AUDIOSCAPE-TTA: A STRUCTURED SOUNDSCAPE BENCHMARK FOR FINE-GRAINED TEXT-TO-AUDIO EVALUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Text-to-audio (TTA) generation has made substantial progress in synthesizing realistic audio from natural-language descriptions. However, determining whether generated audio faithfully satisfies complex instructions remains challenging. Widely used global text–audio similarity metrics provide limited insight into which specific semantic requirements are satisfied or omitted. We introduce **AudioScape-TTA**, a structured and complexity-aware benchmark for fine-grained TTA evaluation. The benchmark represents soundscapes through verified annotations of scene context, sound effects, background music, and speech, and characterizes sample complexity using annotation-derived event density and audio-component structure. From these annotations, we derive fixed semantic rubrics covering event presence, acoustic attributes, and specified speech content. Event and attribute requirements are evaluated by an audio-language model, while speech content is assessed through a separate ASR-based coverage pathway. AudioScape-TTA contains 2,258 audio–text pairs and 25,707 semantic rubrics, enabling scalable and interpretable analysis of TTA systems. Evaluation of 13 open-source models reveals distinct profiles in event realization, attribute control, speech-content coverage, and compositional soundscape coverage. On a human-study subset, the proposed scores achieve higher model-level rank correlation with human semantic judgments than CLAP similarity, supporting structured semantic assessment beyond global text–audio alignment. Anonymous project page.

1 INTRODUCTION

Recent advances in text-to-audio (TTA) generation have enabled models to synthesize diverse audio content, including sound effects, music, and speech, from natural-language instructions (Vyas et al., 2023; Yang et al., 2024). As generation quality and controllability improve, TTA systems are increasingly expected to realize coherent soundscapes containing multiple heterogeneous audio components rather than isolated events. Consequently, evaluation must assess not only perceptual plausibility but also whether the structured requirements specified by the text are faithfully realized.

However, existing TTA benchmarks provide limited support for evaluating structured soundscape generation. First, existing benchmarks do not consistently provide realistic and compositional instructions that explicitly represent heterogeneous audio components within a single scene. As summarized in Table 21, prior studies investigate caption-level generation, conditional generation, temporal control, and event-level relations (Kim et al., 2019; Drossos et al., 2020; Xie et al., 2025; He et al., 2025; Wang et al., 2026; 2025a). However, these resources do not jointly provide scalable coverage of sound effects, background music, and target-utterance-constrained speech within a unified soundscape. In particular, speech is often treated as an audio event rather than a linguistic signal with target-utterance constraints, limiting the evaluation of speech-content preservation in complex soundscapes. Second, existing benchmarks provide limited complexity-aware analysis of structured soundscape generation. Although recent studies consider multi-event prompts and event-level relations, evaluation is typically summarized as aggregate performance. Such evaluation does not characterize sample-level complexity, which may vary with acoustic content, active audio categories, and their clip-level composition. Consequently, existing benchmarks provide limited insight into how semantic instruction following varies as soundscape complexity increases.

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.

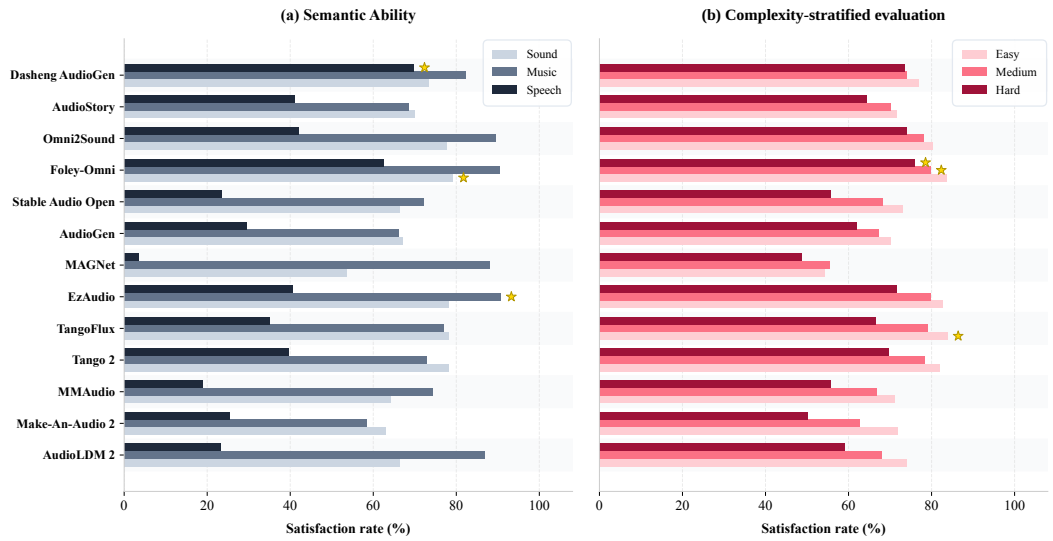

Figure 1: Component-specific and complexity-stratified satisfaction rates. In panel (a), Sound and Music denote rubrics associated with SFX and BGM, respectively; the speech-component score includes speech-related requirements and is distinct from the speech-content-only SCCA metric. Yellow stars mark the highest observed score within each component or complexity stratum. The strata in panel (b) differ in composition as well as complexity, so these comparisons do not isolate a causal effect of complexity.

Beyond benchmark construction, automatic evaluation remains challenging. Global metrics such as CLAP (Wu et al., 2023) and holistic audio-language judgments summarize overall alignment but do not reveal which requested components or attributes are satisfied. Recent QA- and rubric-based audio evaluators provide finer-grained alternatives. Our focus is complementary: we establish an annotation-grounded requirement set before model outputs are scored, making requirement coverage auditable and evaluation questions identical across systems. This fixed protocol supports reproducibility and comparison consistency rather than guaranteeing evaluator accuracy, while providing requirement-level outputs for analyzing structured soundscape generation.

To address these limitations, we introduce **AudioScape-TTA**, a structured and complexity-aware benchmark for fine-grained TTA evaluation. AudioScape-TTA contains 2,258 audio–text instances and 25,707 fixed semantic rubrics. Each instance is represented by a verified structured annotation containing scene context and the SFX, BGM, and Speech components present in the clip. We characterize sample complexity through two complementary dimensions: *event density*, which counts verified SFX event nodes together with clip-level Speech and BGM presence indicators, and *structural complexity*, which combines the number of active audio categories with a speech co-occurrence indicator. This design supports stratified analysis across different levels of event-node density and audio-component structure. Figure 1 previews the component-specific and complexity-stratified results. Building upon these annotations, we derive a fixed set of fine-grained semantic requirements covering event presence, acoustic attributes, and specified speech content. Our audio-grounded rubric-based framework evaluates event and attribute requirements using an audio-language model, while a separate ASR-based coverage pathway assesses specified speech content. This design provides requirement-level outputs for diagnosis beyond global text–audio similarity and supports analysis across audio categories, soundscape compositions, and complexity levels.

We evaluate 13 open-source TTA models spanning diverse generation paradigms. Results show that global similarity and perceptual-quality metrics provide complementary but incomplete views of requirement satisfaction. Stratified results show performance differences across benchmark-specific bins that vary jointly in complexity and component composition. On the 10-prompt human-study subset, model-level Overall SR correlates more strongly with human semantic judgments than MS-CLAP similarity. This result supports the utility of structured semantic assessment at the model level; it does not certify individual questions or Yes/No decisions.

Our contributions are summarized as follows:

- We introduce **AudioScape-TTA**, a structured and complexity-aware benchmark that represents realistic audio scenes as structured soundscapes and enables fine-grained evaluation across different complexity levels.
- We propose a rubric-based audio-grounded evaluation that decomposes semantic fidelity into event presence, acoustic attributes, and speech content, enabling scalable and interpretable evaluation.
- We conduct a comprehensive evaluation of 13 open-source TTA models, revealing key limitations in fine-grained semantic instruction following.

2 Related Work

2.1 Text-to-Audio Generation

Text-to-audio (TTA) generation aims to synthesize realistic audio from natural-language descriptions. Early approaches, such as AudioGen (Kreuk et al., 2023), formulate TTA as autoregressive generation over discrete audio tokens, demonstrating the feasibility of large-scale audio synthesis. Recent progress has been largely driven by diffusion-based models. AudioLDM (Liu et al., 2023) introduces latent diffusion for audio generation, while subsequent approaches, including Make-An-Audio (Huang et al., 2023b), Make-An-Audio 2 (Huang et al., 2023a), Auffusion (Xue et al., 2024), and EzAudio (Hai et al., 2025), further improve semantic alignment, temporal modeling, and generation efficiency. More recently, flow-based methods such as TangoFlux (Hung et al., 2024) and MeanAudio (Li et al., 2026b) achieve faster generation with competitive quality. In addition, unified audio generation frameworks, including Foley-Omni (Tao et al., 2026), Omni2Sound (Dai et al., 2026), and AudioStory (Guo et al., 2026), extend TTA toward multimodal audio synthesis.

TTA benchmarks and controllability. AudioCaps (Kim et al., 2019) and Clotho (Drossos et al., 2020) provide audio–caption pairs, while AudioCondition (Guo et al., 2024) and RiTTA (He et al., 2025) study targeted conditioning or event relations. AudioTime (Xie et al., 2025) benchmarks event ordering, duration, frequency, and timestamps with temporally aligned text and the grounding-based STEAM metric. T2A-EpicBench, released with T2A-Feedback (Wang et al., 2025b), targets long multi-event narratives. AudioAtlas (Wang et al., 2025a) and VinTAGe-Bench (Kushwaha & Tian, 2025) broaden scene and prompt coverage, whereas TTA-Bench (Wang et al., 2026) studies holistic functional quality, reliability, and responsibility. AudioX (Tian et al., 2025) is a broad generation and editing model rather than an evaluation benchmark, and Aurelius (He et al., 2026) contributes relation-aware audio corpora.

Fine-grained and rubric-based evaluation. T2A-Feedback evaluates event occurrence, event sequence, and acoustic/harmonic quality (Wang et al., 2025b). AQAScore formulates text–audio alignment as probabilistic audio question answering (Kuan et al., 2026), while AnyAudio-Judge dynamically decomposes instructions into binary rubric items (Li et al., 2026a). Structured criteria and fine-grained judging have also been studied across text, image, and video (Kim et al., 2024; Leng et al., 2025; Hayes et al., 2025; Corneanu et al., 2025; Deng et al., 2025; Lin et al., 2024; He et al., 2024; Guan et al., 2025). Our contribution is protocol design, not a claim of higher judge accuracy: human-verified, annotation-grounded requirements are fixed before scoring, making coverage auditable and questions identical across systems. The protocol separates Speech, Music, and SFX, routes speech content through ASR coverage, and stratifies by complexity and component structure. Fixed targets support reproducibility and comparison consistency, not necessarily evaluator accuracy; unlike T2A-Feedback, we do not score event sequence.

3 The AudioScape-TTA Benchmark

Task definition. In the current empirical setting, *omni-audio generation* denotes one evaluated model system receiving a single text prompt and producing one complete soundscape waveform in one generation run; the waveform may jointly contain Speech, BGM, and SFX with the requested attributes, but need not contain all three. External pipelines that generate category-specific stems

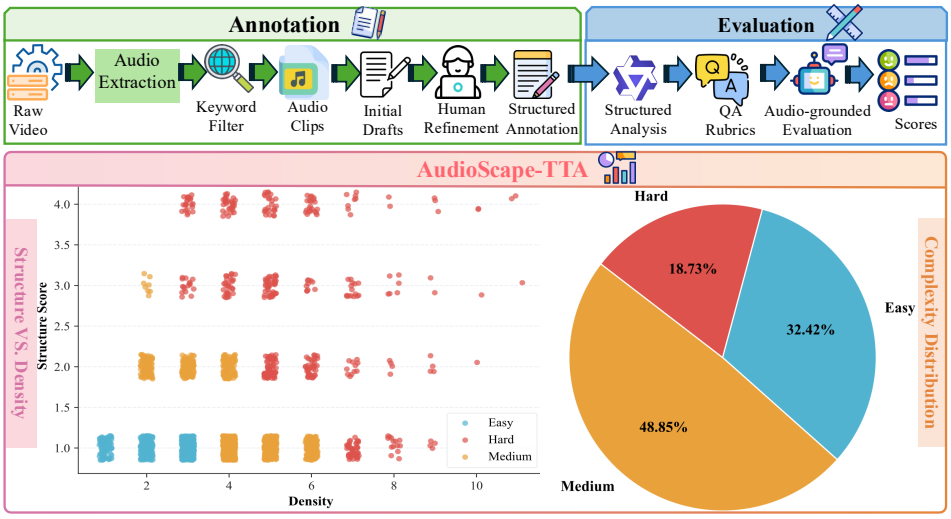

Figure 2: **Overview of AudioScape-TTA**, a complexity-aware benchmark for text-driven omni-audio generation. The top panel illustrates benchmark construction and evaluation: event-presence and event-attribute rubrics are judged by the audio-language evaluator, while speech-content rubrics follow the separate ASR coverage path defined in Section 4. The bottom panel organizes samples within the proposed Density–Structure complexity space.

separately and then mix them are outside the present experiments, although the output-based rubrics could score their final waveform in future work. Figure 2 summarizes the benchmark construction, evaluation workflow, and complexity-aware organization.

Complexity definition. For clip i, let $I_{c,i} \in \{0,1\}$ indicate whether audio category $c \in \{\text{SFX}, \text{Speech}, \text{BGM}\}$ is present, and let $N_{\text{SFX},i}$ be the number of verified SFX event nodes. Event density is $D_i = N_{\text{SFX},i} + I_{\text{Speech},i} + I_{\text{BGM},i}$. Let $K_i = I_{\text{SFX},i} + I_{\text{Speech},i} + I_{\text{BGM},i}$ be the number of active categories. Structural complexity is $S_i = K_i + I_{\text{Speech},i}\,\mathbb{I}[I_{\text{SFX},i} + I_{\text{BGM},i} \geq 1]$, where $\mathbb{I}[\cdot]$ is the indicator function. Thus, the interaction bonus applies specifically when speech co-occurs with SFX or BGM. The core score is

$$C_i = Z(D_i) + Z(S_i) = \frac{D_i - \mu_D}{\sigma_D} + \frac{S_i - \mu_S}{\sigma_S}, \tag{1}$$

where μ_D, σ_D and μ_S, σ_S are the benchmark-wide means and standard deviations of D_i and S_i, respectively. We assign *Easy* when $C_i \leq -1$, *Medium* when $-1 < C_i < 1$, and *Hard* when $C_i \geq 1$. Appendix A.2 provides the detailed counting rationale.

