DASHENG AUDIOGEN: A UNIFIED MODEL FOR GENERATING COHERENT AUDIO SCENES FROM TEXT

Anonymous authors
Paper under double-blind review

ABSTRACT

Audio generation has long been fragmented, with speech, music, and sound effects produced by domain-specific models that fail to jointly generate coherent audio scenes from a single description. The key obstacles are insufficient fine-grained supervision for real-world mixed audio and limited acoustic representations for modeling concurrent audio components. We present Dasheng AudioGen, a unified framework for generating general mixed-audio scenes from text. Dasheng AudioGen introduces structured multi-view captions, which explicitly decouple complex acoustic scenes into complementary description views, thereby enabling fine-grained control over audio layers. Furthermore, we employ a high-dimensional unified semantic-acoustic representation as the shared latent space. It injects semantic priors that facilitate cross-modal training convergence, while its high-dimensional feature space provides sufficient capacity to disentangle and fuse concurrent audio components effectively. With these designs, a simple flow-matching DiT achieves high-quality end-to-end audio scene generation. We also establish a comprehensive evaluation pipeline for audio scene generation. Experiments demonstrate that Dasheng AudioGen achieves performance approaching real-world recordings in mixed-audio categories, while remaining competitive with specialized models in single-type generation tasks. Demos are available at *https://anonymous2232330.github.io/Dasheng-AudioGen/.*

1 INTRODUCTION

Current audio generation research is largely divided into separate domains, with independent architectures, methods, and datasets for speech, music, and sound effects. Text-to-speech (TTS) models synthesize clean speech without modeling the acoustic environment Hu et al. (2026); Ren et al. (2019); text-to-music (TTM) models generate instrumental music Copet et al. (2023); Agostinelli et al. (2023); and text-to-audio (TTA) models generate sound effects Liu et al. (2023a;b); Hung et al. (2024) but not intelligible speech. However, real-world audio rarely occurs as a single domain. For example, a news broadcast combines speech, background music, transition effects, and ambient sounds into one coherent *audio scene*[1]. This requires general audio generation systems to jointly model temporal relations, energy balance, masking effects, environmental consistency, and overall realism among different audio components within the same scene.

Unified audio scene generation faces two key challenges. The first is data and textual supervision. Existing high-quality audio datasets are typically domain-specific. For example, TTS datasets such as LibriTTS Zen et al. (2019) provide accurate transcripts and clean speech, but contain little music, sound events, or ambient sounds, thereby lacking acoustic diversity. In contrast, in-the-wild audio contains richer acoustic scenes, but is often annotated only with coarse global captions. For complex mixed audio, a global caption is insufficient to provide the fine-grained supervision needed for controllable and coherent scene generation.

The second challenge is audio representation. General audio scenes contain heterogeneous and overlapping sound sources, making them harder to model than single-domain TTS, TTM, or TTA. Traditional generation systems use low-dimensional VAE acoustic latents as targets, forcing the model to learn a

[1]In this paper, we use *audio scene* to denote a coherent mixed-audio clip that may contain intelligible speech, music, and sound effects.

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

Table 1: Generation capabilities of representative audio generation models. The last group supports coherent mixed-audio scene generation. Open indicates publicly available model weights; Appendix Table A1 compares generation paradigms, training data, and evaluation benchmarks.

Model	Sound Effect	Music	Intel. Speech	Audio Scene	Open
MusicGen Copet et al. (2023)	×	✓	×	×	✓
TangoFlux Hung et al. (2024)	✓	✓	×	×	✓
Qwen3-TTS Hu et al. (2026)	×	×	✓	×	✓
UniFlow-Audio Xu et al. (2025)	✓	✓	✓	×	✓
BagPiper Tian et al. (2026)	✓	✓	✓	✓	×
AudioChat Chen et al. (2026)	✓	✓	✓	✓	×
WavJourney Liu et al. (2025)	✓	✓	✓	✓	✓
Dasheng AudioGen	✓	✓	✓	✓	✓

difficult mapping from semantic text conditions to low-level acoustic spaces. Such compact latents may also lack the capacity to represent multiple coexisting audio components and their interactions. Thus, unified audio generation requires not only a unified architecture, but also a high-capacity representation space that preserves both semantic structure and acoustic details.

To address these challenges, we propose Dasheng AudioGen, a unified text-to-audio framework for general audio scene generation. Our key insight is that unified generation does not require separate modules for different sound types; instead, it requires structured conditioning and a unified semantic-acoustic latent space. Specifically, we introduce structured multi-view captions, which decompose a complex audio scene into complementary textual views, including global scene description, speaker style, speech transcript, sound events, music description, and acoustic environment. Compared with a single caption, this format provides finer-grained supervision and control. We further adopt a unified semantic-acoustic representation based on DashengTokenizer, which reduces the difficulty of text-to-audio mapping and provides sufficient capacity for jointly modeling speech, music, and sound effects. With these designs, a simple flow-matching DiT achieves high-quality end-to-end audio scene generation. Table 1 compares the generation capabilities of representative models; to our knowledge, Dasheng AudioGen is the first model to jointly generate speech, music, and sound effects end to end using a non-autoregressive flow-matching DiT in a shared high-dimensional semantic-acoustic space. It is specifically designed and systematically evaluated for coherent mixed-audio scenes.

Our evaluation pipeline covers AudioCaps Kim et al. (2019), MusicCaps Agostinelli et al. (2023), and LibriTTS Zen et al. (2019) for single-type generation, and MECAT Niu et al. (2025) for both single-type and mixed-audio scenes. We construct a strong Expert-Pipeline baseline and use human evaluation and PAFI (Physical Acoustic Fidelity Index), an LLM-as-a-judge metric, to assess perceptual quality, text relevance, and scene realism.

Our contributions are summarized as follows:

1. **End-to-end audio scene generation.** We propose Dasheng AudioGen, a unified text-to-audio framework that jointly generates intelligible speech, music, sound effects, and environmental acoustics within one audio scene, without domain-specific modeling.
2. **Structured multi-view captions.** We introduce a layered caption design that provides fine-grained supervision and disentangled control over different audio components, while remaining naturally compatible with agentic systems.
3. **Generation with semantic-acoustic representation.** Instead of relying on low-dimensional acoustic VAEs, we introduce a unified semantic-acoustic representation based on DashengTokenizer Dinkel et al. (2026) as the shared latent space for flow matching. It provides semantic priors for efficient training and sufficient capacity to disentangle and fuse concurrent audio components.
4. **Evaluation pipeline for audio scene generation.** We establish a comprehensive evaluation pipeline for audio scene generation. Experiments show that Dasheng AudioGen substantially outperforms Expert-Pipeline in mixed-audio scenes while remaining competitive with specialized models.

2 Related Work

Unified Audio Generation. AudioX Tian et al. (2025), UniAudio Yang et al. (2024), and UniFlow-Audio Xu et al. (2025) unify multiple audio generation tasks with task-specific conditioning modules, such as phoneme encoders for speech and MIDI encoders for music. However, these designs introduce architectural complexity and scale poorly to new audio types. They also only support multiple tasks separately. BagPiper Tian et al. (2026) adopts an autoregressive framework for joint audio understanding and generation, and demonstrates the ability to generate mixed audio compositions. However, it remains closed-source and relies on very long unstructured captions, approximately 500 words for a 10-second clip, which makes fine-grained and disentangled control difficult. Moreover, BagPiper does not provide a dedicated evaluation protocol for mixed-audio generation, making direct comparison on coherent mixed-audio scenes difficult. AudioChat Chen et al. (2026) and WavJourney Liu et al. (2025) also target mixed-audio generation. AudioChat is a closed-source autoregressive audio language model, and its StoryGen-Eval benchmark is unavailable. The open-source WavJourney uses LLM-based script planning followed by independent track generation and mixing. It thus lacks joint modeling of speech, music, and sound-effect interactions in a shared representation and incurs substantially higher inference costs.

These methods generate low-dimensional acoustic VAE latents or discrete audio tokens. To our knowledge, Dasheng AudioGen is the first model to jointly generate speech, music, sound effects, and ambient sounds end to end using a non-autoregressive flow-matching DiT in a shared high-dimensional semantic-acoustic space (1280-dimensional DashengTokenizer representations). It is specifically designed and systematically evaluated for coherent mixed-audio scenes. Table 1 and Appendix Table A1 compare these models; related work on single-type generation (TTS, TTM, and TTA) is reviewed in Appendix A.

3 Method

Given a text description y, our goal is to generate an audio scene x, where x may simultaneously contain speech, music, sound effects, and environmental acoustics. We formulate this task as conditional generation $p_\theta(x \mid y)$. Figure 1 illustrates the design of structured multi-view captions and the agentic inference pipeline.

