "during this period" (within a speech segment), place
     it at a reasonable midpoint of that speech segment (e.g., average of the speech
     start and end).
   - If no time is given for a sound, infer its placement based on the narrative order (
     e.g., first sound before first speech, second sound between speeches, etc.).
     Distribute events evenly across the 10-second timeline, respecting the described
     order.

3. Generate one JSON object per event. Order events by their start time (chronological).
      For overlapping events (e.g., speech with background SFX), they can be output in
     any order, but start times should be sorted.

4. For SFX, each JSON object is a separate event. **And ignore irrelevant expressions
     such as room tone and silence.**

5. Ensure clip_id increments globally starting from 1.

6. Do not output any Markdown code block markers (like ```json). Output raw JSONL text,
     each line a valid JSON object.

Example Output (for the given input):

{"clip_id":1,"audio_type":"sfx","desc":"Click sound","vol":-10,"start":0.0,"end":0.3}
{"clip_id":2,"audio_type":"speech","text":"And then you can add your contacts and I'm
     going to hit not now for now","vol":-2,"start":1.29,"end":4.48}
{"clip_id":3,"audio_type":"sfx","desc":"Click sound","vol":-10,"start":2.8,"end":3.1}
     <!-- inferred midpoint of the first speech -->
{"clip_id":4,"audio_type":"sfx","desc":"Click sound","vol":-10,"start":4.8,"end":5.1}
{"clip_id":5,"audio_type":"speech","text":"And then from here, you can pick some
     interests that you're interested in so like","vol":-2,"start":5.84,"end":9.26}
{"clip_id":6,"audio_type":"sfx","desc":"Click sound","vol":-10,"start":7.5,"end":7.8}
     <!-- middle of second speech -->

Note: The above is an illustration. Your output must be based on the actual input
     description. For background SFX, if described as "minimal background noise", you
     may emit a room tone or omit it if not explicitly called out. However, if the
     description mentions a continuous ambiance (e.g., "wind blowing"), you must include
      it.
```

```
Final Instruction:
Read the input description carefully and produce the JSONL output. Only output the JSON
    lines, nothing else. Do not include explanations or extra text.
```

F.2 MING-OMNI-TTS FORMAT-CONVERSION PROMPT

The runtime prompt contains literal Chinese JSON field names. The listing below reproduces these field names directly; no prompt content is changed.

Listing 5: Prompt used to convert MMAG captions to the Ming-Omni-TTS input schema.

```
You are a precise information extraction and format conversion assistant. Convert the
    given audio caption into a JSON object with the exact schema below.

# Input:
A text string describing an audio clip. It may contain:
    1) speech transcript, typically enclosed in double quotes ("...")
    2) speaker age, gender, emotion, dialect/accent
    3) sound event descriptions
    4) music genre, mood, instruments

# Output schema:
{
    "风格": "",
    "方言": "",
    "情感": "",
    "BGM": {
        "ENV": "",
        "Genre": "",
        "Mood": "",
        "Instrument": ""
    },
    "text": ""
}

# Rules:
1) Extract only information explicitly present in the caption. **Do not infer, invent,
    or add extra content.**
2) If a field is not present, set it to null. Do not use empty strings.
3) Keep the Chinese field names exactly as shown. Always keep the "BGM" object and set
    missing subfields to null.
4) Content correspondence
    - "风格": the speaker-related description, including age, gender, emotion, dialect/
    accent, language, and speaking action if stated. Preserve the original wording as
    much as possible. Do not put the speech transcript here.
    - "方言": the speaker's dialect/accent as a bare normalized label without the word "
    accent". Examples: "American accent" -> "American". If not stated, null.
    - "情感": the speaker's emotion as a base adjective, not an adverb or noun. Examples
    : "happily" -> "happy". If not stated, null.
    - "BGM.ENV": sound event description / environmental audio events. Extract concrete
    sound events. If not stated, null.
    -"BGM.Genre": music genre/style. If not stated, null.
    -"BGM.Mood": music mood. If not stated, null.
    -"BGM.Instrument": musical instruments. If not stated, null.
    -"text": the speech transcript exactly as spoken, without surrounding quotation
    marks or speaker description. Preserve punctuation and capitalization. If there is
    no speech transcript, null.

Return only valid JSON. Do not include markdown code fences, explanations, or extra text.


# Example:
Input:
```

```
An elderly man with an American accent happily remarked in English: "As the Grand
    National and the GNX. In 1987, buick introduced the GNX." Subsequently, a rhythmic
    electronic music with synthesizer and drum plays. Next, the car's engine roars,
    especially the piercing sound of the tires. The professionally mixed audio creates
    the atmosphere of a studio-produced commercial environment.

Output:
{
"风格": "An elderly man with an American accent happily remarked in English",
"方言": "American",
"情感": "happy",
"BGM": {
"ENV": "the car's engine roars, the piercing sound of the tires",
"Genre": "electronic",
"Mood": null,
"Instrument": "synthesizer, drum"
},
"text": "As the Grand National and the GNX. In 1987, buick introduced the GNX."
}
```

G LIMITATIONS

Despite its comprehensive scope, MMAG has several limitations.

The 10-second clip duration, single-speaker English restriction, and imbalanced sound/music categories limit generalizability and may bias fine-grained evaluation on underrepresented classes.

In addition, even with a multi-expert pipeline and manual verification, sound events, music genres, and precise timestamps remain inherently subjective, introducing unavoidable uncertainty into the ground truth.

For evaluation, we incorporate LLM-based semantic metrics, but current multimodal LLMs still show limited reliability on fine-grained audio tasks, such as timestamp-aware event localization. This is expected to improve as foundation models continue to advance.