3.1 Collection and Structured Representation

We collect target 10-second clips from major video-sharing platforms, including YouTube, TikTok, and Bilibili, spanning both movie/television content and everyday environmental recordings. Keyword retrieval provides candidate scenes; Gemini-2.5-Pro generates initial structured drafts, and trained annotators verify the audio, correct event and attribute descriptions, and remove ambiguous or unsupported cases.

Compact annotation schema. Each instance pairs a source clip with its caption, stable identifier, scene description, and verified event nodes. Every node records an `event_id`, a record type (Scene, SFX, Speech, or BGM), normalized event text, and applicable source/action, acoustic, temporal, spatial, or repetition attributes. Speech nodes additionally store speaker descriptors and, when specified, the verbatim target utterance; unknown fields remain empty rather than being inferred. Appendix A describes the collection and annotation procedure, record schema, provenance fields, and planned release. Because the platform-collected data mix professionally produced and everyday recordings, they should not be interpreted as an unbiased sample of all real-world audio. Table 1 summarizes the benchmark scale, duration, rubric counts, and finalized speech-related statistics.

Table 1: Key statistics of the AudioScape-TTA benchmark. Speech-annotated clips are those covered by at least one final Speech-category rubric.

Statistic	Number	Statistic	Number
Samples	2,258	Semantic rubrics	25,707
Mean duration (s)	9.95	Mean annotation length (words)	32.45
Mean rubrics per clip	11.38	Mean event density	3.89
Speech-related statistics			
Speech-annotated clips	252 (11.16%)	Clips with target utterances	83
Speech-content rubrics	163	Speech-attribute rubrics	931

4 Audio-Grounded Evaluation

4.1 From annotations to verifiable requirements

The construction pipeline organizes each instance into a sample–component–event–attribute hierarchy. Structured semantic targets are first derived deterministically from the refined annotations. Qwen3.5-27B (Qwen Team, 2026) rewrites these targets into concise yes/no questions without adding or removing requirements. Each rubric stores its question, type, event identifier, and supporting annotation fields. Validation filters malformed, duplicated, or unsupported questions and uses fallback templates when rewriting leaks internal field names or repeats target descriptions excessively. The final set contains 25,707 rubrics and remains fixed for all evaluated systems; Appendix A.5 gives the schema and rewriting instructions. Fixing the rubric set makes coverage auditable and keeps the same questions across systems, supporting reproducibility and comparison consistency rather than implying higher evaluator accuracy than existing QA or rubric protocols.

Rubric terminology. A *semantic rubric* is one field-level requirement rewritten as a yes/no question and linked to a verified event-presence target, attribute field, or target utterance; coordinated descriptors stored in one field are judged jointly. Examples include "Is dog barking audible?" and "Does the dog sound distant?". For instance i, $\mathcal{R}_i$ denotes the complete *instance-level rubric set*; this is a grouping, not a second rubric class. For example, one caption may yield a set containing dog-presence, dog-distance, music-presence, and specified-speech-content requirements. Rubrics may share an event identifier but test distinct requirements and are scored separately.

Table 6 shows selected questions corresponding to the displayed annotation fields. The evaluator receives only the generated audio and the stored question. Coordinated descriptors stored in one field are judged jointly (Appendix F).

4.2 Scoring and aggregation

For rubric $r \in \mathcal{R}_i$ attached to instance i, Qwen3-Omni-Instruct (Xu et al., 2025) receives the generated audio x_i, fixed question q_{ir}, and choices $\{\texttt{A: Yes}, \texttt{B: No}\}$, and returns $\hat{a}_{ir} \in \{A, B\}$. For event-presence and event-attribute rubrics, $s_{ir} = 1$ when $\hat{a}_{ir} = A$ (Yes) and $s_{ir} = 0$ otherwise. All main-benchmark targets are positive requirements derived from annotations. Consequently, the score measures requested-content satisfaction; it is not conventional classification accuracy and does not penalize arbitrary additional sounds that the prompt does not forbid.

Speech-content rubrics use Qwen3-ASR (Shi et al., 2026). After text normalization, clipped multiset overlap is computed between the target and recognized text using Latin-word and CJK-character units. Let R_{ir} denote the resulting target-content coverage for speech-content rubric r of instance i. Its binary satisfaction decision is $s_{ir} = \mathbb{I}[R_{ir} \geq 0.60]$. This coverage is order-agnostic and does not directly penalize insertions, so SCCA@0.60 should not be interpreted as exact transcript accuracy. For any set $\mathcal{S}$ of instance–rubric pairs, the satisfaction rate is

$$\mathrm{SR}(\mathcal{S}) = \frac{100}{|\mathcal{S}|} \sum_{(i,r)\in\mathcal{S}} s_{ir}. \tag{2}$$

Overall SR uses all rubrics; Event SR and Attribute SR restrict to the corresponding types, while SCCA reports speech-content rubrics separately. Overall SR is a *micro-average*: it weights require-

Table 2: Fine-grained results on AudioScape-TTA. SR and SCCA are percentages; higher is better. Overall SR is rubric-micro averaged. Bold and underline denote the best and second-best observed values.

Model	Overall SR	Event SR	Attribute SR	SCCA @0.60	$\mathrm{CLAP}_{\mathrm{MS}}$
AudioLDM 2	66.92	68.43	66.57	0.00	0.4193
Make-An-Audio 2	61.33	61.48	61.89	0.00	0.3321
AudioGen	66.53	68.40	65.92	0.00	0.3664
Tango 2	76.82	<u>80.84</u>	74.78	0.00	0.4145
TangoFlux	76.77	80.55	74.89	0.00	<u>0.4736</u>
EzAudio	<u>78.20</u>	80.83	<u>77.17</u>	0.00	0.4713
MAGNet	53.39	49.89	56.50	0.00	0.2827
Stable Audio Open	65.95	67.57	65.51	0.00	0.3480
MMAudio	64.73	67.80	63.24	0.00	**0.5169**
Foley-Omni	**79.62**	**81.74**	**78.34**	<u>57.06</u>	0.4378
Omni2Sound	77.59	79.98	76.72	0.00	0.4370
AudioStory	69.03	70.09	69.02	0.00	0.3989
Dasheng AudioGen	74.50	74.86	74.21	**77.30**	0.4139

ments, not components equally. Speech-content rubrics constitute only $163/25{,}707 \approx 0.634\%$ of all items, so success on that subset can change the overall score by at most approximately 0.634 percentage points when other decisions are fixed. We therefore interpret the overall score alongside, not instead of, its separate diagnostic dimensions. Evaluator option confidence is an uncalibrated diagnostic and is reported separately in Table 7.

5 Experiments

5.1 Experimental Setup

We evaluate 13 open-source systems: AudioLDM 2 (Liu et al., 2024), Make-An-Audio 2, AudioGen, Tango 2 (Majumder et al., 2024), TangoFlux, EzAudio, MAGNet (Ziv et al., 2024), Stable Audio Open (Evans et al., 2025), MMAudio (Cheng et al., 2025), Foley-Omni, Omni2Sound, AudioStory, and Dasheng AudioGen (Mei et al., 2026). Each system generates one sample per prompt without manual resampling, selection, or best-of ranking. Most systems use the free-form caption; Make-An-Audio 2 and Dasheng AudioGen use model-native structured prompts derived from the same annotation. Checkpoints, generation settings, architectures, and input formats are documented in Appendix D and Table 20. AudioScape-TTA tests end-to-end generation of a complete soundscape from one text prompt. We evaluate only released model systems through their native single-prompt generation interfaces; no externally composed modular pipeline is included. A modular TTS–music–SFX pipeline remains a complementary design direction, but additionally requires text decomposition, temporal scheduling, spatialization, and loudness-aware mixing. We therefore distinguish *scope-aligned* evaluation from the *compositional/OOD challenge* using each checkpoint's declared capabilities (Appendix Table 10). Scores on unsupported categories measure transfer to the benchmark task rather than failure on a model's native task, and the Overall ranking is not a ranking of systems on their original tasks.

5.2 Fine-grained Semantic Evaluation

Overall Results. Overall SR summarizes aggregate requirement coverage; evidence about unified soundscape generation is interpreted primarily from the component-, composition-, scope-, and speech-specific results below. Foley-Omni achieves the highest observed Overall SR (79.62), Event SR (81.74), and Attribute SR (78.34), followed by different systems across dimensions (Table 2). Attribute SR is lower than Event SR for most models, but not for Make-An-Audio 2 or MAGNet. The two scores are field-partitioned subsets rather than a clean split between event recognition and attribute control: some presence questions already bundle descriptors, and some attribute items overlap those fields or leave the source implicit (Appendix F). Speech-content results differ sharply from overall rankings. Dasheng AudioGen reaches 77.30 SCCA@0.60 and Foley-Omni reaches 57.06;

Table 3: Scope-aware Overall SR (%). Scope-aligned contains supported single-component clips; compositional/OOD contains multi-component clips or clips with unsupported categories. Δ is scope-aligned minus compositional/OOD SR, computed before rounding. Subsets are model-specific and are not directly comparable across models.

Model	Scope-aligned SR	Compositional/OOD SR	Δ
AudioLDM 2	69.52	61.45	8.07
Make-An-Audio 2	65.65	52.79	12.86
AudioGen	67.86	63.90	3.96
Tango 2	79.55	71.08	8.47
TangoFlux	81.42	67.56	13.86
EzAudio	80.22	74.21	6.02
MAGNet	86.85	52.91	33.94
Stable Audio Open	70.08	57.24	12.84
MMAudio	67.36	59.54	7.82
Foley-Omni	80.72	77.32	3.41
Omni2Sound	78.64	75.37	3.27
AudioStory	70.76	65.39	5.36
Dasheng AudioGen	74.45	74.63	-0.19

the remaining evaluated configurations score zero at this threshold. This establishes limited target-utterance coverage under the tested interfaces, not a blanket claim that all these systems should support controllable speech. Threshold sensitivity and complementary transcription diagnostics are reported in Appendix E.

Audio-component and Complexity Analysis. Figure 1(a) shows different component-specific profiles. The composition breakdown in Table 4 further distinguishes single-component and mixed scenes: for the sound+speech+music subset, Dasheng AudioGen and Foley-Omni obtain 74.88 and 72.69 SR, respectively. Figure 1(b) shows performance differences across complexity strata, including lower scores on harder strata for many systems. These strata also contain different proportions of speech, music, and requirements; the results do not isolate complexity from composition or demonstrate a causal architectural advantage.

Scope-aware Analysis. Table 3 shows lower compositional/OOD SR for 12 of 13 systems; Dasheng AudioGen is the only exception, with nearly unchanged performance (74.45 vs. 74.63). Each gap is a within-model descriptor under that checkpoint's own partition: supported-category sets differ, so neither the columns nor the gaps form a common-subset cross-model ranking. For example, MAGNet's 33.94-point gap contrasts a music-only aligned subset with a broad challenge subset, whereas other systems have different aligned coverage. Dasheng AudioGen's near-zero gap should likewise be read in light of its declared support for speech, music, and sound effects, not as evidence of better transfer than other systems. The challenge split measures coverage of the benchmark's complete-soundscape setting rather than performance on a model's native task.

5.3 Comparison with Conventional Evaluation Metrics

The MS-CLAP ranking differs substantially from the rubric-based ranking. MMAudio achieves the highest $\mathrm{CLAP}_{\mathrm{MS}}$ cosine similarity (0.5169), yet obtains an Overall SR of only 64.73, whereas Foley-Omni achieves the highest Overall SR with a lower $\mathrm{CLAP}_{\mathrm{MS}}$ score of 0.4378. This discrepancy highlights the different evaluation targets: MS-CLAP measures holistic semantic correspondence between the caption and audio, while Overall SR directly evaluates whether individual annotation-derived requirements are satisfied. Thus, a generated clip may achieve high global semantic similarity while missing specific events or attributes, or satisfy many explicit requirements without maximizing embedding similarity.

These metrics are therefore complementary rather than interchangeable. This distinction is further supported by the human-alignment analysis below, where Overall SR shows stronger model-level association with human semantic ratings than MS-CLAP on the same study samples. Appendix E further reports perceptual and distributional metrics, including Audiobox-Aesthetic, Fréchet dis-

Table 4: Semantic Satisfaction Rate (SR, %) of open-source text-to-audio generation models across samples with different audio-component compositions. Sp. denotes speech; SFX and BGM denote sound effects and background music. The best and second-best results are highlighted in **bold** and underline, respectively.

Model	SFX	BGM	SFX+Sp.	SFX+BGM	Sp.+BGM	SFX+Sp. +BGM
AudioLDM 2	69.19	84.93	52.37	70.54	56.25	48.60
Make-An-Audio 2	65.65	72.33	50.50	54.44	56.25	47.18
AudioGen	67.86	74.25	53.78	71.25	62.50	54.35
Tango 2	79.53	80.27	60.29	77.93	75.00	65.46
TangoFlux	**81.42**	76.16	65.61	72.12	62.50	56.70
EzAudio	80.22	**87.40**	63.11	81.08	75.00	66.28
MAGNet	51.96	86.85	26.47	73.05	50.00	39.90
Stable Audio Open	70.01	73.42	48.57	65.13	56.25	46.91
MMAudio	67.36	77.53	48.46	68.08	43.75	46.63
Foley-Omni	80.63	84.93	69.36	**82.54**	**87.50**	72.69
Omni2Sound	78.51	84.66	66.08	80.70	75.00	71.98
AudioStory	70.73	72.02	62.12	69.90	43.75	57.91
Dasheng AudioGen	74.22	84.93	**70.71**	76.20	75.00	**74.88**

tance, KL divergence, and Inception Score, to characterize generation quality and distributional similarity from complementary perspectives.

5.4 User Study

To test whether automatic scores track human ratings, we conduct a user study on AudioScape-TTA. We randomly select 10 prompts covering diverse audio-component compositions and complexity levels, and collect generated audio from all 13 evaluated models, resulting in 130 samples. Each sample is rated by 26 participants on three dimensions: *event completeness*, *attribute satisfaction*, and *audio quality*. Event completeness measures whether required events are generated, attribute satisfaction evaluates whether fine-grained acoustic properties are preserved, and audio quality assesses perceptual naturalness and generation artifacts. Participants rate each dimension using a 5-point Likert scale. Model identities are anonymized, and model-to-label assignments are shuffled within each prompt; Appendix C describes the protocol.

As shown in Figure 3, human ratings show model-level trends broadly consistent with the rubric-based SR evaluation. Foley-Omni achieves the highest scores in event completeness and attribute satisfaction, consistent with its leading Overall SR, Event SR, and Attribute SR. Attribute satisfaction is generally lower than event completeness, in the same direction as the automatic Event–Attribute SR gap. This human trend does not isolate pure event recognition from attribute control, because the automatic subsets are not a clean analogue of the two rating dimensions. Dasheng AudioGen obtains the highest audio quality score, indicating stronger perceptual naturalness among the evaluated models, but does not lead the semantic dimensions.