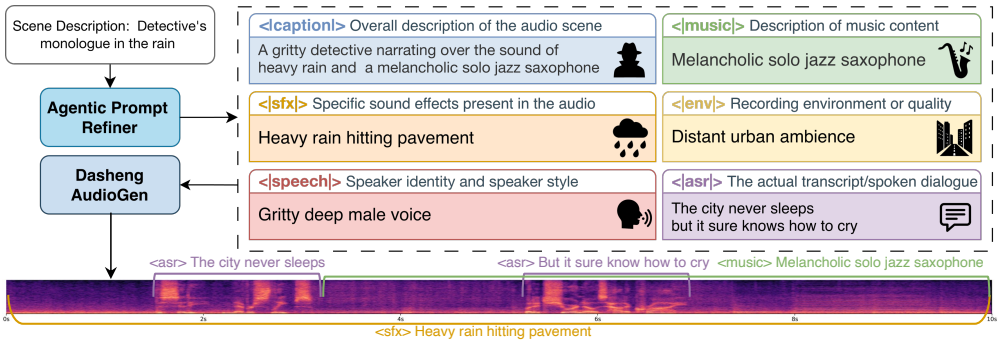

Figure 1: Structured multi-view audio scene captioning and agentic inference pipeline. Special tokens such as <|music|> describe different components of the target audio scene. An agentic prompt refiner automatically converts a simple scene description into a structured caption for fine-grained control.

3.1 Structured Multi-View Audio Scene Captioning

Prior audio generation systems usually rely on a single coarse text prompt or label, forcing the model to infer multiple audio layers from an entangled global description. For unified audio generation, such coarse conditioning limits fine-grained control and joint modeling of different audio components.

To address this, we introduce structured multi-view captions, which decompose an audio scene into six complementary views, as shown in Figure 1. Each view is associated with a dedicated special token,

such as <|caption|> and <|speech|>. Every sample contains the <|caption|> field for the global scene description, while the remaining fields are used only when applicable. For example, pure speech samples do not contain <|music|> or <|sfx|>.

This structured condition provides factorized supervision for complex audio scenes. Instead of using separate text encoders or task-specific modules for different sound types, we expose different semantic views to a unified text encoder via explicit special tokens. Compared with a single global caption, multi-view captions reduce semantic entanglement among control factors and enable fine-grained control.

Structured captions contain no explicit temporal annotations, such as timestamps, ordering labels, or durations. Temporal relations among components are learned implicitly from waveform-level supervision on large-scale real audio.

At inference, the structured format supports an optional Agentic Prompt Refiner that populates the fields from a short scene description. The refiner uses DeepSeek-V4-flash (Appendix J). Users can also provide only <|caption|> directly to the generation model, leaving other fields unspecified. None of the reported experiments uses the refiner or any LLM-based prompt rewriting.

3.2 View-Aware Conditioning

Dasheng AudioGen uses a single T5 Chung et al. (2024) text encoder to condition the flow-matching DiT. Given a structured multi-view caption, we represent it as a sequence of view segments $y = [s_1, y_1, s_2, y_2, \ldots, s_K, y_K]$, where s_k is the special token for the k-th view, such as <|asr|>, and y_k is its textual content. These special tokens define explicit semantic boundaries and expose view identity to the text encoder. The full sequence is encoded by T5 as $C \in \mathbb{R}^{L \times d_c}$, where L is the number of text tokens and d_c is the text embedding dimension.

The DiT generates DashengTokenizer latent sequences. Let $H_l \in \mathbb{R}^{T \times d}$ denote the audio hidden states at the l-th DiT block, where T is the latent length and d is the hidden dimension. Each block uses self-attention to model temporal dependencies among audio latents and cross-attention to inject multi-view text conditions:

$$\text{CrossAttn}(H_l, C) = \text{softmax}\left(\frac{Q(H_l)K(C)^\top}{\sqrt{d}}\right)V(C). \tag{1}$$

This allows each audio latent token to softly select relevant information from different caption views according to the current generation state. Dasheng AudioGen thus achieves view-aware conditioning with only special tokens and cross-attention, without view-specific encoders or task-specific modules.

3.3 Semantic-Acoustic Latent Space

Many prior models Liu et al. (2023b); Xu et al. (2025) use low-dimensional acoustic VAE latents as the generation space. However, general audio scenes contain concurrent heterogeneous components and require joint modeling of their interactions. This makes it difficult to map semantic text conditions to purely acoustic latents, while the low-dimensional VAE bottleneck may discard details needed for overlapping components.

To introduce semantic priors into the generation space, we use the unified semantic-acoustic representation from DashengTokenizer Dinkel et al. (2026). This choice addresses the need for both semantic priors and sufficient representational capacity in mixed-scene generation. Given an audio waveform x, the DashengTokenizer encoder produces a continuous latent representation $z = E_{\text{DS}}(x) \in \mathbb{R}^{T \times 1280}$ with a frame rate of 25 Hz. Unlike low-dimensional acoustic VAE latents, DashengTokenizer representations contain both semantic information and acoustic detail. The controlled experiments in Section 4.4 support this choice: replacing acoustic VAE latents with the unified representation while keeping the DiT backbone unchanged yields an average gain of approximately 20% on MECAT.

3.4 Flow Matching Objective

As shown in Figure 2, we perform generation in the DashengTokenizer latent space using standard flow matching Lipman et al. (2024). Let $z_1 = E_{\text{DS}}(x)$ be the real audio latent and $z_0 \sim \mathcal{N}(0, I)$ be Gaussian noise. For $t \sim \mathcal{U}(0, 1)$, we construct $z_t = (1 - t)z_0 + tz_1$. The DiT learns a conditional

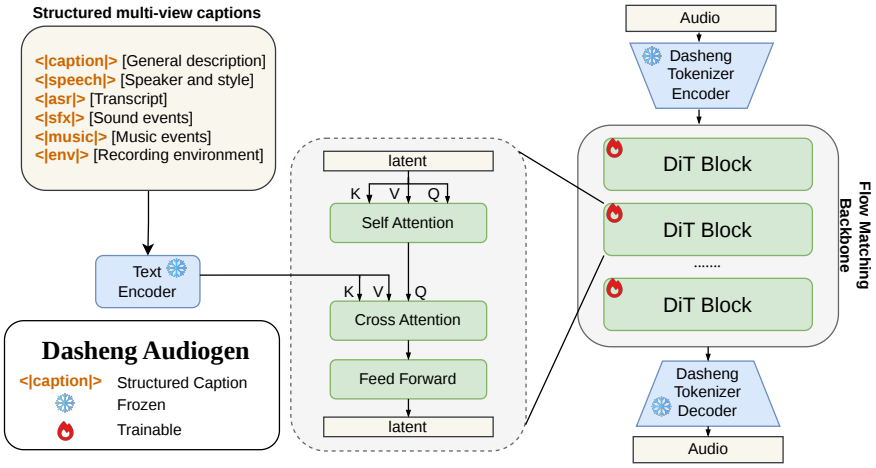

Figure 2: Overview of Dasheng AudioGen. A structured multi-view caption is encoded by T5 and used to condition a DiT that generates DashengTokenizer latents via flow matching. The DashengTokenizer decoder converts latents to waveforms.

vector field $v_\theta(z_t, t, C)$, where C is the text condition encoded from the structured caption. The training objective is

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{z_0, z_1, t, y}\left[\|v_\theta(z_t, t, C) - (z_1 - z_0)\|_2^2\right]. \tag{2}$$

At inference, we start from Gaussian noise and solve $dz_t/dt = v_\theta(z_t, t, C)$ to obtain the generated latent $\hat{z}$, which is decoded into a waveform by the DashengTokenizer decoder. We use classifier-free guidance for stronger conditioning and randomly drop caption fields during training to support inputs with different levels of detail.

Compared with previous diffusion-based unified models, our architecture is deliberately simple. UniAudio Yang et al. (2024) and UniFlow-Audio Xu et al. (2025) rely on multiple task-specific encoders, which complicates optimization. In contrast, Dasheng AudioGen requires only structured audio captions as input.

3.5 Implementation Details

The DiT has width 1536 and 32 layers with approximately 2B trainable parameters. The DashengTokenizer decoder has 173M parameters with 12 layers and hidden dimension 1280. We use Flan-T5-Large Chung et al. (2024) as the text encoder, with 780M parameters. We train with AdamW, batch size 256, and learning rate 5×10^{-4} with cosine decay to 10% for 800k steps, which takes 10 days on 8 H200 GPUs. Every multi-view caption field except `<|caption|>` is randomly dropped with probability 0.2 during training to improve robustness.

At inference, we use 25 flow-matching steps with classifier-free guidance scale 5.0.

4 Experiments

4.1 Experimental Setup

Our training dataset is ACAVCaps Niu et al. (2026), a large-scale audio captioning dataset derived from ACAV100M Lee et al. (2021). We train on a private superset with 77k hours covering speech, music, and sound effects. ACAVCaps uses a multi-expert annotation pipeline that analyzes each audio clip from six domain-specific perspectives, which we convert into our structured multi-view caption format. Detailed caption construction method and examples are provided in Appendix F.2.

Table 2: Standard benchmark results on AudioCaps, MusicCaps, and LibriTTS. Best in **bold** and second-best underlined.

	AudioCaps				MusicCaps				LibriTTS	
Method	FAD_{VGG} ↓	FD_{PANNS} ↓	KL ↓	CLAP ↑	FAD_{VGG} ↓	FD_{PANNS} ↓	KL ↓	CLAP ↑	WER(%) ↓	UTMOSv2 ↑
GT	0	0	0	53.0	0	0	0	37.0	2.82	3.14
AudioLDM2 Liu et al. (2023b)	2.29	21.84	1.41	47.6	3.13	-	1.20	30.1	-	-
TangoFlux Hung et al. (2024)	**2.26**	19.13	**1.19**	**58.0**	-	-	-	-	-	-
AudioX Tian et al. (2025)	2.45	24.69	-	44.0	1.42	18.53	**1.12**	**38.6**	-	-
UniFlow-Audio Xu et al. (2025)	5.74	**17.18**	1.43	47.6	4.05	27.12	1.87	24.1	-	-
MusicGen Copet et al. (2023)	-	-	-	-	3.80	-	1.31	28.0	-	-
Qwen3-TTS Hu et al. (2026)	-	-	-	-	-	-	-	-	**2.15**	**3.40**
Ours	3.19	26.06	1.86	43.8	**1.37**	**18.45**	1.37	33.4	10.77	3.12

Our main evaluation benchmark is MECAT Niu et al. (2025), which is a held-out test set of ACAV-Caps that categorizes audio into single-type (S00 = speech, 0M0 = music, 00A = sound effects) and mixed categories (0MA, S0A, SM0, SMA). We use a compact notation where S, M, and A denote the presence of speech, music, and sound effects, respectively, and 0 denotes absence. Unlike single-type benchmarks such as AudioCaps, MECAT contains both single and mixed-type audio samples together with rich multi-view annotations, making it especially suitable for assessing mixed audio scene generation. We also report on AudioCaps Kim et al. (2019), MusicCaps Agostinelli et al. (2023), and LibriTTS Zen et al. (2019) for comparability with prior work. Across these benchmarks, we report audio distribution metrics (FAD, FD, and KL), text similarity metrics (CLAP Elizalde et al. (2023), GLAP Dinkel et al. (2025), and multiFLAM Chen et al. (2026); Wu et al. (2025)), and speech-related metrics (WER and UTMOSv2 Baba et al. (2024)).

Details of the evaluation dataset statistics and metrics are provided in Appendix G.1 and Appendix G.2.

For single-type audio evaluation, we compare Dasheng AudioGen against representative specialized models, including AudioLDM2 Liu et al. (2023b), TangoFlux Hung et al. (2024), MusicGen Copet et al. (2023), and Qwen3-TTS Hu et al. (2026). We also include unified audio generation models that integrate multiple audio generation tasks, such as AudioX Tian et al. (2025) and UniFlow-Audio Xu et al. (2025). For mixed-audio evaluation, we further construct a strong Expert-Pipeline baseline. This baseline uses different expert models (Qwen3-TTS, MusicGen, and TangoFlux) to generate the speech, music, and sound-effect components of a mixed audio scene separately, and then mixes them into the final audio output. Because these models prefer different input formats, we construct model-specific evaluation captions; details are provided in Appendix G.3. We also compare against WavJourney Liu et al. (2025), an open-source compositional system that uses an LLM to plan a script, generates each track independently, and mixes the tracks into a scene.

In addition, we conduct ablation studies to quantify the contribution of structured multi-view captions and the unified semantic-acoustic representation. We also recruit 20 audio-domain professionals to evaluate overall quality (OVL) and text relevance (REL) on MECAT. The human evaluation and results are presented in Appendix H.1. On the LLM-based Physical Acoustic Fidelity Index (PAFI), Dasheng AudioGen outperforms Expert-Pipeline overall (3.57 vs. 3.42, $p < 0.001$), and PAFI system preferences correlate with human OVL ($r = 0.822$). Full protocols and results are provided in Appendix I.

4.2 Standard Generation Benchmarks

To validate Dasheng AudioGen's capabilities on single-type audio generation, we report results on AudioCaps, MusicCaps, and LibriTTS in Table 2.

Sound Effect Generation. On AudioCaps, Dasheng AudioGen (FAD 3.19) slightly trails task-optimized models such as AudioLDM2 (2.29) and TangoFlux (2.26), but performs comparably to the unified multi-task model AudioX (2.45) and substantially outperforms UniFlow-Audio (5.74). This performance gap can be attributed to three factors: (1) unlike our model, the other TTA baselines include AudioCaps in their training sets, yielding an in-domain advantage; (2) pure sound effects (00A) comprise only 1.34% of our training data (Appendix Table A6), leading to data scarcity in this category; and (3) our minimalist architecture does not include modality-specific inductive biases or targeted optimizations, such as the CLAP-oriented preference optimization used in TangoFlux.

Table 3: Results on MECAT mixed-audio categories. Best values are in **bold** and second-best values are underlined. 0MA=Music+Audio, S0A=Speech+Audio, SM0=Speech+Music, SMA=Speech+Music+Audio.

Category	Method	FAD_{VGG} ↓	FD_{PANNS} ↓	KL ↓	CLAP ↑	multiFLAM ↑	WER(%) ↓	UTMOSv2 ↑
0MA	TangoFlux	9.71	58.44	2.45	31.6	-	-	-
	MusicGen	13.06	60.98	2.86	19.3	-	-	-
	Qwen3-TTS	23.17	149.24	5.28	-3.0	-	-	-
	Expert-Pipeline	5.55	43.79	1.83	**35.2**	-	-	-
	WavJourney Liu et al. (2025)	9.89	58.09	2.10	23.4	75.55	-	-
	Ours	**3.25**	**30.58**	**1.42**	31.9	**76.95**	-	-
S0A	TangoFlux	10.92	43.94	2.23	36.0	-	99.30	1.84
	MusicGen	30.38	89.09	5.60	0.2	-	100.00	1.65
	Qwen3-TTS	19.62	62.84	2.42	15.3	-	**13.63**	**3.48**
	Expert-Pipeline	7.10	32.36	1.86	**37.9**	-	49.22	2.22
	WavJourney	9.15	31.66	1.58	34.9	**94.76**	31.37	2.45
	Ours	**1.75**	**8.56**	**0.69**	36.3	90.19	22.98	2.60
SM0	TangoFlux	11.57	40.77	1.11	33.7	-	99.44	1.40
	MusicGen	19.83	57.28	2.38	13.8	-	99.99	1.65
	Qwen3-TTS	10.20	63.15	2.16	18.6	-	**15.60**	**3.49**
	Expert-Pipeline	9.55	24.10	0.69	30.9	-	24.31	2.26
	WavJourney	4.88	22.73	0.67	**33.8**	**81.23**	26.49	2.66
	Ours	**1.70**	**6.69**	**0.33**	32.7	73.85	21.96	2.72
SMA	TangoFlux	11.40	50.58	2.05	29.0	-	99.72	1.63
	MusicGen	20.18	69.19	3.41	10.7	-	100.00	1.69
	Qwen3-TTS	16.53	83.01	2.45	10.2	-	**14.92**	**3.22**
	Expert-Pipeline	6.38	30.10	1.08	37.7	-	62.14	2.24
	WavJourney	6.95	35.87	1.18	33.1	**86.09**	34.06	2.51
	Ours	**2.17**	**17.75**	**0.63**	**38.3**	71.40	28.98	2.46

Music and Speech Generation. Our method exhibits strong competitiveness in music and speech generation. On MusicCaps, Dasheng AudioGen achieves an FAD of 1.37, substantially outperforming AudioLDM2 (3.13) and MusicGen (3.80). It also slightly surpasses the unified multi-task model AudioX (1.42) and clearly outperforms UniFlow-Audio (4.05). On LibriTTS, its WER (10.77%) is higher than that of the optimized Qwen3-TTS system (2.15%). This is primarily because our model currently generates audio with a fixed 10-second duration, truncating longer texts and artificially inflating WER.

On UTMOSv2, which better reflects overall speech perception and naturalness, we achieve a score of 3.12, approaching Qwen3-TTS (3.40).

In conclusion, despite being designed for complex mixed-audio scenes, Dasheng AudioGen maintains highly competitive performance across standard TTA, TTM, and TTS benchmarks, with particularly strong results in music quality and speech naturalness. This demonstrates that a unified representation and minimalist architecture do not compromise foundational single-type generation capabilities, providing a solid basis for its strong performance in mixed-audio scenarios.

4.3 MECAT Benchmark

To comprehensively evaluate model performance in complex mixed-audio scenes, we conduct experiments on the MECAT benchmark.

Single-Type Categories. Results on the single-type categories (00A, 0M0, and S00) are reported in Appendix Table A2. Dasheng AudioGen achieves the best performance on all acoustic distribution metrics (FAD, FD, and KL) across all three categories, substantially outperforming all baselines. In addition, our model attains the best or second-best results on the text-audio similarity metrics CLAP and GLAP. On the speech-only S00 category, Dasheng AudioGen has higher WER (22.96% vs. 13.14%) and lower UTMOSv2 (2.92 vs. 3.46) than Expert-Pipeline, but markedly better acoustic distribution metrics (e.g., FAD 1.76 vs. 8.46). We attribute the Expert-Pipeline's weaker distributional performance mainly to its lack of environmental awareness. Unlike the clean studio-style speech in LibriTTS, the MECAT speech categories contain rich vocal and environmental details that reflect realistic acoustic conditions (see Appendix F.2 for representative examples). This indicates that Dasheng AudioGen better models speech together with its acoustic context, whereas specialized TTS systems typically ignore environmental descriptions and generate speech detached from the physical scene.

Mixed Categories. Table 3 reports results on the mixed-audio categories 0MA, S0A, SM0, and SMA. Across all mixed categories, Dasheng AudioGen achieves the best FAD, FD, and KL by a large margin,

and the second-best WER on every speech-containing category. It ranks first or second on CLAP in three of the four mixed categories and on UTMOSv2 in two of the three speech-containing mixed categories. Dasheng AudioGen substantially outperforms WavJourney on distributional metrics (FAD, FD, KL) and speech intelligibility (WER), whereas WavJourney scores higher on frame-level component alignment (multiFLAM) in most categories. This suggests a trade-off: joint generation improves distributional fidelity and intelligibility, while independently controllable tracks favor component alignment.