5.5 Human Alignment of Evaluation Metrics

We next test whether rubric-based scores track the user-study ratings more closely than global similarity. We compute model-level Spearman correlations over the 13 systems using automatic scores on the same human-study samples. The human semantic score is the equal-weight average of event completeness and attribute satisfaction; audio quality is excluded because it measures perceptual fidelity rather than semantic instruction following. Event SR, Attribute SR, and Overall SR correlate with their corresponding human ratings at 0.743, 0.825, and 0.879, respectively, compared with 0.312 for MS-CLAP (Appendix Table 8). At the clip level, automatic semantic scores correlate with mean human semantic ratings at $\rho = 0.632$ over 130 generated clips from 10 prompts (prompt-clustered bootstrap 95% CI $[0.477, 0.719]$). Event- and attribute-level associations reach $\rho = 0.644$ and 0.458, respectively (Appendix Table 9). These results support using SR as a model- and clip-level associate of human semantic ratings, not as a certification of individual questions or Yes/No decisions. The study does not tell whether a No is a generation miss or an evaluator miss,

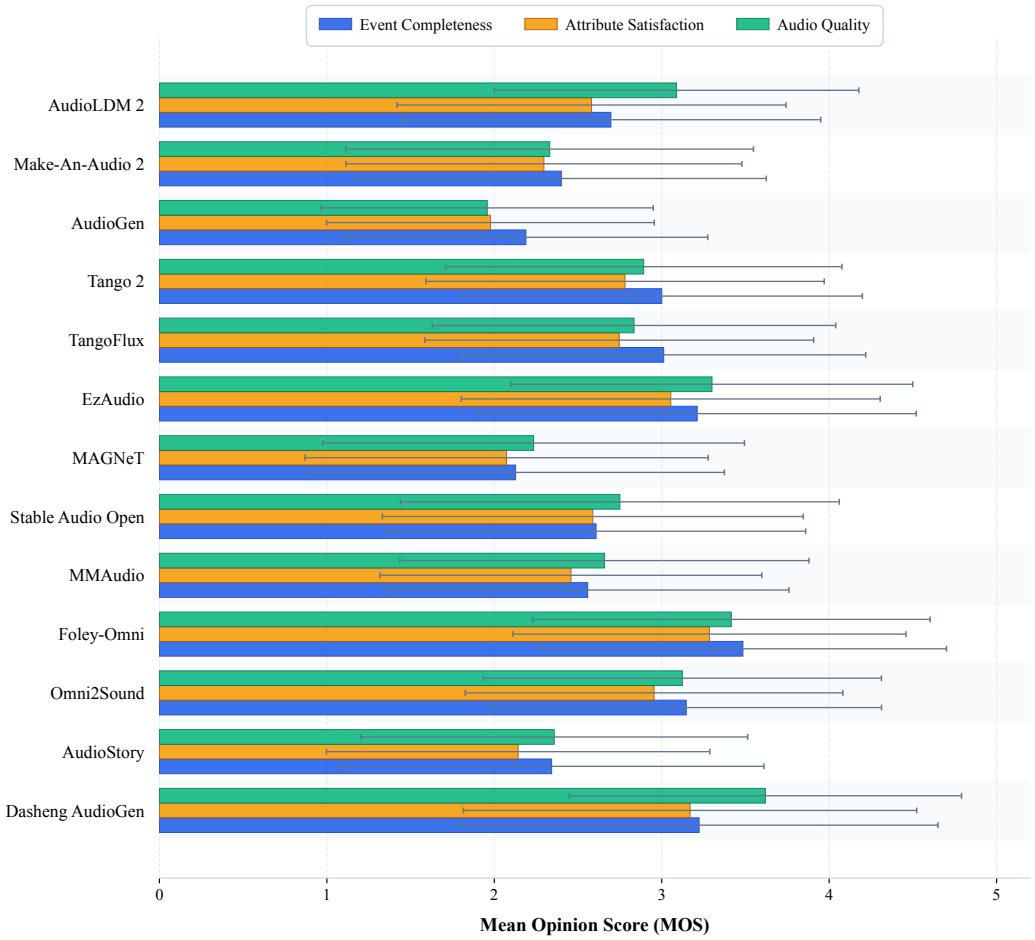

Figure 3: Human ratings of event completeness, attribute satisfaction, and audio quality. Bars show means and error bars show standard deviations, not confidence intervals.

whether a Yes attaches to the intended source, or whether residual underspecified items move ranks. Additional evaluator controls and sensitivity analyses are reported in Appendix E.3. These controls test sensitivity to acoustic evidence but do not constitute generated-audio calibration.

6 Conclusion

AudioScape-TTA introduces a structured and complexity-aware benchmark for evaluating how text-to-audio (TTA) systems realize heterogeneous soundscapes. It comprises 2,258 clips with human-verified SFX, BGM, and Speech annotations and 25,707 fixed, annotation-derived field-level rubrics covering event presence, acoustic attributes, and specified speech content. The benchmark combines audio-grounded binary judgments with a separate ASR-based pathway for speech-content evaluation, enabling requirement-level assessment and analysis across components, compositions, and soundscape complexity. Across 13 open-source TTA systems, AudioScape-TTA reveals distinct semantic-following profiles and persistent challenges in multi-component scenes and target speech content that are not fully captured by global similarity metrics. Human validation further supports the association of Overall SR with human semantic judgments, showing stronger model-level correlation than MS-CLAP on the study subset. Overall, AudioScape-TTA provides requirement-level outputs that can support diagnosis of TTA systems beyond holistic audio-text similarity.

AI USE STATEMENT

Generative AI models were used both within the research methodology and for authoring support. Gemini-2.5-Pro generated initial structured-annotation drafts, which trained annotators verified and corrected. Qwen3.5-27B rewrote deterministic annotation-derived targets into rubric questions; Qwen3-Omni-Instruct evaluated event-presence and event-attribute rubrics; and Qwen3-ASR transcribed generated speech for coverage-based speech-content evaluation. Separately, ChatGPT supported methodological discussion, interpretation and qualification of results, literature checking, language editing, and LaTeX preparation, while Codex assisted implementation and deployment of the demonstration website. The authors verified the methodology, results, references, code, and final manuscript and remain responsible for all submitted and released content.

ETHICS STATEMENT

The benchmark uses platform-collected audio spanning movie/television productions and everyday environmental recordings, and includes human listening judgments. Source material may be subject to third-party copyright and access conditions; the benchmark should not be interpreted as granting redistribution rights to that material. The work targets research evaluation rather than speaker identification or imitation of an identifiable person. Model-judged scores can reflect evaluator and domain biases, so they should not be used as definitive evidence about individual speakers or unrestricted real-world audio quality. Human-study procedures are described in Appendix C.

REPRODUCIBILITY STATEMENT

Section 3 specifies the representation and complexity descriptors, and Section 4 defines rubric construction and aggregation. Appendix A provides annotation details and examples; Appendix B gives metric definitions; Appendix D records checkpoints and inference configurations; and Appendix C describes the listening study. Additional controls and diagnostic results are retained in Appendix E. The anonymous project page linked in the abstract contains the demonstrations. We plan to release the benchmark annotations and evaluation code for research use upon paper acceptance; this is distinct from a claim that all third-party source audio can be redistributed.

REFERENCES

Ho Kei Cheng, Masato Ishii, Akio Hayakawa, Takashi Shibuya, Alexander Schwing, and Yuki Mitsufuji. MMAudio: Taming multimodal joint training for high-quality video-to-audio synthesis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 28901–28911, 2025.

Ciprian A. Corneanu, Qianli Feng, and Aleix M. Martinez. Structured human assessment of text-to-image generative models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 4481–4490, 2025. doi: 10.1109/WACV61041.2025.00440.

Yusheng Dai, Zehua Chen, Yuxuan Jiang, Qiuhong Ke, Jianfei Cai, and Jun Zhu. Omni2Sound: Towards unified video-text-to-audio generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1661–1671, 2026.

Xueqing Deng, Linjie Yang, Qihang Yu, Chenglin Yang, and Liang-Chieh Chen. Leveraging panoptic scene graph for evaluating fine-grained text-to-image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15107–15116, 2025.

Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. Clotho: An audio captioning dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 736–740. IEEE, 2020.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. CLAP: Learning audio concepts from natural language supervision. In *2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023. doi: 10.1109/ICASSP49357.2023.10095889.

Zach Evans, Julian D Parker, CJ Carr, Zack Zukowski, Josiah Taylor, and Jordi Pons. Stable Audio Open. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

Kaisi Guan, Zhengfeng Lai, Yuchong Sun, Peng Zhang, Wei Liu, Kieran Liu, Meng Cao, and Ruihua Song. ETVA: Evaluation of text-to-video alignment via fine-grained question generation and answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21299–21309, 2025.

Yuxin Guo, Teng Wang, Yuying Ge, Shijie Ma, Yixiao Ge, and Wei Zou. AudioStory: Generating long-form narrative audio with large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 37735–37744, 2026.

Zhifang Guo, Jianguo Mao, Rui Tao, Long Yan, Kazushige Ouchi, Hong Liu, and Xiangdong Wang. Audio generation with multiple conditional diffusion model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 18153–18161, 2024.

Jiarui Hai, Yong Xu, Hao Zhang, Chenxing Li, Helin Wang, Mounya Elhilali, and Dong Yu. EzAudio: Enhancing Text-to-Audio Generation with Efficient Diffusion Transformer. In *Proc. Interspeech 2025*, pp. 4233–4237, 2025.

Kevin David Hayes, Micah Goldblum, Vikash Sehwag, Gowthami Somepalli, Ashwinee Panda, and Tom Goldstein. FineGRAIN: Evaluating failure modes of text-to-image models with vision language model judges. In *Advances in Neural Information Processing Systems 38*, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/accaab47784b91d9d381868eb638be0a-Abstract-Datasets_and_Benchmarks_Track.html.

Xuan He, Dongfu Jiang, Ge Zhang, Max Ku, Achint Soni, Sherman Siu, Haonan Chen, Abhranil Chandra, Ziyan Jiang, Aaran Arulraj, Kai Wang, Quy Duc Do, Yuansheng Ni, Bohan Lyu, Yaswanth Narsupalli, Rongqi Fan, Zhiheng Lyu, Bill Yuchen Lin, and Wenhu Chen. VideoScore: Building automatic metrics to simulate fine-grained human feedback for video generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 2105–2123, 2024. doi: 10.18653/v1/2024.emnlp-main.127.

Yuhang He, Yash Jain, Xubo Liu, Andrew Markham, and Vibhav Vineet. RiTTA: Modeling Event Relations in Text-to-Audio Generation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 3497–3511, 2025.

Yuhang He, He Liang, Yash Jain, Andrew Markham, and Vibhav Vineet. Aurelius: Relation Aware Text-to-Audio Generation at Scale. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=LAYCYiIgZ1.

Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, R. Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron J. Weiss, and Kevin W. Wilson. CNN architectures for large-scale audio classification. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 131–135, 2017. doi: 10.1109/ICASSP.2017.7952132.

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

Rongjie Huang, Jiawei Huang, Dongchao Yang, Yi Ren, Luping Liu, Mingze Li, Zhenhui Ye, Jinglin Liu, Xiang Yin, and Zhou Zhao. Make-An-Audio: Text-To-Audio Generation with Prompt-Enhanced Diffusion Models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 13916–13932, 2023b. URL https://proceedings.mlr.press/v202/huang23i.html.

Chia-Yu Hung, Navonil Majumder, Zhifeng Kong, Ambuj Mehrish, Amir Ali Bagherzadeh, 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. *arXiv preprint arXiv:2412.21037*, 2024.

Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. Fréchet audio distance: A reference-free metric for evaluating music enhancement algorithms. In *Interspeech 2019*, pp. 2350–2354, 2019. doi: 10.21437/Interspeech.2019-2219.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. AudioCaps: Generating captions for audios in the wild. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 119–132, 2019.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. Prometheus 2: An open source language model specialized in evaluating other language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 4334–4353, 2024. doi: 10.18653/v1/2024.emnlp-main.248.

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. doi: 10.1109/TASLP.2020.3030497.

Khaled Koutini, Jan Schlüter, Hamid Eghbal-zadeh, and Gerhard Widmer. Efficient training of audio transformers with patchout. In *Interspeech 2022*, pp. 2753–2757, 2022. doi: 10.21437/Interspeech.2022-227.

Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi. AudioGen: Textually guided audio generation. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=CYK7RfcOzQ4.

Chun-Yi Kuan, Kai-Wei Chang, and Hung-yi Lee. AQAScore: Evaluating Semantic Alignment in Text-to-Audio Generation via Audio Question Answering. *arXiv preprint arXiv:2601.14728*, 2026.

Saksham Singh Kushwaha and Yapeng Tian. VinTAGe: Joint video and text conditioning for holistic audio generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 13529–13539, 2025.

Yongqi Leng, Renren Jin, Yue Chen, Zhuowen Han, Ling Shi, Jianxiang Peng, Lei Yang, Juesi Xiao, and Deyi Xiong. Praetor: A fine-grained generative LLM evaluator with instance-level customizable evaluation criteria. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10386–10418, 2025. doi: 10.18653/v1/2025.acl-long.513.

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*, 2026a.

Xiquan Li, Junxi Liu, Yuzhe Liang, Zhikang Niu, Wenxi Chen, and Xie Chen. MeanAudio: Fast and Faithful Text-to-Audio Generation with Mean Flows. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14378–14393, 2026b. URL https://aclanthology.org/2026.acl-long.654/.

Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. In *European Conference on Computer Vision*, pp. 366–384, 2024. doi: 10.1007/978-3-031-72673-6_20.

Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley. AudioLDM: Text-to-Audio Generation with Latent Diffusion Models. In *International Conference on Machine Learning*, pp. 21450–21474. PMLR, 2023.

Haohe Liu, Yi Yuan, Xubo Liu, Xinhao Mei, Qiuqiang Kong, Qiao Tian, Yuping Wang, Wenwu Wang, Yuxuan Wang, and Mark D. Plumbley. AudioLDM 2: Learning Holistic Audio Generation With Self-Supervised Pretraining. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:2871–2883, 2024. doi: 10.1109/TASLP.2024.3399607.

Navonil Majumder, Chia-Yu Hung, Deepanway Ghosal, Wei-Ning Hsu, Rada Mihalcea, and Soujanya Poria. Tango 2: Aligning Diffusion-based Text-to-Audio Generations through Direct Preference Optimization. In *ACM Multimedia 2024*, 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.