For example, in the most challenging SMA setting, which contains concurrent speech, music, and sound effects, Dasheng AudioGen achieves a remarkably low FAD of 2.17 together with a WER of 28.98%. By comparison, the Expert-Pipeline reaches an FAD of 6.38 and a WER of 62.14%, even though the standalone Qwen3-TTS baseline attains 14.92% WER on the same category. MECAT's in-the-wild audio includes multiple speakers, low signal-to-noise ratios, and rich paralinguistic content. These conditions can impair end-of-sequence (EOS) prediction in autoregressive TTS systems such as Qwen3-TTS, leading to excessively long outputs, hallucinated speech, and higher WER. The Expert-Pipeline's increase in WER after mixing indicates that independently generated components are difficult to combine into a coherent mixed-audio scene. In particular, the expert models lack global coordination, which leads to severe acoustic masking and mutual interference between speech, music, and sound effects after mixing. WavJourney reduces WER to 34.06%, suggesting that its LLM-based planning and track-mixing strategy help mitigate the interference associated with directly superimposing independently generated components. Nevertheless, its WER remains higher than the 28.98% achieved by Dasheng AudioGen, supporting the benefit of end-to-end joint modeling for speech intelligibility in mixed-audio scenes. By contrast, the unified representation helps coordinate energy distribution and cross-component interactions at the scene level, leading to more natural and coherent disentanglement and fusion of overlapping audio layers.

4.4 Ablation Experiments

We first compare training on the public subset and the full dataset, then ablate two core designs: structured multi-view captions and the unified semantic-acoustic representation.

Public vs. Full Training Data. The public ACAVCaps release contains approximately 10K hours, compared with the private 77K-hour superset used in our main experiments. We compare the two to assess whether effective training is possible using only publicly available data. At the same number of training steps, the public ACAVCaps subset yields comparable performance on several metrics, while the full dataset improves music generation and speech naturalness. These results support effective training with public data; experimental settings and full results are provided in Appendix D.

Table 4: Comparison between structured and unstructured captions on the non-speech MECAT categories and LibriTTS. Best in **bold**.

	Structured						Unstructured					
Category	$FAD_{VGG}\downarrow$	$FD_{PANNS}\downarrow$	KL ↓	GLAP ↑	WER(%) ↓	UTMOSv2 ↑	$FAD_{VGG}\downarrow$	$FD_{PANNS}\downarrow$	KL ↓	GLAP ↑	WER(%) ↓	UTMOSv2 ↑
00A	**4.07**	**17.86**	**1.35**	**10.59**	-	-	4.86	22.73	1.51	9.82	-	-
0M0	**1.68**	**15.44**	**0.61**	9.52	-	-	2.34	15.78	0.66	**9.60**	-	-
0MA	**3.25**	**30.58**	**1.42**	**8.02**	-	-	5.04	33.93	1.61	7.89	-	-
LibriTTS	-	-	-	-	**10.77**	**3.12**	-	-	-	-	52.0	2.70

Structured vs. Unstructured Captions. Table 4 compares structured and unstructured captions on MECAT and LibriTTS. In the unstructured setting, the model is trained with only the `<|caption|>` field. Because transcripts cannot be integrated fairly into plain unstructured text for MECAT speech subsets such as S00 and SMA, we use LibriTTS as the speech benchmark; prompt construction details are provided in Appendix G.3. On the non-speech MECAT subsets 00A, 0M0, and 0MA, structured captions outperform unstructured captions on 11 of the 12 reported metrics; for example, on 0MA they reduce FAD from 5.04 to 3.25. On LibriTTS, structured captions reduce WER from 52.0% to 10.77% and improve UTMOSv2 from 2.70 to 3.12, confirming that explicit transcript conditioning is critical for intelligible speech generation.

To separate the effects of information content and structure, we construct an equal-information unstructured baseline by retaining all fields but replacing field markers with spaces. At 400K training steps, structured conditioning performs better on 9 of 10 metrics, supporting the benefit of explicit field markers; full results are in Appendix E.

Acoustic vs. Unified Embeddings. To validate the benefit of unified semantic-acoustic embedding, we train flow-matching DiT models on both the large-scale mixed-audio dataset ACAVCaps Niu et al. (2026) and single-type datasets (WavCaps Mei et al. (2024), LP-MusicCaps Doh et al. (2023), LibriTTS Zen et al. (2019)), using either a VAE-based acoustic representation (d=128) or DashengTokenizer's unified representation (d=1280). Appendix C provides the relative gains (Figure A1), detailed model configurations, and absolute scores.

When trained on ACAVCaps, the unified representation exhibits stable and substantial advantages. Across MECAT subsets, it outperforms the acoustic representation on nearly all metrics, with an average gain of approximately 20%. Substantial improvements are also observed on the single-type evaluation sets: AudioCaps (+33.3%), MusicCaps (+27.0%), and LibriTTS (+86.8%). This advantage cannot be explained by stronger acoustic reconstruction, as the unified representation does not surpass the VAE in reconstruction quality Dinkel et al. (2026). Instead, we attribute it to the semantic priors embedded in the unified space, which shorten the cross-modal mapping from text to audio and facilitate more stable generation alignment in mixed-audio scenes. A localized exception is the FD metric on MECAT 0M0 subset, which does not alter the overall dominance of the unified representation.

By contrast, when trained on single-type datasets, the unified representation still performs better on AudioCaps (+16.0%) and MusicCaps (+5.6%). However, on LibriTTS, a trade-off between speech intelligibility and quality emerges: the unified representation degrades WER (-90.6%) while substantially improving UTMOSv2 (+104.1%). This contrasts with ACAVCaps training, where both WER (+67.3%) and UTMOSv2 (+106.3%) improve consistently. The speech quality (UTMOSv2) gain likely arises from the unified representation's larger capacity, whereas the WER behavior is less straightforward.

Unlike TTA and TTM, which focus on global acoustic scene rendering, TTS demands strict local temporal alignment between transcripts and audio. When trained on the clean speech dataset LibriTTS, the unified representation's semantic priors provide little additional advantage, and its higher dimensionality instead increases the difficulty of learning fine-grained pronunciation alignments. Conversely, training on mixed-audio data such as ACAVCaps requires extracting speech-relevant components from complex mixed representations and aligning them with input transcripts, significantly increasing alignment difficulty. The acoustic representation is especially vulnerable to this shift: its WER on the LibriTTS test set surges from 6.4% to 32.9% when training moves from clean LibriTTS to mixed ACAVCaps. The unified representation's WER on LibriTTS, however, drops from 12.2% to 10.77% under the same shift, suggesting that its semantic priors effectively disentangle audio layers and alleviate the alignment burden.

5 Limitations

Dasheng AudioGen has several limitations. First, because Dasheng AudioGen's training data consists entirely of 10-second audio clips, the current model is limited to 10-second generation. Second, it lacks explicit timestamp control. Third, in the TTS setting, the model currently supports only coarse speaker-style control from text and does not support voice cloning or explicit speaker-identity conditioning. As a result, speaker similarity metrics are not applicable. Moreover, although the generated speech is natural, its intelligibility still lags behind specialized TTS systems. Fourth, full reproducibility is currently limited because training relies on a much larger private superset of ACAVCaps Niu et al. (2026) rather than the public release of approximately 10K hours. However, Appendix D demonstrates effective training on the public subset alone, partially mitigating this limitation.

6 Conclusion

We presented Dasheng AudioGen, a unified text-to-audio model for coherent audio scene generation. By introducing structured multi-view captions and semantic-acoustic latents, Dasheng AudioGen achieves high-quality end-to-end complex audio scene generation using a simple flow-matching DiT framework. Comprehensive evaluations show that our method substantially outperforms expert pipelines in complex mixed-audio scenes while remaining competitive on single-domain tasks. Future work will explore variable-length generation, improved speech intelligibility, and finer-grained controllability such as audio editing and explicit temporal control.

AI Use Statement

We disclose the following uses of generative AI. PAFI uses Gemini-3.1-Pro to assess the physical acoustic fidelity of generated audio (Appendix I.3). For training annotations, Whisper Radford et al. (2023) generates <|asr|> transcripts, and DeepSeek-R1 organizes smaller models' outputs into the remaining fields Niu et al. (2025). The optional agentic prompt refiner uses DeepSeek-V4-flash to convert natural-language descriptions into structured multi-view captions (Appendix J); it is not used in any reported experiment. Beyond these uses, generative AI was not used to synthesize training data, develop theoretical frameworks, formulate or prove mathematical results, propose hypotheses, or design experiments. We also used large language models to translate and polish parts of the manuscript. The authors reviewed all AI-assisted content and take responsibility for the final text, claims, and artifacts.

Ethics Statement

Our subjective evaluation involved 20 audio professionals who rated overall quality and text relevance (see Appendix H for instructions, criteria, and the interface). The evaluation was low-risk, with no harmful, sensitive, or distressing listening material. We collected no personally identifiable information, biometric data, or other private attributes and report only aggregate results. Participants were informed of the purpose and procedure, and risks were minimized. On this basis, the study did not undergo IRB or equivalent ethics review. Realistic audio with intelligible speech poses risks of misinformation, fraud, and impersonation; mixed scenes may make fabricated content more convincing and harder to detect than single-type audio. The current model supports only coarse text-based speaker-style control, not voice cloning or explicit speaker-identity conditioning, and cannot directly reproduce a specific person's voice. Training data may be imbalanced across languages, regions, and acoustic scenes; music and sound-effect generation also raise copyright and provenance concerns. We recommend synthetic-audio labeling, watermarking, and misuse detection for future extensions and deployments, with restrictions on fraud, fabricated evidence, and impersonation. This work studies modeling, representations, and evaluation for unified audio scene generation; we do not endorse deceptive, privacy-infringing, or otherwise harmful uses.