Qwen Team. Qwen3.5: Towards Native Multimodal Agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Xian Shi, Xiong Wang, Zhifang Guo, Yongqi Wang, Pei Zhang, Xinyu Zhang, Zishan Guo, Hongkun Hao, Yu Xi, Baosong Yang, et al. Qwen3-ASR Technical Report. *arXiv preprint arXiv:2601.21337*, 2026.

Ye Tao, Lupeng Liu, Xuenan Xu, Jiasun Feng, Jiarui Wang, Ying Qin, Shuiyang Mao, Wei Liu, and Shuai Wang. Foley-Omni: A Unified Multimodal Generation Model from Task-Level Audio Synthesis to Complete Video Soundtrack Generation. *arXiv preprint arXiv:2606.03672*, 2026.

Zeyue Tian, Yizhu Jin, Zhaoyang Liu, Ruibin Yuan, Xu Tan, Qifeng Chen, Wei Xue, and Yike Guo. AudioX: A Unified Framework for Anything-to-Audio Generation. *arXiv preprint arXiv:2503.10522*, 2025.

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 assessment for speech, music and sound. In *2025 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 1–8, 2025. doi: 10.1109/ASRU65441.2025.11434623.

Apoorv Vyas, Bowen Shi, Matthew Le, Andros Tjandra, Yi-Chiao Wu, Baishan Guo, Jiemin Zhang, Xinyue Zhang, Robert Adkins, William Ngan, Jeff Wang, Ivan Cruz, Bapi Akula, Akinniyi Akinyemi, Brian Ellis, Rashel Moritz, Yael Yungster, Alice Rakotoarison, Liang Tan, Chris Summers, Carleigh Wood, Joshua Lane, Mary Williamson, and Wei-Ning Hsu. Audiobox: Unified Audio Generation with Natural Language Prompts. *arXiv preprint arXiv:2312.15821*, 2023.

Chenxi Wang, Yusheng Dai, Lei Sun, Jun Du, and Jianqing Gao. AudioAtlas: A Comprehensive and Balanced Benchmark Towards Movie-Oriented Text-to-Audio Generation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 13266–13272, 2025a.

Hui Wang, Cheng Liu, Junyang Chen, Haoze Liu, Yuhang Jia, Shiwan Zhao, Jiaming Zhou, Haoqin Sun, Hui Bu, and Yong Qin. TTA-Bench: A comprehensive benchmark for evaluating text-to-audio models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 33512–33520, 2026.

Zehan Wang, Ke Lei, Chen Zhu, Jiawei Huang, Sashuai Zhou, Luping Liu, Xize Cheng, Shengpeng Ji, Zhenhui Ye, Tao Jin, et al. T2A-Feedback: Improving Basic Capabilities of Text-to-Audio Generation via Fine-grained AI Feedback. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 23535–23547, 2025b. doi: 10.18653/v1/2025.acl-long.1147. URL https://aclanthology.org/2025.acl-long.1147/.

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 *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Zeyu Xie, Xuenan Xu, Zhizheng Wu, and Mengyue Wu. AudioTime: A temporally-aligned audio-text benchmark dataset. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 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.

Jinlong Xue, Yayue Deng, Yingming Gao, and Ya Li. Auffusion: Leveraging the power of diffusion and large language models for text-to-audio generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:4700–4712, 2024.

Dongchao Yang, Jinchuan Tian, Xu Tan, Rongjie Huang, Songxiang Liu, Haohan Guo, Xuankai Chang, Jiatong Shi, Sheng Zhao, Jiang Bian, Zhou Zhao, Xixin Wu, and Helen M. Meng. UniAudio: Towards universal audio generation with large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 56422–56447. PMLR, 2024. URL https://proceedings.mlr.press/v235/yang24x.html.

Alon Ziv, Itai Gat, Gael Le Lan, Tal Remez, Felix Kreuk, Jade Copet, Alexandre Défossez, Gabriel Synnaeve, and Yossi Adi. Masked audio generation using a single non-autoregressive transformer. In *International Conference on Learning Representations*, volume 2024, pp. 43249–43268, 2024.

A Additional Dataset Details

A.1 Data Collection and Annotation Pipeline

Source videos are collected from major video-sharing platforms, including YouTube, TikTok, and Bilibili, and span movie/television content and everyday environmental recordings. Scene- and event-oriented keyword queries retrieve candidate segments, from which we extract target 10-second windows. Thus, "short clip" refers operationally to this target duration rather than an unspecified qualitative notion of shortness.

Gemini-2.5-Pro generates an initial scene description and candidate structured annotation for each segment. Trained annotators then listen to the extracted audio, correct component events and attribute fields, and retain a clip only when its scene and event nodes are audibly verifiable. Ambiguous, unsupported, or unreliable cases are removed. This model-assisted procedure reduces annotation cost, while human verification determines the final retained content. The resulting collection combines professionally produced soundscapes with everyday acoustic scenes and may retain platform, production, and selection biases.

Structured records and provenance. Each retained record stores a stable clip/row identifier, caption, scene description, verified event list, component counts and composition, complexity metadata, source platform and source index, and clip timestamps. Event-level fields and missing-value handling follow the schema summarized in the main paper and detailed in Appendix A.4.

Planned release. Upon paper acceptance, we will release the exact search-query list, source platform, video identifier and URL, clip timestamp, and filtering metadata, together with the text prompts, structured annotations, derived rubrics, evaluation code, source indices, and corresponding raw audio clips where redistribution is permitted. For any clip that cannot be redistributed, we will provide its source index, timestamp, and reconstruction metadata. These materials document the benchmark construction and support reuse by new model developers.

A.2 Complexity Score Definition

We characterize sample complexity using two complementary dimensions: *event density* and *structural complexity*. While event density measures annotation-derived semantic event load, structural complexity characterizes the diversity and clip-level composition of active audio categories. Together, they capture two complementary aspects of generation complexity.

Event Density. Event density measures annotation-derived semantic event load in a soundscape. For clip i, it is defined as

$$D_i = N_{\mathrm{SFX},i} + I_{\mathrm{Speech},i} + I_{\mathrm{BGM},i}. \tag{3}$$

This is an annotation-derived count, not an acoustic-onset count. Specifically, $N_{\mathrm{SFX},i}$ is the number of SFX event nodes in the verified structured annotation. Each node represents one semantically coherent source–action occurrence. Multiple emissions encoded within the same node (e.g., a dog barks once, twice, or continuously, with repetition stored as an attribute) contribute one count; they contribute multiple counts only when annotators represent them as distinct event nodes because they are temporally separable or differ in source or action. Thus, the operational boundary is the structured event node rather than every audible onset. In contrast, $I_{\mathrm{Speech},i} \in \{0, 1\}$ and $I_{\mathrm{BGM},i} \in \{0, 1\}$ collapse all nodes in those categories to clip-level presence indicators.

This asymmetric definition reflects the different temporal characteristics of the three audio categories. SFX typically consists of discrete and countable event instances, whereas speech usually forms a continuous dialogue layer and BGM serves as a sustained background layer. Counting individual speech utterances or music segments would depend heavily on annotation granularity and could artificially inflate the density score. We therefore count SFX at the event-instance level while representing speech and BGM solely by their clip-level presence. Note that D_i does not represent physical loudness or duration; it only measures the number of annotated semantic components.

Structural Complexity. Structural complexity measures how diverse the composition of a soundscape is, independent of the number of event instances. Let $K_i = I_{\mathrm{SFX},i} + I_{\mathrm{Speech},i} + I_{\mathrm{BGM},i}$ be the number of active audio-component categories among Speech, SFX, and BGM. Then

$$S_i = K_i + I_{\mathrm{Speech},i}\,\mathbb{I}[I_{\mathrm{SFX},i} + I_{\mathrm{BGM},i} \geq 1], \tag{4}$$

where $\mathbb{I}[\cdot]$ is the indicator function. The interaction bonus applies specifically when speech co-occurs with SFX or BGM. We treat speech as a distinct component because, unlike non-speech audio components, speech generation requires not only acoustic realization but also preservation of linguistic content and speaker-related characteristics. Therefore, S_i ranges from 1 to 4, where larger values indicate more diverse audio-component combinations.

Unlike event density, structural complexity does not depend on the number of SFX instances. Instead, it captures the diversity of active audio categories and whether heterogeneous audio components must be generated jointly. We explicitly consider speech-related co-occurrence because unified speech-aware audio generation introduces additional requirements beyond non-speech sound synthesis, including linguistic content preservation, speaker identity modeling, and coordination between vocal and non-vocal components. Therefore, soundscapes containing speech together with SFX or BGM typically require joint modeling of semantic, acoustic, and linguistic factors.

The proposed structural complexity is designed as a practical benchmark-oriented complexity heuristic rather than a claim that speech is universally more complex than other audio categories. The co-occurrence term characterizes category-level composition only; it does not explicitly model temporal overlap, event ordering, causal relations, or semantic interactions among components. Accordingly, the proposed complexity characterization captures clip-level audio-category composition rather than explicit temporal overlap, event ordering, causal relations, or semantic interactions among individual events. References to compositional difficulty in this paper should therefore be interpreted under this benchmark-specific Density–Structure characterization.

Complexity Score. Since D_i and S_i are measured on different scales, we first normalize them using z-score normalization and then compute the overall complexity score as

$$C_i = Z(D_i) + Z(S_i) = \frac{D_i - \mu_D}{\sigma_D} + \frac{S_i - \mu_S}{\sigma_S}, \tag{5}$$

where μ_D, σ_D and μ_S, σ_S are the benchmark-wide means and standard deviations of D_i and S_i, respectively.

Finally, samples are partitioned into three complexity levels according to

$$\begin{cases} \text{Easy}, & C_i \leq -1, \\ \text{Medium}, & -1 < C_i < 1, \\ \text{Hard}, & C_i \geq 1. \end{cases}$$

The thresholds are selected based on standardized complexity scores to separate low-, medium-, and high-complexity samples while maintaining sufficient samples in each subset. This formulation enables complexity-aware evaluation by jointly considering annotation-derived event load and compositional diversity of structured soundscapes.

Complementarity Analysis. Across all 2,258 samples, event density and structural complexity exhibit only weak correlation (Pearson $r = 0.221$; Spearman $\rho = 0.153$), indicating that the two metrics capture complementary aspects of audio-scene complexity. Specifically, event density measures annotation-derived semantic event load, whereas structural complexity characterizes the diversity and co-occurrence patterns of audio components.

A.3 Data Statistics

To further demonstrate the diversity and complexity of the proposed AudioScape-TTA benchmark, we provide additional annotation examples and dataset statistics in the supplementary material. Ta-

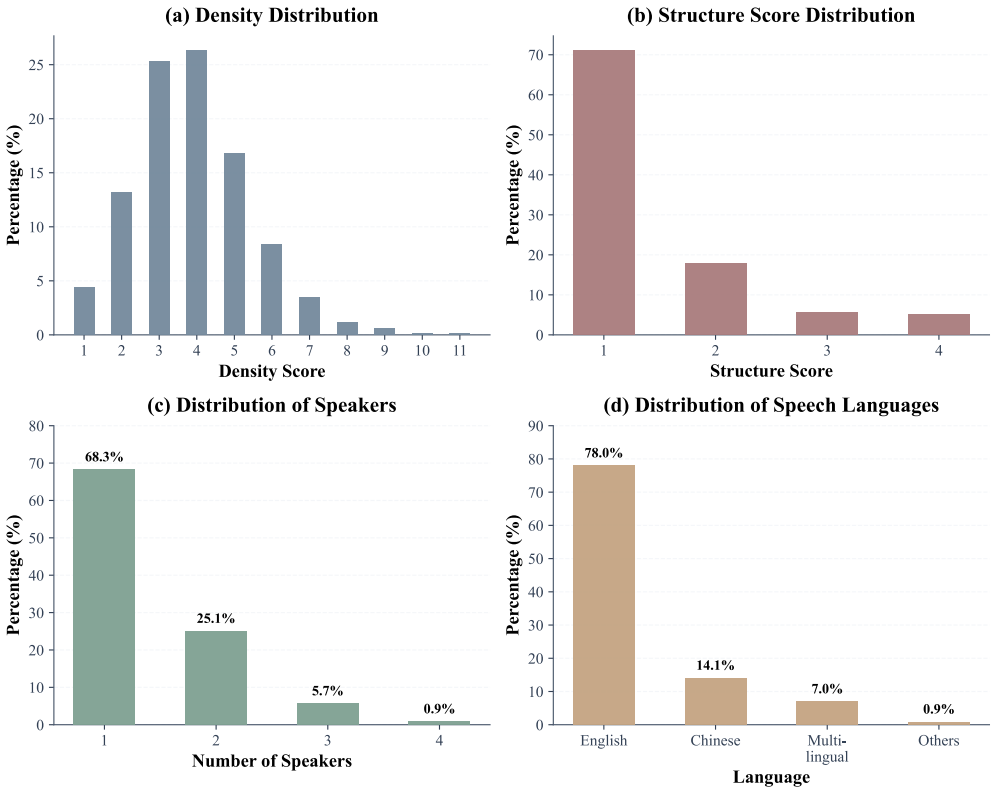

Figure 4: **Statistics of the proposed AudioScape-TTA benchmark.** (a) The D_i distribution reflects the combination of discrete SFX-instance counts and clip-level presence indicators for speech and BGM. (b) The S_i distribution characterizes the complexity and organization of multi-event soundscapes. (c) and (d) summarize speaker-count and speech-language distributions over the 227 speech clips with complete metadata for both fields; percentages in each panel are normalized within this subset. The remaining Speech-category clips are retained in the benchmark but omitted from these two metadata summaries. These statistics describe the audio-component diversity and annotated composition of the benchmark.

Table 5: Composition statistics of the Easy, Medium, and Hard subsets. Avg. SFX denotes the mean number of annotated SFX instances per clip. Speech and BGM indicate the percentages of clips with at least one final rubric in the corresponding category. Multi-comp. denotes the percentage of clips covered by final rubrics from at least two audio categories among Speech, SFX, and BGM. Avg. Rubrics is the average number of QA rubrics per clip in the benchmark.

Subset	# Samples	Avg. SFX	Speech (%)	BGM (%)	Multi-comp. (%)	Avg. Rubrics
Easy	732	2.36	0.00	10.52	0.14	7.62
Medium	1,103	3.91	0.91	25.57	26.11	11.92
Hard	423	4.48	57.21	56.50	85.58	16.51

ble 6 presents representative samples with different complexity levels. Each row shows the annotation fields used in that example together with a selected subset of questions derived from those fields, typed as event presence, attributes, or speech content. The illustrated clips progress from an SFX-only water scene to a cafe scene with background music, then to a mixture of speech, score, and sound effects.