Reproducibility Statement

We provide training and inference code, model weights, and evaluation scripts. Training configurations are in Section 3.5; data and caption construction in Appendix F; objective evaluation, baseline checkpoints, and prompts in Appendix G; and human evaluation and PAFI protocols in Appendices H and I. MECAT Niu et al. (2025) and the public ACAVCaps Niu et al. (2026) subset are available; Appendix D demonstrates effective model training using only approximately 10K hours of public data. However, the full 77K-hour training set cannot be released, so this validation does not fully reproduce all main-experiment results.

References

Andrea Agostinelli, Timo I. Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, Matt Sharifi, Neil Zeghidour, and Christian Frank. MusicLM: Generating music from text. In *International Conference on Machine Learning (ICML Workshop)*, 2023.

Kaito Baba, Wataru Nakata, Yuki Saito, and Hiroshi Saruwatari. The t05 system for the VoiceMOS Challenge 2024: Transfer learning from deep image classifier to naturalness MOS prediction of high-quality synthetic speech. In *IEEE Spoken Language Technology Workshop (SLT)*, pp. 818–824, 2024. doi: 10.1109/SLT61566.2024.10832315.

William Chen, Prem Seetharaman, Rithesh Kumar, Oriol Nieto, Shinji Watanabe, Justin Salamon, and Zeyu Jin. Audiochat: Unified audio storytelling, editing, and understanding with transfusion forcing. *arXiv preprint arXiv:2602.17097*, 2026.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53, 2024.

Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. Simple and controllable music generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Heinrich Dinkel, Zhiyong Yan, Tianzi Wang, Yongqing Wang, Xingwei Sun, Yadong Niu, Jizhong Liu, Gang Li, Junbo Zhang, and Jian Luan. GLAP: General contrastive audio-text pretraining across domains and languages. *arXiv preprint arXiv:2506.11350*, 2025.

Heinrich Dinkel, Xingwei Sun, Gang Li, Jiahao Mei, Yadong Niu, Jizhong Liu, Xiyang Li, Yifan Liao, Jiahao Zhou, Junbo Zhang, and Jian Luan. DashengTokenizer: One layer is enough for unified audio understanding and generation. *arXiv preprint arXiv:2602.23765*, 2026.

SeungHeon Doh, Keunwoo Choi, Jongpil Lee, and Juhan Nam. Lp-musiccaps: LLM-based pseudo music captioning. *arXiv preprint arXiv:2307.16372*, 2023.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. CLAP: Learning audio concepts from natural language supervision. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron Weiss, and Kevin Wilson. Cnn architectures for large-scale audio classification. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

Hangrui Hu, Xinfa Zhu, Ting He, Dake Guo, Bin Zhang, Xiong Wang, Zhifang Guo, Ziyue Jiang, Hongkun Hao, Zishan Guo, Xinyu Zhang, Pei Zhang, Baosong Yang, Jin Xu, Jingren Zhou, and Junyang Lin. Qwen3-TTS technical report. *arXiv preprint arXiv:2601.15621*, 2026.

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 *International Conference on Machine Learning (ICML)*, 2023.

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.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. AudioCaps: Generating captions for audios in the wild. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2019.

Jaehyeon Kim, Jungil Kong, and Juhee Son. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *International Conference on Machine Learning (ICML)*, 2021.

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

Sangho Lee, Jiwan Chung, Youngjae Yu, Gunhee Kim, Thomas Breuel, Gal Chechik, and Yale Song. Acav100m: Automatic curation of large-scale datasets for audio-visual video representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10274–10284, 2021.

Peike Patrick Li, Boyu Chen, Yao Yao, Yikai Wang, Allen Wang, and Alex Wang. Jen-1: Text-guided universal music generation with omnidirectional diffusion models. In *2024 IEEE Conference on Artificial Intelligence (CAI)*, pp. 762–769. IEEE, 2024.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Matthew Nickel, and Matthew Le. Flow matching for generative modeling. In *International Conference on Learning Representations (ICLR)*, 2024.

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 (ICML)*, 2023a.

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*, 2023b.

Xubo Liu, Zhongkai Zhu, Haohe Liu, Yi Yuan, Qiushi Huang, Meng Cui, Jinhua Liang, Yin Cao, Qiuqiang Kong, Mark D Plumbley, et al. Wavjourney: Compositional audio creation with large language models. *IEEE Transactions on Audio, Speech and Language Processing*, 33:2830–2844, 2025.

Xinhao Mei, Chutong Meng, Haohe Liu, Qiuqiang Kong, Tom Ko, Chengqi Zhao, Mark D Plumbley, Yuexian Zou, and Wenwu Wang. Wavcaps: A ChatGPT-assisted weakly-labelled audio captioning dataset for audio-language multimodal research. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:3339–3354, 2024.

Yadong Niu, Tianzi Wang, Heinrich Dinkel, Xingwei Sun, Jiahao Zhou, Gang Li, Jizhong Liu, Xunying Liu, Junbo Zhang, and Jian Luan. MECAT: A multi-experts constructed benchmark for fine-grained audio understanding tasks. *arXiv preprint arXiv:2507.23511*, 2025.

Yadong Niu, Tianzi Wang, Heinrich Dinkel, Xingwei Sun, Jiahao Zhou, Gang Li, Jizhong Liu, Junbo Zhang, and Jian Luan. ACAVCaps: Enabling large-scale training for fine-grained and diverse audio understanding. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 15347–15351. IEEE, 2026.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. *International Conference on Machine Learning*, 2023.

Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. FastSpeech: Fast, robust and controllable text to speech. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 4779–4783. IEEE, 2018.

Jinchuan Tian, Haoran Wang, Bo-Hao Su, Chien-yu Huang, Qingzheng Wang, Jiatong Shi, William Chen, Xun Gong, Siddhant Arora, Chin-Jou Li, et al. Bagpiper: Solving open-ended audio tasks via rich captions. *arXiv preprint arXiv:2602.05220*, 2026.

Zeyue Tian, Yizhu Jin, Zhaoyang Liu, Ruibin Yuan, Xu Tan, Qifeng Chen, Wei Xue, and Yike Guo. Audiox: Diffusion transformer for anything-to-audio generation. *arXiv preprint arXiv:2503.10522*, 2025.

Yusong Wu, Christos Tsirigotis, Ke Chen, Cheng-Zhi Anna Huang, Aaron Courville, Oriol Nieto, Prem Seetharaman, and Justin Salamon. Flam: Frame-wise language-audio modeling. *arXiv preprint arXiv:2505.05335*, 2025.

Xuenan Xu, Jiahao Mei, Zihao Zheng, Ye Tao, Zeyu Xie, Yaoyun Zhang, Haohe Liu, Yuning Wu, Ming Yan, Wen Wu, Chao Zhang, and Mengyue Wu. UniFlow-Audio: Unified flow matching for audio generation from omni-modalities. *arXiv preprint arXiv:2509.24391*, 2025.

Dongchao Yang, Jinchuan Tian, Xu Tan, Rongjie Huang, Songxiang Liu, Haohan Guo, Xuankai Chang, Jiatong Shi, Jiang Bian, Zhou Zhao, et al. Uniaudio: Towards universal audio generation with large language models. In *International Conference on Machine Learning*, 2024.

Heiga Zen, Rob Clark, Ron J. Weiss, Viet Dang, Ye Jia, Yonghui Wu, Yu Zhang, and Zhifeng Chen. Libritts: A corpus derived from librispeech for text-to-speech. In *Interspeech*, 2019.

Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. *arXiv preprint arXiv:2510.11690*, 2025.

A Extended Related Work

Text-to-Speech. TTS has progressed from concatenative and statistical parametric synthesis to neural models like Tacotron Shen et al. (2018), FastSpeech Ren et al. (2019), VITS Kim et al. (2021), and more recent systems such as Qwen3-TTS Hu et al. (2026). Recent systems achieve high naturalness on clean speech, but typically ignore acoustic context—generated speech sounds like a studio recording regardless of the described environment.

Text-to-Music. Text-to-music generation follows two dominant paradigms. Autoregressive language models operate over discrete audio tokens: MusicGen Copet et al. (2023) uses a single-stage transformer to predict EnCodec tokens from text, while MusicLM Agostinelli et al. (2023) generates music via hierarchical sequence-to-sequence modeling. Diffusion-based approaches instead operate in continuous latent spaces: AudioLDM2 Liu et al. (2023b) applies latent diffusion with a joint audio-text representation, and JEN-1 Li et al. (2024) uses flow matching for efficient generation. These systems produce high-quality music but cannot generate speech or integrate with spoken content.