Moreover, Figure 4 summarizes the benchmark statistics from multiple perspectives. The density and structure score distributions characterize annotation-derived event load and clip-level audio-

component composition. The speaker and language distributions further demonstrate the diversity of speech-related scenarios. Together, these results highlight that AudioScape-TTA covers diverse audio components and challenging compositional sound environments, providing a structured testbed for evaluating TTA generation models.

Table 5 summarizes the composition of the Easy, Medium, and Hard subsets. As the difficulty level increases, audio scenes become progressively more complex, with more annotated SFX events, higher proportions of speech and background music, more multi-component compositions, and a larger number of QA rubrics per clip. These statistics show that the strata differ in content composition and rubric count as well as their defined complexity score; they do not isolate an independent effect of complexity.

A.4 End-to-End Annotation and Evaluation Specification

Structured annotation schema. A benchmark *instance* is one source clip paired with its caption. Its record contains a stable clip/row identifier, the caption, a scene description, and an event list. Each event has an `event_id`, a record type (Scene, SFX, Speech, or BGM), normalized event text, and applicable attributes such as source/action, acoustic quality, temporal or spatial context, and repetition. Speech events additionally contain speaker descriptors and, when specified, the verbatim target utterance. Clip-level fields store component counts, component composition, event density D_i, structural complexity S_i, standardized complexity C_i, difficulty bin, and source provenance (platform/source index and clip timestamps). Unknown fields are left empty rather than inferred. Model-generated drafts are retained only after human verification and correction.

Rubric construction and terminology. A *benchmark instance* is one audio–text pair together with its verified structured annotation. A *semantic rubric* is a field-level requirement rewritten as a yes/no question derived from an annotated event-presence target, attribute field, or target utterance and linked to the corresponding instance and `event_id`. An attribute field may contain coordinated descriptors, which receive one joint binary decision rather than separate decisions. Examples include event presence ("Is a dog barking audible?"), an event attribute ("Does the dog sound distant?"), and specified speech content ("Does the speaker say 'The store is closed'?"), rather than a vague judgment of overall similarity. For instance i, $\mathcal{R}_i$ denotes its complete set of semantic rubrics; "instance-level rubric set" refers only to this grouping and is not a second rubric class. The deterministic extractor emits one rubric for each eligible event-presence target, selected attribute, and target utterance. Each rubric belongs to exactly one instance and one annotated event. Multiple rubrics may share an `event_id` and supporting evidence—for example, presence and distance questions for the same dog-barking event—but they target separately stored fields and are scored separately. Exact duplicate targets are removed, but semantically overlapping presence, source, and action targets may remain and therefore receive separate weight in the rubric-micro average.

Rubrics are constructed as follows: (1) the deterministic extractor reads each verified event and eligible attribute; (2) it creates a target record with `qa_type`, component category, event identifier, target text, attribute field/value, and evidence; (3) an LLM rewrites only that fixed target into a concise yes/no question; and (4) validators check schema completeness, event linkage, answer format, target preservation, and leakage. A failed rewrite is replaced by a deterministic template. The LLM never chooses the reference answer or introduces a new target. Positive annotation-derived questions use choices `A: Yes` and `B: No`, with reference answer `A`.

Evaluator interface and aggregation. For an event-presence or event-attribute rubric $r \in \mathcal{R}_i$, Qwen3-Omni receives the generated audio x_i, question q_{ir}, and choices $\{$`A: Yes`, `B: No`$\}$, and returns $\hat{a}_{ir} \in \{A, B\}$ together with the raw response and, when available, option log-probabilities. For a speech-content rubric, Qwen3-ASR first transcribes x_i; target-content coverage R_{ir} is then computed against the annotated utterance, and $s_{ir} = \mathbb{I}[R_{ir} \geq 0.60]$. For event-presence and event-attribute rubrics, $s_{ir} = 1$ when $\hat{a}_{ir} = A$ and $s_{ir} = 0$ otherwise. For instance i, let $N_i = |\mathcal{R}_i|$. Its instance-level score is $\mathrm{SR}_i = N_i^{-1} \sum_{r \in \mathcal{R}_i} s_{ir}$. The primary reported Overall SR is the rubric-micro average $\mathrm{SR}(\mathcal{Q})$ defined in Appendix B. It weights each field-level rubric equally, rather than each event or instance equally; consequently, an event with several distinct attribute rubrics contributes several separately scored requirements. Given option log-probabilities $\ell_{A,ir}, \ell_{B,ir}$, evalu-

ator confidence is $c_{ir} = \max(q_{A,ir}, q_{B,ir})$, where $q_{k,ir} = \exp(\ell_{k,ir})/(\exp(\ell_{A,ir}) + \exp(\ell_{B,ir}))$. Speech-content coverage and its ASR preprocessing are defined separately below.

Core complexity formula and counting rule. Event density $D_i = N_{\mathrm{SFX},i} + I_{\mathrm{Speech},i} + I_{\mathrm{BGM},i}$ is annotation-derived rather than onset-derived: it counts verified SFX event nodes and clip-level Speech/BGM presence, while S_i captures component diversity and interaction/compositional structure under the rules in the preceding subsection. Repeated emissions within one SFX node count once; only distinct annotated nodes count separately. Since D_i and S_i have different scales, each is standardized over the benchmark and combined with equal weight:

$$C_i = Z(D_i) + Z(S_i) = \frac{D_i - \mu_D}{\sigma_D} + \frac{S_i - \mu_S}{\sigma_S}.$$

Difficulty bins are assigned from this final score as detailed below.

Worked example: caption to score. Consider the caption: *"On a quiet street, a distant dog barks twice while a man says 'The store is closed,' over soft piano music."* The structured record contains the scene `quiet street`; e_1 = `dog barking` (SFX; `distance=distant`; `repetition=two`); e_2 = `male speech` (Speech; target utterance `The store is closed`); and e_3 = `soft piano music` (BGM). Here the repeated barks remain one event, so $D_i = 3$. For a compact scoring illustration, we select four rubrics from the potentially larger set produced by deterministic extraction: r_1: "Is dog barking audible?", r_2: "Does the dog sound distant?", r_3: "Is soft piano music audible?", and r_4: "Does the speaker say 'The store is closed'?"; each has reference answer A. Suppose Qwen3-Omni returns A for r_1–r_3. For r_4, Qwen3-ASR first produces a transcript; the subsequent clipped multiset overlap matches two of the four target word units, giving $R_{i4} = 2/4 = 0.50 < 0.60$ and therefore $s_{i4} = 0$. The resulting decisions are $(s_{i1}, s_{i2}, s_{i3}, s_{i4}) = (1, 1, 1, 0)$. Within this simplified four-rubric illustration, $N_i = 4$ and $\mathrm{SR}_i = 3/4 = 0.75$. Its three active categories give $K_i = 3$, and speech co-occurrence contributes the interaction bonus, so $S_i = 4$ and $C_i = Z(3) + Z(4)$ using the benchmark-wide means and standard deviations. This example separates the annotation target, rubric wording, evaluator decision, semantic score, and complexity metadata.

A.5 Semantic Rubric Construction Details

We construct rubric pairs from structured event-level annotations using a deterministic target-generation pipeline followed by natural-language rewriting. The semantic target of each rubric pair is derived directly from the annotation, while the LLM is only used to rewrite the target into a concise yes/no question. Therefore, the LLM does not decide the answer or introduce new semantic targets. All rubric pairs use the same two-choice format, `A: Yes` and `B: No`, and the reference answer is set to `A` because each question is generated from a positive annotated event or attribute.

Rubric Types. We generate three types of rubric pairs:

- **Event presence**: typed from an event-presence field and intended to check whether an annotated scene or event is present; the question may still include descriptors from that field.
- **Event attribute**: typed from an attribute field (source, material, intensity, temporal or spatial context, instrument, emotion, speaker, or language).
- **Speech content**: verifies whether the spoken content of a speech event is preserved.

Structured rubric schema. The stored key `component` identifies the record type (Scene, SFX, Speech, or BGM). Each QA rubric pair stores both the natural-language question and its structured evidence:

```
{
  "qa_id": string,
  "qa_type": "event_presence | event_attribute | speech_content",
  "answer_type": "two_choice",
  "question": string,
  "choices": {"A": "Yes", "B": "No"},
  "answer": "A",
  "component": "SFX | BGM | Speech | Scene",
  "event_id": int,
  "event_text": string,
  "attribute_field": string,
  "attribute_value": string,
  "evidence": object
}
```

Table 6: Representative structured annotations and selected questions across complexity levels. The questions shown are a selected subset for the illustrated clip and correspond to the displayed annotation fields.

Easy
Annotation. *[Scene] Outdoor water source environment; [SFX] Gurgling and trickling of flowing water; [SFX] Water pouring out of the spout.*
Selected questions. Event presence: *1. Is there an outdoor water source environment? 2. Is there water pouring out of the spout?* **Event attributes:** *1. Can gurgling and trickling be heard?*

Medium
Annotation. *[Scene] Indoor cafe/kitchen environment; [BGM] Gentle piano music; [SFX] Grinding or blending machinery sound.*
Selected questions. Event presence: *1. Is there an indoor cafe/kitchen environment? 2. Is there a grinding or blending machinery sound?* **Event attributes:** *1. Is the music gentle in tone? 2. Is the music played on a piano?*

Hard
Annotation. *[Scene] Indoor closed environment; [BGM] Tense atmospheric dramatic score with synthesized tones; [SFX] Male speaking; Baby crying; [Speaker A, elderly male, speaking calmly]: "Tell the story. He said he could hear that baby cry just as plain. And um..."*
Selected questions. Event presence: *1. Is there an indoor closed environment? 2. Is there a male speaking? 3. Is there a baby crying?* **Speech content:** *Does the speech include the exact phrase 'tell the story. He said he could hear that baby cry just as plain. And um...'?* **Event attributes:** *1. Is the speaker elderly? 2. Is the speech delivered calmly? 3. Is the genre of the score atmospheric dramatic? 4. Is the mood of the score tense? 5. Does the score include synthesized tones?*

Target generation. For each annotated event, we first construct structured QA targets. For non-speech events, we create an `event_presence` target. For all events, we select informative attributes according to component-specific priority lists and create `event_attribute` targets. Generic or uninformative values are removed. For speech events, if the `action_or_content` field contains explicit spoken content, we create a `speech_content` target.

The attribute priorities are:

- **Speech**: speaker gender, age, role, language, emotion, style, prosody, source.
- **SFX**: source, action/content, intensity, spatial property, temporal pattern, material, style.
- **BGM**: genre, mood, instrument, style, intensity, temporal pattern, spatial property.
- **Scene**: spatial property, temporal pattern, intensity, mood, style.

Question rewriting. Structured targets are rewritten into natural questions using an LLM with strict constraints. The model is instructed to preserve the `qa_id`, `qa_type`, event ID, component, and attribute field; not to add or remove targets; and not to include answer choices in the question. The prompt also prevents exposing internal annotation terms such as `actor_or_source`

or `action_or_content`. For `speech_content`, the rewritten question must quote the exact spoken content or a long exact phrase.

The core rewriting instruction is:

```
Rewrite each structured QA target as one natural yes/no question for
    evaluating a generated audio clip.

Requirements:
- Answerable solely from the audio.
- Preserve the original target exactly.
- Do not add or omit information.
- Do not expose internal field names.
- Keep the question concise and natural.

Return strict JSON:
{
  "questions": [
    {
      "qa_id": "same id as input",
      "question": "natural yes/no question"
    }
  ]
}
```

Leakage prevention and fallback templates. To reduce question leakage, we apply rule-based checks after rewriting. A question is replaced by a safe template if it copies the full event text into an attribute question, repeats the attribute value excessively, exposes internal field names, or contains forbidden words such as `reference`, `annotation`, `caption`, or `metadata`. We use component-aware fallback templates, *e.g.,*

```
    Event presence:  Is there a sound event described as {event_text}?
BGM genre:       Is there {value} background music in the audio?
SFX material:    Does the sound have a {value} quality?
Speech language: Is the speech in {value}?
Speech content:  Does the speaker say "{spoken_content}" in the audio?
```

Validation and filtering. Finally, each generated question is validated against its structured evidence. The validation checks whether the question refers to the same event, component, and attribute, whether the attribute value is correctly represented, and whether unsupported details are introduced. These checks target evidence alignment and malformed wording; they do not guarantee that every retained question is purely acoustic, source-explicit, or free of overlap with other fields. Questions with severe validation errors are discarded. We also remove duplicated questions, questions with internal field wording, invalid `Scene` event-attribute questions, and malformed questions. After filtering, the final benchmark contains **25,707** rubric pairs over **2,258** audio-text instances.

B Detailed Metric Definitions

This section defines the fine-grained metrics. SR and SCCA are expressed as percentages in both the formulas and tables. Coverage R_{ir} remains a fraction in $[0, 1]$; word error rate (WER) and character error rate (CER) are reported as unscaled edit-error ratios.

Semantic Satisfaction Rate. For rubric $r \in \mathcal{R}_i$ attached to instance i, Qwen3-Omni-Instruct receives the generated audio x_i, fixed question q_{ir}, and choices $\{\texttt{A: Yes}, \texttt{B: No}\}$, and returns $\hat{a}_{ir} \in \{A, B\}$. For event-presence and event-attribute rubrics, $s_{ir} = 1$ when $\hat{a}_{ir} = A$ (Yes) and $s_{ir} = 0$ otherwise. All main-benchmark targets are positive requirements derived from annotations.

Table 7: Model metadata and the evaluator-confidence diagnostics accompanying Table 2. Confidence is shown as a percentage and is an uncalibrated binary-option preference, not a probability of semantic correctness. The generation configurations are detailed in Appendix D.