Text-to-Audio. Text-to-audio generation for general sound effects splits into diffusion and flow-matching approaches. AudioLDM Liu et al. (2023a) and Make-An-Audio Huang et al. (2023) use latent diffusion to generate sound effects from text, operating on mel-spectrogram or VAE latents. More recently, flow-matching methods have emerged as efficient alternatives: TangoFlux Hung et al. (2024) uses conditional flow matching for fast, high-quality TTA without the iterative sampling overhead of diffusion. Both paradigms focus exclusively on non-speech, non-music sounds.

Mixed-audio generation systems differ substantially in openness, generation paradigm, and evaluation conditions, complicating direct comparison. Table A1 compares BagPiper, AudioChat, WavJourney, and our method along these dimensions.

Table A1: Comparison of systems for mixed-audio scene generation.

	BagPiper	AudioChat	WavJourney	**Ours**
Mixed-audio scenes	✓	✓	✓	✓
Open-source model	×	×	✓	✓
Generation paradigm	Audio LM	Audio LM	Pipeline	FM DiT
Public training data	×	×	-	Partial
Mixed-scene eval.	×	✓	✓	✓
Public benchmark	×	×	×	✓
Latency	-	32 s	144 s	1.5 s

BagPiper uses 271K annotated samples and AudioChat uses AudioCopilot-6M; neither training set is public. ACAVCaps has a public subset and is therefore marked "Partial"; WavJourney is an agentic system that requires no additional training. The "Mixed-scene eval." row indicates whether a system has a dedicated mixed-audio evaluation: BagPiper has no dedicated benchmark, whereas AudioChat uses StoryGen-Eval and WavJourney uses a storytelling evaluation set, neither of which is publicly available. We use the fully public MECAT benchmark. Latency denotes generation time per sample.

B MECAT Single-Type Results

Table A2: Results on MECAT single-type categories. Best values are in **bold** and second-best values are underlined. S00 denotes speech only audio category, 0M0 denotes music only, and 00A denotes sound effects only.

Category	Method	FAD_{VGG} ↓	FD_{PANNS} ↓	KL ↓	CLAP ↑	GLAP ↑	WER(%) ↓	UTMOSv2 ↑
00A	TangoFlux	15.68	48.81	2.30	29.9	4.89	-	-
	MusicGen	43.68	87.41	5.99	-2.8	-5.78	-	-
	Qwen3-TTS	51.42	153.55	5.90	-3.5	-8.56	-	-
	Expert-Pipeline	7.35	29.38	2.17	**38.1**	9.56	-	-
	Ours	**4.07**	**17.86**	**1.35**	36.5	**10.59**	-	-
0M0	TangoFlux	5.04	24.97	0.82	**32.2**	6.70	-	-
	MusicGen	4.53	19.75	0.88	26.0	5.77	-	-
	Qwen3-TTS	21.68	137.10	4.99	-2.7	-8.79	-	-
	Expert-Pipeline	3.93	16.77	0.87	24.1	6.50	-	-
	Ours	**1.68**	**15.44**	**0.61**	27.0	**9.52**	-	-
S00	TangoFlux	10.54	53.15	1.63	29.1	-0.91	99.03	1.71
	MusicGen	26.74	89.71	5.38	4.4	-9.54	100.00	1.61
	Qwen3-TTS	8.75	27.55	0.84	31.4	0.20	14.82	**3.60**
	Expert-Pipeline	8.46	25.84	0.84	28.4	-0.89	**13.14**	3.46
	Ours	**1.76**	**3.93**	**0.40**	**33.9**	**11.45**	22.96	2.92

Table A2 shows the results on MECAT single-type categories. Dasheng AudioGen achieves the best results on all audio distribution metrics (FAD, FD, and KL), ranks first or second on the text relevance metrics (CLAP and GLAP), and remains competitive with specialized models on the speech-related metrics (WER and UTMOSv2). Overall, these results indicate that Dasheng AudioGen matches or surpasses specialized models on single-type audio generation tasks.

C Acoustic vs. Unified Embedding Ablation Results

Figure A1 summarizes the relative gains of unified over acoustic representations across training and evaluation sets, while Table A3 reports the corresponding absolute scores. Positive gains indicate improvement according to each metric's direction. The heatmap highlights broad gains under ACAVCaps training and the WER–UTMOSv2 trade-off under LibriTTS-only training.

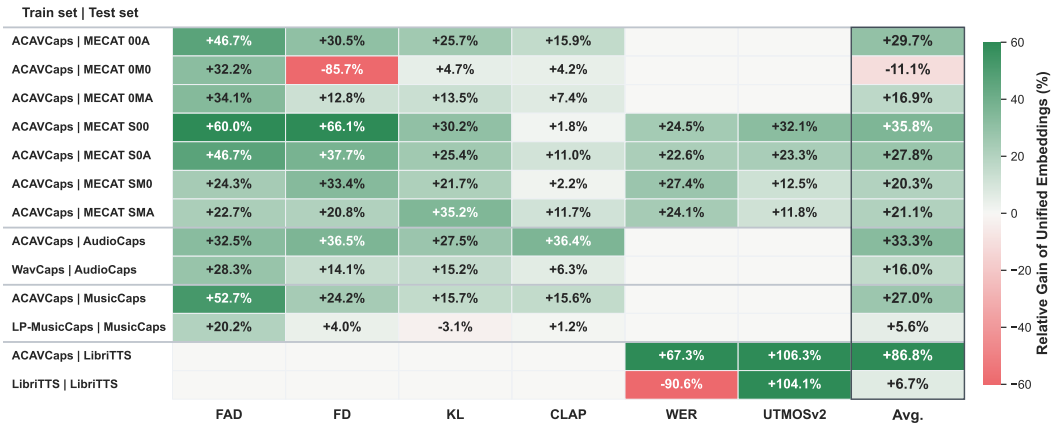

Figure A1: Comparison of relative percentage gains of unified embeddings (DashengTokenizer) over acoustic embeddings (VAE). Each row is labeled as training set | evaluation set.

When trained on ACAVCaps, the acoustic and unified representations use the same DiT backbone (32 layers, hidden size 1536, ∼2B parameters). When trained on single-type datasets, we select different DiT configurations with approximately 750M total parameters to match the preference of each representation. The acoustic representation uses a 24-layer DiT with hidden size 1024. For the unified

Table A3: Acoustic (VAE) vs. unified (DashengTokenizer) embedding results under ACAVCaps and single-type dataset training. Better value is highlighted in **bold**.

		FAD_{VGG} ↓		FD_{PANNS} ↓		KL ↓		CLAP ↑		WER(%) ↓		UTMOSv2 ↑	
Train set	Test set	Acou.	Uni.	Acou.	Uni.	Acou.	Uni.	Acou.	Uni.	Acou.	Uni.	Acou.	Uni.
ACAVCaps	MECAT-00A	7.63	**4.07**	25.69	**17.86**	1.82	**1.35**	31.5	**36.5**	-	-	-	-
ACAVCaps	MECAT-0M0	2.48	**1.68**	**8.32**	15.44	0.64	**0.61**	25.9	**27.0**	-	-	-	-
ACAVCaps	MECAT-0MA	4.93	**3.25**	35.07	**30.58**	1.64	**1.42**	29.7	**31.9**	-	-	-	-
ACAVCaps	MECAT-S00	4.39	**1.76**	11.60	**3.93**	0.57	**0.40**	33.3	**33.9**	30.41	**22.96**	2.21	**2.92**
ACAVCaps	MECAT-S0A	3.28	**1.75**	13.74	**8.56**	0.92	**0.69**	32.7	**36.3**	29.68	**22.98**	2.11	**2.60**
ACAVCaps	MECAT-SM0	2.24	**1.70**	10.04	**6.69**	0.43	**0.33**	32.0	**32.7**	30.24	**21.96**	2.42	**2.72**
ACAVCaps	MECAT-SMA	2.81	**2.17**	22.42	**17.75**	0.98	**0.63**	34.3	**38.3**	38.17	**28.98**	2.20	**2.46**
ACAVCaps	AudioCaps	4.73	**3.19**	41.06	**26.06**	2.56	**1.86**	32.1	**43.8**	-	-	-	-
WavCaps	AudioCaps	4.27	**3.06**	24.72	**21.23**	1.84	**1.56**	41.2	**43.8**	-	-	-	-
ACAVCaps	MusicCaps	2.90	**1.37**	24.34	**18.45**	1.62	**1.37**	28.9	**33.4**	-	-	-	-
LP-MusicCaps	MusicCaps	4.35	**3.47**	23.68	**22.74**	**1.94**	2.00	25.8	**26.1**	-	-	-	-
ACAVCaps	LibriTTS	-	-	-	-	-	-	-	-	32.90	**10.77**	1.52	**3.12**
LibriTTS	LibriTTS	-	-	-	-	-	-	-	-	**6.40**	12.20	1.48	**3.02**

representation, following the finding of RAE Zheng et al. (2025) that the DiT hidden size should exceed the representation dimensionality, we use an 11-layer DiT with hidden size 1536.

D Training on the Public Subset of ACAVCaps

We compare training on the public ACAVCaps subset (approximately 10K hours) and the full dataset (77K hours). Both models use structured multi-view captions and are evaluated at 400K training steps; the full-data model's final 800K checkpoint is not used in this comparison. We compute CLAP scores using `630k-best.pt`, which differs from the CLAP checkpoint used in the main experiments; the scores are therefore not directly comparable across the two evaluations.

Table A4: Public ACAVCaps subset (approximately 10K hours) versus the full training set (77K hours). Both use structured multi-view captions and 400K-step checkpoints; CLAP scores are percentages.