Model	Year	Params	Overall Conf.	Event Conf.	Attribute Conf.
AudioLDM 2	2023	1.5B	93.85	94.05	93.71
Make-An-Audio 2	2023	937M	92.99	93.26	92.80
AudioGen	2023	1.5B	93.62	93.99	93.35
Tango 2	2024	866M	93.33	93.97	92.88
TangoFlux	2024	515M	94.27	94.56	94.07
EzAudio	2025	875M	94.14	94.38	93.97
MAGNet	2024	1.5B	92.44	92.38	92.48
Stable Audio Open	2025	1.1B	93.59	93.98	93.32
MMAudio	2025	1.03B	94.15	94.52	93.89
Foley-Omni	2026	5.5B	94.05	94.66	93.61
Omni2Sound	2026	1.33B	94.47	94.86	94.20
AudioStory	2026	4.36B	93.04	93.36	92.80
Dasheng AudioGen	2026	2.19B	94.17	94.27	94.09

For a speech-content rubric r of instance i, the satisfaction indicator uses the ASR-based coverage criterion:

$$s_{ir} = \mathbb{I}[R_{ir} \geq 0.60].$$

For any set $\mathcal{S}$ of instance–rubric pairs, the satisfaction rate is

$$\mathrm{SR}(\mathcal{S}) = \frac{100}{|\mathcal{S}|} \sum_{(i,r)\in\mathcal{S}} s_{ir}, \tag{6}$$

$$\text{Overall SR} = \mathrm{SR}(\mathcal{Q}), \qquad \text{Event SR} = \mathrm{SR}(\mathcal{Q}_{\text{event}}),$$
$$\text{Attribute SR} = \mathrm{SR}(\mathcal{Q}_{\text{attr}}). \tag{7}$$

Here $\mathcal{Q}$ is the complete set of instance–rubric pairs, while $\mathcal{Q}_{\text{event}}$ and $\mathcal{Q}_{\text{attr}}$ are its event-presence and event-attribute subsets. Overall SR is a micro-average: it weights requirements, not instances or components equally.

Evaluator confidence. For event-presence and event-attribute rubric r of instance i, we additionally report an uncalibrated evaluator-confidence diagnostic. Let $\ell_{A,ir}$ and $\ell_{B,ir}$ be the log-probabilities assigned to the two answer options. We normalize them over the binary choice set:

$$q_{A,ir} = \frac{\exp(\ell_{A,ir})}{\exp(\ell_{A,ir}) + \exp(\ell_{B,ir})}, \quad q_{B,ir} = \frac{\exp(\ell_{B,ir})}{\exp(\ell_{A,ir}) + \exp(\ell_{B,ir})}. \tag{8}$$

The per-rubric confidence is $c_{ir} = \max(q_{A,ir}, q_{B,ir})$. For a set $\mathcal{S}$ of instance–rubric pairs, Table 7 reports $100|\mathcal{S}|^{-1} \sum_{(i,r)\in\mathcal{S}} c_{ir}$. This quantity measures the evaluator's relative preference between the two candidate answers only. It is not calibrated as a probability of semantic correctness, is not an audio-quality metric, and should not be compared directly across different evaluator models.

Speech-content preprocessing. Speech content is evaluated by comparing the Qwen3-ASR transcript with the annotated target text. Before scoring, we remove a leading bracketed tag when present, lowercase Latin text, and collapse consecutive whitespace. Punctuation does not form an evaluation unit. We retain numbers, repeated words, and filler expressions; no stop-word removal, duplicate suppression, or prefix-based shortcut is applied.

We extract two unit types from the normalized text:

- **Word units**: lowercased Latin alphanumeric tokens, optionally containing internal hyphens or apostrophes;
- **Character units**: individual Chinese CJK characters.

This mixed-unit formulation supports English, Chinese, and code-switched speech under a single protocol.

Mixed-unit coverage. Let t and $\hat{t}$ denote the normalized target text and ASR transcript, respectively. Let $c_t^w(u)$ and $c_{\hat{t}}^w(u)$ be the occurrence counts of word unit u in the target and transcript; $c_t^c(v)$ and $c_{\hat{t}}^c(v)$ are defined analogously for Chinese character unit v. We use clipped multiset overlap:

$$m_w = \sum_u \min\left(c_t^w(u), c_{\hat{t}}^w(u)\right), \tag{9}$$

$$m_c = \sum_v \min\left(c_t^c(v), c_{\hat{t}}^c(v)\right). \tag{10}$$

The total number of matched units is $m = m_w + m_c$. Let N_{ir} denote the total number of word and character units in the target of speech-content rubric r of instance i. We define target-content coverage as

$$R_{ir} = \frac{m}{N_{ir}}. \tag{11}$$

Clipped multiset overlap prevents a repeated transcript unit from matching multiple occurrences in the target. Importantly, this score is an order-agnostic target-unit coverage measure: it does not require the matched units to occur contiguously or in the same order as in the target, and it does not penalize additional transcript units. This coverage score is used only to determine SCCA@0.60; mixed-unit precision and F1 are not reported as benchmark metrics.

Speech-content coverage satisfaction rate. Our primary speech metric is speech-content coverage satisfaction rate at threshold τ, denoted as SCCA@τ. A speech-content pair (i, r) satisfies this criterion if its target-content coverage reaches the threshold:

$$s_{ir} = \mathbb{I}\left[R_{ir} \geq \tau\right]. \tag{12}$$

For the speech-content pair set $\mathcal{Q}_{\text{speech}}$ with $N_{\text{speech}} = |\mathcal{Q}_{\text{speech}}|$, we compute

$$\text{SCCA@}\tau = \frac{100}{N_{\text{speech}}} \sum_{(i,r)\in\mathcal{Q}_{\text{speech}}} s_{ir}. \tag{13}$$

We use $\tau = 0.60$ for the main comparison and vary it only in the threshold-sensitivity analysis. Consequently, SCCA@0.60 measures whether the transcript contains clipped matches for at least 60% of the target units under the mixed-unit coverage criterion. It is not an order-sensitive exact-transcription metric. The score pools clipped matches over Latin word units and Chinese character units; it does not use a maximum across unit types or any fixed-length prefix shortcut.

Supplementary WER and CER diagnostics. WER and CER are computed from the same Qwen3-ASR transcripts used for speech-content evaluation. In contrast to SCCA@0.60, which is a rubric-level binary target-coverage metric over $N_{\text{speech}} = 163$ speech-content rubrics, WER and CER are clip-level corpus transcription diagnostics. For unit type $t \in \{\text{word}, \text{character}\}$, we pool edit counts across the applicable clips and compute

$$E_t = \frac{n_t^{\text{sub}} + n_t^{\text{del}} + n_t^{\text{ins}}}{N_t^{\text{ref}}}, \tag{14}$$

where n_t^{sub}, n_t^{del}, and n_t^{ins} are the Levenshtein substitution, deletion, and insertion counts, and N_t^{ref} is the number of reference units. We report $\text{WER} = E_{\text{word}}$ and $\text{CER} = E_{\text{character}}$.

Consistent with the main-text statistics, the final benchmark contains 163 speech-content rubrics associated with 83 clips containing target utterances. For the supplementary clip-level transcription diagnostics, we construct a caption-derived reference transcript by concatenating quoted `[Speaker ...]` utterances in annotation order. This construction is available for 80 of the 83 target-utterance clips. The remaining three clips contain speech-related targets without a usable quoted utterance reference (e.g., indistinct or non-lexical vocal activity) and are excluded only from the WER/CER computation; they remain included in the speech-content benchmark and in SCCA@0.60.

In the final evaluation cache, WER is computed over $|\mathcal{J}_{\text{WER}}| = 69$ valid reference clips containing a total of 1,503 Latin word units, and CER is computed over $|\mathcal{J}_{\text{CER}}| = 11$ valid reference clips containing a total of 309 Chinese CJK character units. A code-switched clip may contribute to both sets when it contains both unit types. Because CER is computed on a smaller Chinese-reference subset, it should be interpreted as a complementary diagnostic with additional caution. The resulting WER and CER values are reported in Table 13.

Component, composition, and complexity subsets. For component-level analysis, rubric items are grouped according to the queried component, including scene, sound effects, background music, and speech. In particular, *Speech-component rubric satisfaction rate* denotes rubric satisfaction rate over rubrics associated with the speech component; it is distinct from SCCA@0.60, which evaluates the linguistic content of synthesized speech through ASR transcripts.

For composition analysis, samples are grouped by the set of audio components required by their prompts, such as Sound-only, Sound+Music, or Sound+Speech+Music. For complexity analysis, samples are grouped into Easy, Medium, and Hard subsets according to the benchmark's event-density and structural-complexity annotations. All subset results use the same s_{ir} definition as the overall rubric-based evaluation.

Complementary global metrics. We additionally report global quality and alignment metrics as complementary measurements. Audiobox-Aesthetic (Tjandra et al., 2025) provides content enjoyment (CE), content usefulness (CU), production complexity (PC), and production quality (PQ), where higher values indicate better predicted perceptual quality. $\mathrm{CLAP}_{\mathrm{MS}}$ follows the AV-Benchmark implementation (Cheng et al., 2025) and is the mean cosine similarity between paired text and audio embeddings from Microsoft CLAP (MS-CLAP) (Elizalde et al., 2023); higher values indicate stronger coarse semantic alignment. For the distribution-based metrics, we follow the AudioLDM definitions (Liu et al., 2023) and AV-Benchmark implementation (Cheng et al., 2025). Fréchet distance (FD) compares the feature distributions of generated and reference audio, using VGGish (Hershey et al., 2017; Kilgour et al., 2019), PANNs (Kong et al., 2020), or PaSST (Koutini et al., 2022) features as indicated. Mean Kullback–Leibler distance (KL) averages the divergence between paired generated/reference classifier outputs from PANNs or PaSST, whereas Inception Score (ISC) measures the classifiability and diversity of generated samples without comparing them with the reference distribution. Lower FD and KL are preferred, whereas higher ISC is preferred. These global metrics are not included in SR and are intended to complement, rather than replace, the proposed fine-grained semantic evaluation.

C Human Evaluation Protocol

To validate the effectiveness of our automatic evaluation framework, we conduct a human evaluation study on a subset of AudioScape-TTA. We randomly select 10 prompts covering different audio-component compositions and complexity levels, and collect generated audio from all 13 evaluated TTA models, resulting in 130 audio samples. The identities of the models are anonymized, and within each prompt the model-to-label assignments are randomly shuffled before evaluation.

Each audio sample is independently rated by 26 participants. For each sample, participants evaluate three aspects: *event completeness*, *attribute satisfaction*, and *audio quality*. Event completeness measures whether the required sound events described in the prompt are present in the generated audio. Attribute satisfaction evaluates whether fine-grained acoustic properties, such as sound source characteristics, temporal patterns, and other specified attributes, are correctly realized. Audio quality measures overall perceptual quality, including naturalness, coherence, and generation artifacts.

Participants are provided with the original text prompt and the generated audio, and rate each dimension independently using a 5-point Likert scale. For semantic dimensions (*event completeness* and *attribute satisfaction*), a higher score indicates that the generated audio better satisfies the corresponding textual requirements. For *audio quality*, a higher score indicates better perceptual quality and fewer noticeable artifacts. The final human score for each sample is computed by averaging the ratings from all participants. These human judgments are then used to analyze the correlation between human perception and automatic evaluation metrics.

C.1 Model-Level Human Alignment

C.2 Clip-Level Human Agreement

For each of the 130 generated clips, we average the ratings from all 26 participants. The human event and attribute scores are the mean event-completeness and attribute-satisfaction ratings, respectively; their equal-weight mean defines the human semantic score, excluding audio quality. We linearly

Table 8: Descriptive model-level Spearman correlations with human judgments across 13 models on the human-study samples. Semantic denotes the equal-weight mean of event-completeness and attribute-satisfaction ratings. These correlations support model-level association and do not certify individual questions or Yes/No decisions.

Human judgment	**Automatic metric**	ρ
Event completeness	Event SR	0.743
Attribute satisfaction	Attribute SR	0.825
Semantic	Overall SR	**0.879**
Semantic	$\mathrm{CLAP}_{\mathrm{MS}}$	0.312

Table 9: Clip-level association between automatic rubric scores and mean human semantic ratings. Confidence intervals are 95% prompt-clustered bootstrap intervals from 10,000 resamples. Attribute results exclude clips without attribute rubrics. These correlations do not certify individual questions or Yes/No decisions.

Comparison	N	**Pearson** r	**95% CI**	**Spearman** ρ	**95% CI**
Event	130	0.656	$[0.544, 0.747]$	0.644	$[0.503, 0.746]$
Attribute	104	0.475	$[0.280, 0.640]$	0.458	$[0.252, 0.625]$
Semantic	130	0.634	$[0.516, 0.721]$	0.632	$[0.477, 0.719]$

map the 1–5 human scale to 0–100 and compare it with Event SR, Attribute SR, and Overall SR for the same clip. Attribute agreement uses the 104 clips with at least one attribute rubric; the remaining 26 clips have no defined Attribute SR. We compute Pearson and Spearman correlations and obtain 95% confidence intervals from 10,000 bootstrap samples clustered by prompt, thereby resampling the 10 prompts rather than treating all model outputs as independent.

After normalization to the same 0–100 range, the mean absolute errors are 26.82, 25.82, and 24.31 points for event, attribute, and semantic scores, respectively. Thus, the correlations indicate moderate-to-strong clip-level association, whereas the absolute-score differences show that the automatic scores are not calibrated substitutes for human Likert ratings. This experiment does not provide human binary labels for individual rubrics and therefore cannot estimate human-grounded rubric-level accuracy, false-positive rate, or false-negative rate.

D TTA Model Generation Protocol

To ensure a model-native comparison across text-to-audio generation models, we follow the native text-conditioning interface of each model and report the input format, output configuration, and sampling settings. Unless otherwise specified, each model generates one audio sample per prompt without manual resampling, manual selection, or best-of output picking.

D.1 Evaluation Input and Runtime Setup

All evaluated models receive the text descriptions provided by AudioScape-TTA. The input CSV contains the fields `name`, `caption`, and `struct_caption`. The `caption` field contains the free-form natural language description and serves as the default generation prompt for most evaluated models. Most models directly use the free-form `caption` field as input. For models supporting structured prompting, i.e., Make-An-Audio-2 and Dasheng-AudioGen, the annotated `struct_caption` is converted into the corresponding model-specific prompt representation.

All models are inferred on NVIDIA H20 GPUs with CUDA Version 12.5.

D.2 Per-Model Generation Protocol

We use the official open-source checkpoints and inference pipelines of all evaluated TTA models. Unless otherwise specified, each model generates one 10-second audio sample per input prompt

without manual selection or resampling. The reported configurations are collected from official checkpoints, released inference scripts, and configuration files.

AudioLDM2[1] We use 200 diffusion steps with a 10-second generation length.

Make-An-Audio-2[2] Generation uses DDIM sampling with 100 steps and guidance scale 4.

AudioGen[3] The text caption is directly used with the default generation configuration of the official checkpoint.

Tango 2[4] We use the default text-to-audio pipeline with 200 sampling steps and guidance scale 3.

TangoFlux[5] Generation uses the default flow-matching configuration with 50 steps and guidance scale 4.5.

EzAudio[6] We use the official inference configuration with 100 DDIM steps and guidance scale 5.

MAGNet[7] The official 10-second checkpoint is used with its default generation configuration.