	AudioCaps		MusicCaps		LibriTTS		MECAT		MECAT speech subset	
Training data	FAD ↓	CLAP ↑	FAD ↓	CLAP ↑	WER(%) ↓	UTMOSv2 ↑	FAD ↓	CLAP ↑	WER(%) ↓	UTMOSv2 ↑
Public subset (∼10K h)	**3.54**	38.3	7.78	41.4	**20.08**	2.413	**2.475**	**47.4**	**24.56**	2.664
Full training set (77K h)	3.67	**39.7**	**3.84**	**47.4**	22.61	**3.146**	2.586	44.0	26.31	**3.053**

Under this setup, each model performs better on 5 of 10 metrics (Table A4). The public-subset model achieves better AudioCaps FAD, LibriTTS WER, and MECAT FAD, CLAP, and WER. The full-data model improves music generation and speech naturalness: MusicCaps FAD is 3.84 versus 7.78, while UTMOSv2 is 3.146 versus 2.413 on LibriTTS and 3.053 versus 2.664 on the MECAT speech subset. At the same number of training steps, the benefits of the full dataset are thus concentrated in music generation and speech naturalness.

The public subset also covers 99.953% of token occurrences in the full dataset when weighted by frequency. It thus covers most frequently occurring tokens, though textual coverage does not imply comparable acoustic diversity. Together, these results support a reproducible path to effective training with public data, while the full dataset retains advantages in music generation and speech naturalness.

E Equal-Information Structured vs. Unstructured Caption Ablation

The caption-only control changes both information content and structure: the Jaccard similarity between `<|caption|>` and the other fields is 9.6%, indicating complementary information. To isolate the effect of structure, the equal-information unstructured baseline retains all field content but removes field markers and joins the text with spaces.

We independently retrain both models on the full 77K-hour dataset and evaluate them at 400K steps, using the same training budget as in Appendix D. These checkpoints come from separate training runs, so their results may differ from those of the main-experiment model. We compute CLAP scores using

`630k-best.pt`, which differs from the CLAP checkpoint used in the main experiments; the scores are therefore not directly comparable across the two evaluations.

Table A5: Structured versus unstructured captions with equal information. Both use the full 77K-hour training set and 400K-step checkpoints; CLAP scores are percentages.

	AudioCaps		MusicCaps		LibriTTS		MECAT		MECAT speech subset	
Caption	FAD ↓	CLAP ↑	FAD ↓	CLAP ↑	WER(%) ↓	UTMOSv2 ↑	FAD ↓	CLAP ↑	WER(%) ↓	UTMOSv2 ↑
Equal-info unstructured	**3.30**	35.3	7.72	41.3	55.28	3.035	2.932	41.0	28.32	2.863
Structured	3.89	**37.5**	**2.60**	**49.7**	**17.14**	**3.160**	**2.451**	**44.3**	**27.31**	**2.976**

Structured conditioning performs better on 9 of 10 metrics in Table A5, reducing LibriTTS WER from 55.28% to 17.14%. With identical information content, these results suggest that explicit field markers help the model use caption information and identify transcripts. The exception is AudioCaps FAD, which is higher under structured conditioning (3.89 vs. 3.30); improvements thus do not extend to every metric.

F Training Details

F.1 Training Data Distribution

Table A6: Training data statistics. Average word counts are computed on the English subset using structured and unstructured captions.

Category	Duration (h)	Proportion (%)	Structured Avg. Words	Unstructured Avg. Words
00A	1036.54	1.34	28.24	11.55
0M0	10492.79	13.52	35.89	13.19
0MA	465.28	0.60	35.97	13.23
S00	37094.54	47.79	52.49	12.64
S0A	7601.22	9.79	49.82	12.89
SM0	19447.17	25.05	55.60	13.61
SMA	1482.54	1.91	54.57	14.20
Total	77620.07	100.00	50.42	13.01

We train on a superset of ACAVCaps, a 77k-hour multilingual mixed-audio dataset with rich annotations, where all audio clips are 10 seconds long. Table A6 summarizes training data distribution statistics and average word counts for structured and unstructured captions across all categories. The training distribution is dominated by speech-containing data, while also including a substantial amount of mixed-audio content. Pure speech (S00) alone accounts for 47.79% of the training set, and the four speech-related categories (S00, S0A, SM0, and SMA) together account for 84.54%. At the same time, the mixed-audio categories (0MA, S0A, SM0, and SMA) account for 37.35% of the total data, providing broad coverage of scenes in which multiple audio components co-occur and interact. By contrast, the proportions of pure music (0M0) and pure sound effects (00A) are smaller, at 13.52% and 1.34%, respectively. This distribution exposes the model to a large amount of speech and mixed-audio data, which is consistent with its strong performance on speech-related and mixed-audio scene generation tasks.

Across categories, structured captions contain roughly two to three times as many words as unstructured captions on average, showing that the structured format provides substantially richer descriptive information.

Table A7 shows the speech language distribution of training data. English is the dominant language in the training data, accounting for 58.86%, followed by Spanish, Portuguese, and Russian. This imbalance is likely to affect model performance across languages.

F.2 Structured Caption Construction Method and Training Cases

ACAVCaps Niu et al. (2026) uses a multi-expert annotation pipeline that analyzes each audio clip from six domain-specific perspectives: `long` and `short` (detailed and summarized scene descrip-

Table A7: Top training languages ranked by duration, followed by all remaining languages grouped as Other.

Language	Duration (h)	Proportion (%)
English	15367.80	58.86
Spanish	2740.96	10.50
Portuguese	1916.24	7.34
Russian	1217.39	4.66
French	933.91	3.58
Japanese	874.51	3.35
Korean	848.15	3.25
German	842.29	3.23
Other	1369.16	5.24

tions), `speech` (speaker characteristics), `music` (music description), `sound` (sound events and effects), and `environment` (acoustic properties such as reverberation and recording quality). This multi-perspective annotation scheme directly motivates our structured caption design. In our training format, the `long`/`short` annotations are mapped into `<|caption|>` to provide an overall scene description, `speech` is mapped to `<|speech|>` to describe speaker style, `music` is mapped to `<|music|>`, `sound` is mapped to `<|sfx|>`, and `environment` is mapped to `<|env|>` to encode acoustic context. For samples that contain speech, we additionally generate `<|asr|>` transcripts using Whisper Radford et al. (2023). This conversion ensures that the model input preserves the annotation granularity available in the source data while matching the structured conditioning format used at training time.

Below we show structured caption training cases from different subcategories without cherry-picking.

Training Case 1: S00 (English)

<|caption|> Male voice analyzing volleyball match outcomes with clear diction. <|speech|> Sportscast-style narration: Two segments discussing team performances (Italy vs. Serbia/US/China) with valence shift from neutral to positive. <|asr|> only two sets left to the opponents' both lost against Serbia. Italy took a second win in a row in 2011, getting the better hand on United States and China <|env|> Studio-quality speech recording with intermittent low-frequency interference.

Training Case 2: 0M0

<|caption|> Upbeat acoustic guitar instrumental with background distortion. <|music|> Folk-inspired guitar performance with rhythmic precision and harmonic resonance. <|env|> Studio recording with persistent electrical interference artifacts.

Training Case 3: 00A

<|caption|> Vehicle engine idling with persistent background interference. <|sfx|> Mechanical engine operation with electronic interference. <|env|> Indeterminate recording space with electrical interference.

Training Case 4: SMA (English)

<|caption|> Neutral speech utterance precedes door sound against melancholic musical accompaniment with audio artifacts. <|speech|> Neutral-toned utterance: 'Okay, open it'. <|asr|> Okay, open it. <|sfx|> Distinct door movement noise preceding vocal utterance. <|env|> Variable acoustic environment with noticeable background interference.

Training Case 5: SMA (Portuguese)

<|caption|> Multiple speakers discussing votes over synth-guitar blend and broadcast artifacts. <|speech|> Multiple Portuguese speakers discussing voting preferences ('Torcerei por Rita Cadillac... meu voto também') with intermittent overlapping dialogue. <|asr|> vai desgastando a pessoa.

Torcerei por Rita Cadillac nessa edição. Quem não torcerá é por Rita Cadillac. Ela tem o meu voto também. Também? Vamos acompanhar o voto do... <|sfx|> Transmission interference noises overlapping content. <|music|> Synthesized instrumentation blending electronic elements and guitar characteristics. <|env|> Broadcast booth setting exhibiting equipment noise and signal degradation.

Training Case 6: SMA (German)

<|caption|> Female voice speaking over synth-driven pop music and subtle ambient sounds. <|speech|> Female voice explaining a hair preparation process in German. <|asr|> lockern und einige Strähnen rausziehen. Dann nehme ich einfach den restlichen Teil meiner Habe und schiebe es in die Hand. <|sfx|> Ambiguous environmental noises suggesting manual activity. <|music|> Electronically processed guitar patterns with pop music structure. <|env|> Indoor recording with mild electrical interference and background artifacts.

G Objective Evaluation

G.1 Evaluation Datasets Statistics

We evaluate model performance on AudioCaps, MusicCaps, LibriTTS, and MECAT. Since Dasheng AudioGen is designed to generate 10-second audio clips, we evaluate on the subset of LibriTTS-test-clean whose utterances are shorter than 10 seconds. MECAT is a multilingual benchmark, and its speech categories contain utterances in multiple languages. To improve evaluation stability and reduce variation from multilingual ASR performance, we report objective results on the English subset of MECAT for the speech-related categories. Table A8 summarizes the number of evaluation samples used in each dataset. For filtered datasets, we report the sample counts before and after filtering. We note that metrics on categories with relatively few samples, such as MECAT 0MA and SMA, may be less stable and reliable than those on larger subsets.