Stable Audio Open[8] We use the official inference configuration with 100 sampling steps and classifier-free guidance scale 7.

MMAudio[9] The model generates audio using 25 sampling steps and guidance strength 4.5.

Foley-Omni[10] We use the default text-to-audio setting with 50 sampling steps and audio guidance scale 3.

Omni2Sound[11] The caption is converted into the required text prompt field. We use the default configuration with 100 steps and guidance scale 2.5.

AudioStory[12] Generation uses 50 inference steps with guidance scale 4.

Dasheng-AudioGen[13] We use the official inference pipeline.

D.3 Model Architecture and Design Observations

The systems differ in architecture, training data, model scale, text supervision, and inference configuration. Table 20 records these characteristics for context; the benchmark is not a controlled architectural ablation. Strong results for individual diffusion, flow-matching, or unified systems should not be attributed to a single design choice. Likewise, zero target-speech coverage under a tested interface does not imply that the system claims to support specified speech. The component and composition breakdowns in Table 4 are therefore important complements to Overall SR.

Task definition and system selection. AudioScape-TTA evaluates complete-soundscape generation from a single text prompt without assuming that all checkpoints were trained for the same task. The present experiments include only released model systems, each run through its native text-conditioning interface to produce one final waveform; we do not implement an external TTS–music–SFX composition pipeline. Such a modular pipeline is a complementary future system design that must additionally perform prompt decomposition, temporal scheduling, spatialization, and loudness-aware mixing. Because scoring consumes the final waveform, the same structured rubrics

[1] https://huggingface.co/cvssp/audioldm2.
[2] https://huggingface.co/ByteDance/Make-An-Audio-2.
[3] https://huggingface.co/facebook/audiogen-medium.
[4] https://huggingface.co/declare-lab/tango2.
[5] https://huggingface.co/declare-lab/TangoFlux.
[6] https://huggingface.co/OpenSound/EzAudio.
[7] https://huggingface.co/facebook/magnet-medium-10secs.
[8] https://huggingface.co/stabilityai/stable-audio-open-1.0.
[9] https://huggingface.co/hkchengrex/MMAudio.
[10] https://huggingface.co/CocoBro/Foley-Omni.
[11] https://huggingface.co/Dalision/Omni2Sound.
[12] https://huggingface.co/TencentARC/AudioStory-3B.
[13] https://huggingface.co/mispeech/Dasheng-AudioGen.

Table 10: Publicly declared task scope of the exact checkpoints evaluated in AudioScape-TTA. ✓ denotes explicit support by the evaluated checkpoint or interface; △ denotes limited, indirect, or family-level support without an explicit guarantee for the evaluated checkpoint; and – denotes that the capability is not claimed (or is explicitly outside scope). *Joint* requires an explicit claim of one-pass mixed generation rather than merely sharing an architecture across tasks. These labels are based on the official papers, repositories, and model cards, not on performance on our benchmark.

Evaluated checkpoint	**Scope label**	**SFX**	**Music**	**Speech**	**Joint**	**Official scope and checkpoint caveat**
AudioLDM 2 (Liu et al., 2024)	General; music-capable	✓	✓	△	–	The evaluated `audioldm2` checkpoint targets general audio; speech uses separate official TTS checkpoints.
Make-An-Audio 2 (Huang et al., 2023a)	SFX-focused; broad TTA	✓	△	△	–	Broad TTA training includes music and human-voice clips, but no explicit intelligible-speech or joint-mix guarantee.
AudioGen (Kreuk et al., 2023)	SFX-focused	✓	–	–	–	The official `audiogen-medium` model is trained for environmental sounds and sound effects; MusicGen is the separate music model.
Tango 2 (Majumder et al., 2024)	SFX-focused; limited music	✓	△	–	–	Its inherited base training includes music corpora, but Tango 2 is aligned and evaluated primarily on AudioCaps event audio rather than as a dedicated music model.
TangoFlux (Hung et al., 2024)	SFX-focused	✓	△	–	–	General TTA with AudioCaps-style event evaluation; music is discussed as an application, not established as a dedicated capability.
EzAudio (Hai et al., 2025)	SFX-focused	✓	–	–	–	The evaluated XL checkpoint targets high-quality, natural-sounding sound-effect generation.
MAGNet (Ziv et al., 2024)	Music-capable	–	✓	–	–	We evaluate `magnet-medium-10secs`, the text-to-music checkpoint; Audio-MAGNet sound checkpoints are separate.
Stable Audio Open (Evans et al., 2025)	SFX-focused; limited music	✓	✓	–	–	Explicitly supports short musical clips as well as samples, Foley, ambience, beats, and riffs; the model card states that realistic vocals and intelligible speech are unsupported.
MMAudio (Cheng et al., 2025)	SFX/Foley-focused	✓	–	–	–	Targets ambient and action-synchronized Foley; its paper explicitly excludes background music and human speech from the Foley scope.
Foley-Omni (Tao et al., 2026)	Unified; speech/music-capable	✓	✓	✓	✓	Explicitly generates synchronized speech, sound effects, and music jointly, and also supports the three component tasks separately.
Omni2Sound (Dai et al., 2026)	General; music-capable	✓	✓	–	–	Official training includes event-audio and music corpora; *unified* refers to V2A, T2A, and video-text-to-audio conditioning, not intelligible speech or a three-component mix.
AudioStory (Guo et al., 2026)	SFX-focused; music-capable	✓	✓	–	–	Targets narrative soundscapes and explicitly trains on background-music data; conversations or voices do not constitute an intelligible or transcript-controlled speech claim.
Dasheng AudioGen (Mei et al., 2026)	Unified; speech/music-capable	✓	✓	✓	✓	Explicitly synthesizes intelligible speech, music, SFX, and environmental acoustics in one coherent scene from structured text.

could evaluate a modular output in future work, but modular systems are outside the current empirical comparison. The benchmark introduces no model training or fine-tuning.

Scope-aligned versus challenge-setting interpretation. Table 10 separates documented native scope from transfer to the benchmark's end-to-end mixed-soundscape task. For reproducibility, the checkpoint-specific supported sets used in the split are: SFX+Music for AudioLDM 2, Tango 2, Stable Audio Open, Omni2Sound, and AudioStory; SFX only for Make-An-Audio 2, AudioGen, TangoFlux, EzAudio, and MMAudio; Music only for the evaluated MAGNet checkpoint; and SFX+Music+Speech for Foley-Omni and Dasheng AudioGen. Tango 2's limited music support is included because its inherited released-checkpoint training explicitly contains music corpora; incidental training content, application-only discussion, or capabilities available only through a different checkpoint are excluded. Every supported single-component clip is *scope-aligned*. Every multi-component clip, and every clip containing an unsupported component, enters the *compositional/OOD challenge* split. The latter is compositional but not necessarily OOD for Foley-Omni and Dasheng AudioGen, which explicitly claim one-pass mixed generation. Overall SR therefore measures coverage under a common end-to-end challenge setting, while this split prevents unsupported-category transfer from being described as failure on a model's originally claimed task. Because the partitions are model-specific, their clip and rubric counts differ; complete per-model counts and assignments accompany the released evaluation statistics.

Most systems score lower on their respective challenge split. For each model, Δ is only a within-system descriptive contrast between that model's own scope-aligned and challenge subsets. For example, the 33.94-point gap for the evaluated music-only MAGNet checkpoint and the nearly unchanged Dasheng AudioGen score (-0.19) describe those checkpoints under their respective partitions. Because supported-category sets, clip counts, prompt content, component count, and rubric composition differ across models, neither the gap magnitudes nor the split scores should be used to rank cross-model transfer ability.

E Additional Experimental Results

E.1 Audio Quality Evaluation

In addition to semantic requirement satisfaction, we report conventional perceptual and distribution-based metrics as complementary analyses. Specifically, Audiobox-Aesthetic (Tjandra et al., 2025) evaluates perceptual quality from multiple dimensions, including content enjoyment, content usefulness, production complexity, and production quality. Distribution-based metrics include Fréchet distance (FD), mean KL distance, and Inception Score (ISC), computed with VGGish, PANNs, or PaSST as indicated in Table 11. FD compares generated and reference feature distributions; KL averages the divergence between paired generated/reference classifier outputs; and ISC measures the classifiability and diversity of generated samples without comparing them with the reference distribution.

Table 12 and Table 11 summarize these complementary evaluations. The rankings obtained from these metrics are not always consistent with semantic requirement satisfaction. For example, models achieving strong distributional similarity or perceptual scores do not necessarily achieve the highest semantic satisfaction rates. This difference is expected because these metrics mainly characterize global audio quality, realism, or feature-level distribution matching, while our rubric-based evaluation directly verifies whether prompt-specified semantic requirements are realized.

Therefore, these metrics are reported to provide a broader view of model performance, but are not used as the primary criteria for evaluating structured soundscape instruction following.

E.2 Additional Speech Content Evaluation

Additional Speech-Content Diagnostics. Besides the binary speech-content satisfaction rate reported in the main paper using SCCA@0.60, we additionally report Word Error Rate (WER) and Character Error Rate (CER) in Table 13. These metrics are computed on target-utterance clips with valid Latin-word or CJK-character references by comparing ASR transcriptions against the reference text.

Table 11: Distribution-based audio quality metrics of open-source TTA models on AudioScape-TTA. These metrics evaluate global distributional similarity and perceptual quality in pretrained feature spaces and are reported as complementary analyses rather than components of the proposed semantic evaluation framework. The best and second-best results are highlighted in **bold** and underline, respectively.

Model	$\mathrm{FD}_{\mathrm{VGG}} \downarrow$	$\mathrm{FD}_{\mathrm{PANN}} \downarrow$	$\mathrm{FD}_{\mathrm{PaSST}} \downarrow$	$\mathrm{KL}_{\mathrm{PANN}} \downarrow$	$\mathrm{KL}_{\mathrm{PaSST}} \downarrow$	$\mathrm{ISC}_{\mathrm{PANN}} \uparrow$	$\mathrm{ISC}_{\mathrm{PaSST}} \uparrow$
AudioLDM 2	4.816	15.342	229.479	1.787	1.694	8.157	7.600
Make-An-Audio 2	5.580	21.137	245.892	2.477	2.432	5.481	5.900
AudioGen	3.976	**2.874**	155.277	2.064	1.930	7.255	7.648
Tango 2	3.965	13.619	227.666	1.919	1.625	7.194	6.306
TangoFlux	4.504	19.405	266.794	1.990	1.755	8.344	7.283
EzAudio	2.514	9.711	**119.163**	1.851	1.740	8.291	8.515
MAGNet	15.711	38.295	509.685	4.015	3.324	1.913	1.736
Stable Audio Open	3.114	14.089	199.270	2.035	1.912	9.418	**8.678**
MMAudio	3.986	11.574	181.498	2.050	1.943	8.088	8.160
Foley-Omni	2.728	16.391	211.450	**1.628**	1.522	8.336	8.475
Omni2Sound	**1.620**	10.160	132.897	1.773	**1.520**	**10.330**	8.600
AudioStory	4.394	21.132	308.961	2.309	2.183	8.379	7.607
Dasheng AudioGen	3.174	17.661	202.953	1.948	1.737	9.527	8.323

Table 12: Audiobox-Aesthetic evaluation results of open-source TTA models on AudioScape-TTA. CE, CU, PC, and PQ denote the aesthetic dimensions measured by Audiobox. These metrics evaluate perceptual audio quality and are reported as complementary analyses rather than semantic requirement evaluation. The best and second-best results are highlighted in **bold** and underline, respectively.

Model	**CE↑**	**CU↑**	**PC↑**	**PQ↑**
AudioLDM 2	4.21	6.04	3.45	6.34
Make-An-Audio 2	3.37	5.14	3.15	5.67
AudioGen	3.24	4.57	3.60	5.26
Tango 2	4.13	5.61	4.32	6.18
TangoFlux	4.27	5.53	4.04	6.02
EzAudio	4.16	5.58	4.07	5.96
MAGNet	4.20	5.11	**5.28**	5.74
Stable Audio Open	3.38	**6.33**	1.96	**6.64**
MMAudio	3.79	5.25	3.33	5.77
Foley-Omni	4.14	5.56	3.65	5.88
Omni2Sound	4.21	5.80	3.95	6.24
AudioStory	3.96	5.36	3.33	5.94
Dasheng AudioGen	**4.29**	5.81	3.56	6.18

Table 13: Supplementary clip-level ASR transcription diagnostics. WER is computed over 69 valid reference clips containing 1,503 Latin word units, while CER is computed over 11 valid reference clips containing 309 Chinese CJK character units. Both metrics are computed from Qwen3-ASR transcripts of generated audio and caption-derived target utterances. They are complementary diagnostics and do not alter the 163-rubric SCCA@0.60 benchmark evaluation. The best and second-best results are highlighted in **bold** and underline, respectively.

Model	WER↓	CER↓
AudioLDM 2	1.1949	1.0000
Make-An-Audio 2	1.0093	1.0000
AudioGen	1.1284	1.0000
Tango 2	1.2448	1.1294
TangoFlux	1.1756	0.9968
EzAudio	1.2582	0.9968
MAGNet	1.8603	1.0000
Stable Audio Open	1.0745	1.0712
MMAudio	1.0153	1.0000
Foley-Omni	<u>0.5469</u>	**0.9191**
Omni2Sound	1.0253	1.0000
AudioStory	1.1564	1.0356
Dasheng AudioGen	**0.2395**	<u>0.9935</u>

Table 14: Speech-content coverage satisfaction rate (%) at different SCCA thresholds. Each value is the percentage of 163 speech-content rubrics whose mixed-unit target coverage is at least τ.

Model	SCCA@0.40	SCCA@0.50	SCCA@0.60	SCCA@0.70	SCCA@0.80
AudioLDM 2	1.23	1.23	0.00	0.00	0.00
Make-An-Audio 2	1.23	1.23	0.00	0.00	0.00
AudioGen	0.00	0.00	0.00	0.00	0.00
Tango 2	0.00	0.00	0.00	0.00	0.00
TangoFlux	0.00	0.00	0.00	0.00	0.00
EzAudio	0.00	0.00	0.00	0.00	0.00
MAGNet	0.00	0.00	0.00	0.00	0.00
Stable Audio Open	0.00	0.00	0.00	0.00	0.00
MMAudio	0.00	0.00	0.00	0.00	0.00
Foley-Omni	<u>68.10</u>	<u>66.26</u>	<u>57.06</u>	<u>50.92</u>	<u>45.40</u>
Omni2Sound	0.00	0.00	0.00	0.00	0.00
AudioStory	0.00	0.00	0.00	0.00	0.00
Dasheng AudioGen	**80.98**	**80.37**	**77.30**	**72.39**	**68.71**

Unlike SCCA@0.60, which measures whether the required speech content is sufficiently covered for semantic evaluation, WER and CER quantify full transcription similarity and are therefore more sensitive to insertion, deletion, and substitution errors. Since many evaluated open-source TTA models are not explicitly designed for controllable speech generation, WER/CER are reported as complementary analyses rather than used in the benchmark's primary semantic ranking.

Speech-content coverage threshold sensitivity. Table 14 shows that the central speech conclusion is not driven by the choice of $\tau = 0.60$. Across the stricter range $\tau \in \{0.60, 0.70, 0.80\}$, Dasheng AudioGen consistently obtains the highest SCCA, while Foley-Omni consistently ranks second. Lowering the threshold to 0.40 or 0.50 yields only marginal non-zero scores (1.23%) for AudioLDM 2 and Make-An-Audio 2, while preserving the same two leading models.

Correlation between SCCA and WER. Table 15 compares SCCA@0.60 with the complementary ASR transcription diagnostics at the model level. We compute Spearman correlation using average ranks for tied SCCA scores and correlate SCCA@0.60 with $-\mathrm{WER}$ and $-\mathrm{CER}$, since lower edit error indicates better transcription fidelity. SCCA@0.60 shows moderate positive agreement with $-\mathrm{WER}$ ($\rho = 0.629$, $p = 0.021$) and $-\mathrm{CER}$ ($\rho = 0.653$, $p = 0.015$). Because most general-purpose models obtain tied zero SCCA@0.60 scores, these results should be interpreted

Table 15: Coarse model-level rank agreement between SCCA@0.60 and complementary clip-level edit-distance diagnostics across 13 TTA models.

Comparison	**Spearman ρ**	**Two-sided p**
SCCA@0.60 vs. $-$WER	0.629	0.021
SCCA@0.60 vs. $-$CER	0.653	0.015

Table 16: Audio-conditioned control experiments using manipulated reference audio. The results demonstrate that rubric-based evaluation requires informative and correctly matched acoustic evidence.

Setting	**Overall SR**	**Event SR**	**Attribute SR**
Reference Audio	**84.88**	**85.60**	**84.37**
Audio Masked	18.31	10.42	23.99
Shuffled Audio	25.88	22.84	28.07

as coarse agreement between target-content coverage and edit-distance-based transcription fidelity, rather than as a fine-grained ranking among lower-performing systems.

E.3 Evaluator Reliability and Sensitivity

Reference-audio controls reduce Overall SR from 84.88 to 18.31 under silence and 25.88 under shuffled, mismatched audio, indicating that the evaluator responds to acoustic evidence rather than relying solely on rubric wording. On a Gemini-verified balanced subset, Qwen achieves 94.91% balanced accuracy and a 4.42% false-positive rate against the retained Gemini labels. However, these labels are not an independent human gold standard. Replacing Qwen with Gemini increases reference-audio SR from 84.88 to 95.03, indicating that absolute SR values are evaluator-dependent. These experiments are sanity checks rather than generated-audio calibration: they do not test distorted-event confusion, source–attribute misattribution, or event-present/attribute-absent near misses, nor do they establish cross-evaluator ranking stability on generated audio. The following analyses provide the complete controls and caveats.

E.3.1 Sanity Checks for Audio-Grounded Rubric-Based Evaluation

Audio-conditioned controls. To verify whether rubric performance genuinely relies on acoustic evidence, we conduct two audio-conditioned control experiments using the reference-audio rubric pairs. These experiments follow the same audio-conditioned setting as our evaluation protocol, where each rubric item is evaluated together with an audio input. In the **audio-masked** setting, the original audio is replaced with a silent clip. In the **shuffled-audio** setting, each question is paired with the reference audio from a randomly selected different sample. As shown in Table 16, both control settings lead to a substantial drop in Overall SR, Event SR, and Attribute SR compared with the correctly matched reference-audio setting. SR follows the benchmark definition and includes event-presence, event-attribute, and thresholded speech-content correctness. These results show that the scores respond to informative and correctly matched audio evidence rather than the question format alone. They are controlled sanity checks, not calibration on ambiguous or distorted generated audio.

Evaluator capability check. To investigate whether remaining errors on reference audio are partially caused by evaluator limitations, we additionally evaluate the same QA pairs using Gemini, an alternative audio-language model. As shown in Table 17, Gemini-2.5-Pro achieves an Overall rubric satisfaction rate of 95.03% on the same rubric annotations and reference audio, substantially outperforming Qwen3-Omni-Instruct. The increase from 84.88% with Qwen3-Omni-Instruct to 95.03% with Gemini-2.5-Pro demonstrates evaluator dependence in absolute reference-audio SR, but does not definitively separate evaluator errors from rubric-construction errors. We do not evaluate cross-evaluator ranking stability on the generated outputs. We adopt Qwen3-Omni-Instruct throughout the

Table 17: Evaluator capability check on the reference-audio QA pairs. Gemini-2.5-Pro is included as an alternative reference evaluator to assess the effect of evaluator capacity on reference-audio rubric scores.

Evaluator	Overall SR	Event SR	Attribute SR
Qwen3-Omni-Instruct	84.88	85.60	84.37
Gemini-2.5-Pro	95.03	96.64	93.87

benchmark for accessibility and reproducibility, while the protocol remains compatible with future audio-language evaluators.

E.3.2 Evaluator Bias and Answer-Position Analysis

Our main *semantic satisfaction* metric measures whether prompt-specified requirements are realized in generated audio. Since the benchmark rubrics are formulated as positive requirement verification questions, we additionally analyze two potential evaluator biases: (i) a tendency to answer "Yes" regardless of audio evidence, and (ii) sensitivity to the position of the affirmative answer. These analyses evaluate evaluator reliability and do not extend the main metric into an open-ended extra-content hallucination measure.

Balanced reference-audio sensitivity analysis. We construct a contrastive subset from reference audio by pairing each event-presence or event-attribute positive rubric with a same-component contrastive target that is absent from the reference clip. To avoid exposing the intended polarity to the evaluator, candidate rubrics are independently verified using two Gemini-2.5-Pro prompts without providing the expected label, candidate source, or polarity. We retain only complete positive-negative pairs for which both items pass verification. The resulting held-out subset contains 452 semantic items from 226 complete pairs, with balanced labels (226 "Yes" and 226 "No" items).

We evaluate `Qwen3-Omni-Instruct` under both answer orders for every item ("Yes" as option A and "Yes" as option B), resulting in 904 evaluator calls. As shown in Table 18, Qwen3-Omni-Instruct achieves high discrimination capability on the reference-audio subset, with 94.25% recall, 4.42% false-positive rate, and 94.91% balanced accuracy. These results rule out a trivial always-"Yes" answering strategy and indicate limited answer-position sensitivity in this controlled reference-audio analysis.

Importantly, this experiment is a *Gemini-verified reference-audio robustness analysis* rather than a human-annotated calibration set. Therefore, the reported false-positive rate, precision, and calibration statistics should be interpreted as evaluator sensitivity estimates with respect to the verified subset, rather than definitive human-grounded error rates. The subset does not test generated-audio boundary cases such as distorted-event confusion, attributes produced by the wrong source, or event-present/attribute-absent near misses.

Answer-position sensitivity. We further randomize whether the affirmative answer appears as option A or option B. On the balanced reference-audio subset, the balanced-accuracy gap between the two answer orders is 1.33 percentage points, and only 14 of 452 semantic items (3.10%) receive different semantic predictions under the two orders. This suggests limited answer-position sensitivity at the item level in the controlled setting.

We further assess whether answer-order randomization affects the relative comparison among evaluated TTA systems. As shown in Table 19, the fixed-order and randomized-order model rankings exhibit strong agreement, with Spearman correlations ranging from 0.937 to 0.993 across semantic dimensions. The top-three models are identical under both protocols for overall, event-level, and attribute-level semantic satisfaction. These results indicate that the benchmark-level comparative conclusions are robust to answer-order randomization.

Scope of contrastive negatives. The contrastive negative rubrics above are used only to measure evaluator presence–absence discrimination on reference audio. They are not used to penalize generated audio for containing arbitrary additional sounds: a sound absent from a reference clip may still be compatible with a generation prompt. Accordingly, our semantic satisfaction metric measures

Table 18: Evaluator sensitivity analysis of `Qwen3-Omni-Instruct` on the Gemini-verified reference-audio subset. TPR (recall) measures the fraction of verified satisfied requirements correctly identified by the evaluator, while FPR measures the fraction of verified unsatisfied requirements incorrectly predicted as satisfied. TNR denotes specificity ($1 - \mathrm{FPR}$). Precision and F1 quantify the discrimination quality of affirmative predictions. ECE and Brier score measure probability calibration. The subset contains 452 balanced semantic items, including 226 positive and 226 negative items.

Metric	**Result**
Recall (TPR)	94.25%
False-positive rate (FPR)	4.42%
Specificity (TNR)	95.58%
Precision	95.52%
F1 score	94.88%
Accuracy	94.91%
Balanced accuracy	94.91%
ECE	0.0197
Brier score	0.0373

Table 19: Robustness of benchmark-level *model rankings* to answer-order randomization across 13 TTA systems. Spearman correlation is computed between the fixed-order and randomized-order model rankings.

Metric	**Spearman ρ**	**Top-3 Overlap**
Overall SR	0.979	3/3
Event SR	0.937	3/3
Attribute SR	0.993	3/3

satisfaction of prompt-specified target content, rather than unrestricted extra-content hallucination. Evaluating the latter would require prompts or annotations with explicit *forbidden* sound constraints, or dedicated human judgments.

F Limitations

AudioScape-TTA covers both professionally produced and everyday environmental soundscapes, but platform-based collection may introduce domain and production biases; the benchmark should therefore be viewed as a representative evaluation suite rather than an unbiased estimate of all real-world audio. The evaluated systems differ in their original scopes, and our scope-aligned and compositional/OOD splits only partially address this heterogeneity. Complexity bins also vary jointly with component composition, so stratified gaps are descriptive rather than causal. Human validation covers 10 prompts and supports using SR as a model- and clip-level associate of human semantic ratings. It does not certify individual questions or Yes/No decisions, tell whether a No is a generation miss or an evaluator miss, bind a Yes to the intended source, or measure how residual underspecified items move ranks. Reference-audio controls are sanity checks rather than generated-audio calibration; their Gemini-verified subset is not an independent human gold standard, and absolute SR is evaluator-dependent. We have not compared the protocol with AQAScore or AnyAudio-Judge on the same generated audio. Field-level grouping records separately stored annotation fields; it does not automatically yield purely acoustic, source-explicit, or non-overlapping questions. Residual wording can remain after automatic filtering. The evaluator sees only the audio and the question, so a stored `event_id` does not resolve reference. Event SR and Attribute SR are therefore field-partitioned subsets rather than a clean split between event recognition and attribute control.

G Additional Benchmark and Model Information

This section retains the benchmark feature inventory and model metadata used to contextualize the experimental results. Table 20 summarizes the evaluated systems, and Table 21 reports the

Table 20: Architectural overview of the 13 open-source text-to-audio (TTA) models evaluated in AudioScape-TTA. We summarize the generation paradigms, text encoders, audio representations, and generation backbones. LD denotes Latent Diffusion, DPO denotes Direct Preference Optimization, DiT denotes Diffusion Transformer, AR denotes Autoregressive, and Non-AR denotes Non-Autoregressive generation. For multimodal systems, only components relevant to text-to-audio generation are listed.

Model	**Paradigm**	**Text Encoder**	**Audio Representation**	**Generation Backbone**
AudioLDM 2	LD	CLAP + FLAN-T5	AudioMAE + Mel VAE	UNet
Make-An-Audio 2	LD	CLAP + T5	Mel VAE	Transformer
AudioGen	Autoregressive	T5 Conditioner	EnCodec Tokenizer	AR Transformer
Tango 2	LD + DPO	FLAN-T5	AudioLDM VAE	UNet
TangoFlux	Flow Matching	FLAN-T5	Stable Audio VAE	Flux DiT
EzAudio	DiT	FLAN-T5	Stable Audio VAE + DAC	EzAudio-DiT
MAGNet	Masked Modeling	T5-Large	EnCodec Tokenizer	Non-AR Transformer
Stable Audio Open	DiT	T5-Base	Waveform VAE	Diffusion Transformer
MMAudio	Flow Matching	CLIP	Mel VAE + BigVGAN	MM-DiT
Foley-Omni	MM Flow Matching	UM-T5	Mel VAE + BigVGAN	Conditional DiT
Omni2Sound	Unified Diffusion	FLAN-T5	Wav VAE	DiT
AudioStory	LLM-guided Diffusion	Qwen2.5 + T5	Whisper + VAE	LLM Planner + DiT
Dasheng AudioGen	Unified Non-AR	FLAN-T5-Large	Dasheng Tokenizer	Flow DiT

feature inventory. The inventory uses the specific definitions in its caption and is not an exhaustive comparison with every recent evaluator.

Table 21: Selected benchmark characteristics under the feature definitions below; this is not an exhaustive ranking of evaluation capability. In particular, partial support does not mean absence of fine-grained evaluation. *Structured Schema* indicates whether audio scenes are represented using predefined components (e.g., sound effects, background music, and speech) rather than only free-form captions. *Speech Content* indicates whether synthesized speech is evaluated against annotated linguistic targets, such as transcripts or specified utterances. *Semantic Diagnosis* indicates whether benchmarks provide annotation-grounded criteria for diagnosing specific semantic requirements. ✓ denotes full support, while △ denotes partial support, such as event-level, segment-level, or attribute-level evaluation without a unified semantic rubric. *Complexity Aware* indicates whether evaluation explicitly characterizes sample complexity or analyzes performance across different complexity levels.

Benchmark	**Samples**	**Schema**	**Speech content**	**Semantic diagnosis**	**Complexity**
AudioCaps 1.0 (Kim et al., 2019)	975	×	×	×	×
AudioCaps 2.0 (Kim et al., 2019)	1,778	×	×	×	×
Clotho (Drossos et al., 2020)	1,045	×	×	×	×
AudioCondition (Guo et al., 2024)	1,110	△	×	△	×
AudioTime (Xie et al., 2025)	500	△	×	△	△
RiTTA (He et al., 2025)	720	△	×	△	△
AudioAtlas (Wang et al., 2025a)	1,165	△	×	△	×
VinTAGe-Bench (Kushwaha & Tian, 2025)	636	△	×	△	×
TTA-Bench (Wang et al., 2026)	2,999	△	×	△	△
T2A-EpicBench (Wang et al., 2025b)	100	×	×	△	×
AudioScape-TTA	**2,258**	✓	✓	✓	✓