Table A8: Sample statistics for the objective evaluation datasets. For filtered datasets, we report the sample counts as "original → used"; otherwise we report the original sample count directly.

Evaluation dataset	Samples
AudioCaps	957
MusicCaps	4950
LibriTTS	4837 → 3835
MECAT 00A	848
MECAT 0M0	2593
MECAT 0MA	199
MECAT S00	7839 → 3838
MECAT S0A	2439 → 1341
MECAT SM0	5312 → 2620
MECAT SMA	643 → 283
MECAT	19873 → 11722

G.2 Objective Metrics

We use the following objective metrics for evaluation:

FAD & FD Following previous work Liu et al. (2023b), we report FAD_{VGG} and FD_{PANNS} to measure the distributional similarity between generated and reference audio based on VGGish Hershey et al. (2017) features and PANNs CNN14 Kong et al. (2020) features, respectively. We compute these metrics using the AudioLDM evaluation toolkit[2].

[2] https://github.com/haoheliu/audioldm_eval

CLAP & GLAP We use CLAP Elizalde et al. (2023)[3] and GLAP Dinkel et al. (2025)[4] to measure semantic relevance between generated audio and the input prompt. Both scores are computed as cosine similarity between audio and text embeddings. On MECAT, all reference texts for these metrics use the overall caption of each sample, i.e., <|caption|>. Compared with CLAP, GLAP is trained with broader speech and multilingual supervision, making it more sensitive to linguistic content while remaining effective for general audio-text matching.

WER We use Word Error Rate (WER) to evaluate how accurately the generated speech matches the target transcription. To reduce transcription hallucinations on acoustically complex mixed-audio samples, we use the NeMo ASR model[5] for transcription.

UTMOSv2 We use UTMOSv2 Baba et al. (2024)[6] as a reference-free metric for overall speech quality assessment. Unlike reference-based metrics, UTMOSv2 predicts a perceptual quality score directly from the generated waveform and does not require paired ground-truth speech. This is particularly useful in our setting, where speech often co-occurs with music or sound effects, making ASR-based metrics more sensitive to background interference. UTMOSv2 therefore provides a complementary view of perceived speech quality beyond transcription accuracy.

multiFLAM Following AudioChat Chen et al. (2026), we use FLAM Wu et al. (2025) to measure frame-level component alignment in mixed audio. We compute frame-level similarities between generated audio and each component description (the <|speech|>, <|music|>, and <|sfx|> fields), then take the maximum over time for each component and average over the components present. The resulting score measures the presence of the described components but does not assess their order, duration, or temporal overlap.

For reproducibility, Table A9 also lists the exact Hugging Face Hub checkpoints used for all baseline systems. We include these identifiers because performance can vary substantially across different releases, scales, and checkpoint variants of the same model family.

Table A9: Exact model versions used for baseline comparison.

Model	Hugging Face Hub name
AudioLDM2	haoheliu/audioldm2-full
TangoFlux	declare-lab/TangoFlux
MusicGen	facebook/musicgen-large
Qwen3-TTS	Qwen/Qwen3-TTS-12Hz-1.7B-VoiceDesign
AudioX	Zeyue7/AudioX
UniFlow-Audio	wsntxxn/UniFlow-Audio-large

G.3 Prompt Details for Objective Evaluation

To ensure a fair evaluation, we adopt different prompt construction strategies for different systems and evaluation datasets, so as to better match each model's input format and preference. Below we summarize the prompt design used for each benchmark.

AudioCaps and MusicCaps. For this benchmark group, the prompt construction is as follows:

- **Dasheng AudioGen:** We prepend the original dataset prompt with the special token <|caption|> to keep the input format consistent with the model's training setup.
- **Other baselines:** We directly use the original text prompt provided by the dataset.

LibriTTS. For LibriTTS, different systems use different prompt formats:

- **Qwen3-TTS:** We feed the transcript directly as input and leave the instruction field empty.
- **Dasheng AudioGen (structured):** We construct a prompt containing both a scene-level description and an <|asr|> field. For example, when the transcript is *"I can't play with you*

[3] https://huggingface.co/lukewys/laion_clap, using 630k-audioset-fusion-best.pt

[4] https://huggingface.co/mispeech/GLAP

[5] https://huggingface.co/nvidia/stt_en_conformer_transducer_xlarge

[6] https://github.com/sarulab-speech/UTMOSv2

like a little boy any more," he said slowly., the structured prompt is <|*caption*|> *Studio-quality, high-fidelity audiobook recording* <|*asr*|> *"I can't play with you like a little boy any more," he said slowly.*

- **Dasheng AudioGen (unstructured):** We use a plain text prompt without special fields, e.g., *Studio-quality, high-fidelity audiobook recording with content "I can't play with you like a little boy any more," he said slowly.*

This design improves transcript adherence while preserving clear, high-quality speech generation.

MECAT. Because MECAT contains more complex multimodal audio scenes, prompt construction differs substantially across systems:

- **Dasheng AudioGen (structured):** We use the full multi-view structured caption.
- **Dasheng AudioGen (unstructured):** We use only the content associated with `<|caption|>`.
- **TangoFlux and MusicGen:** We use all available information from the multi-view structured caption to avoid information loss.
- **Qwen3-TTS:** The transcript input is taken from the content of `<|asr|>`. If this field is absent, the transcript field is left empty. The instruction field is constructed by concatenating all remaining textual content other than `<|asr|>`.
- **Expert-Pipeline:** When `<|sfx|>` is present, its content is used as the prompt for TangoFlux to generate the sound-effects track. When `<|music|>` is present, its content is used as the prompt for MusicGen to generate the music track. When `<|speech|>` is present, Qwen3-TTS is used to generate the speech track, with `<|asr|>` serving as the transcript and `<|speech|>` as the instruction. The generated tracks are aligned at the beginning and mixed into a single final audio clip. Tracks are resampled to 16 kHz and averaged at the waveform level, without professional mixing operations such as loudness normalization, ducking, temporal arrangement, or peak limiting. For example, for the 0MA category, the Expert-Pipeline uses the contents of `<|music|>` and `<|sfx|>` as prompts for MusicGen and TangoFlux, respectively, and then mixes the generated outputs to form the final audio.

H Subjective Evaluation

H.1 Human Evaluation

We evaluate two dimensions: overall quality (OVL), which measures perceived realism and the harmony of mixed-audio scenes, and text relevance (REL), as detailed in Section H.2. We recruit 20 audio-domain professionals to rate a subset covering approximately 6% of the MECAT English subset. Following a within-subject design, we aggregate scores at the subject level and compare systems with paired Wilcoxon signed-rank tests and Holm correction. Figure A2 shows significance results and Figure A3 provides mean scores.

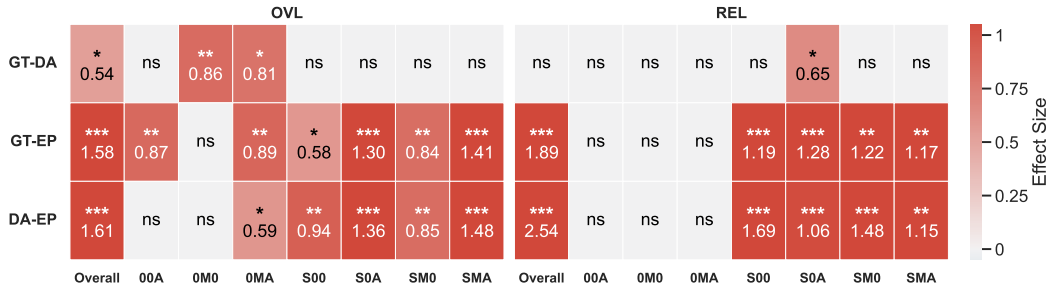

Figure A2: Pairwise human evaluation on MECAT. Cells report significance and Cohen's effect size (d_z) for OVL and REL. GT = Ground Truth, DA = Dasheng AudioGen, EP = Expert-Pipeline. Significance levels, ns: $p > 0.05$, *: $p \leq 0.05$, **: $p \leq 0.01$, ***: $p \leq 0.001$.

For OVL, the gap between ground truth and Expert-Pipeline is very large on the overall set ($d_z = 1.58$, $p < 0.001$) and remains significant in 6 of 7 subcategories, whereas the gap between ground truth and Dasheng AudioGen is much smaller ($d_z = 0.54$, $p < 0.05$) and significant only in 0M0 and 0MA. Dasheng AudioGen also significantly outperforms Expert-Pipeline in all speech-mixed categories (S00, S0A, SM0, SMA) with $p < 0.01$ and $d_z > 0.85$, while showing no significant difference from ground truth.

For REL, Dasheng AudioGen differs significantly from ground truth only in S0A, matching ground-truth text relevance in the other six subcategories and overall. In contrast, the Expert-Pipeline differs significantly from both ground truth and Dasheng AudioGen on the overall set ($d_z = 1.89$ and 2.54) and across all speech-related categories. These findings suggest that while the Expert-Pipeline performs adequately on simple non-speech tasks, it struggles with complex mixed-speech scenarios, which further highlights Dasheng AudioGen's superior controllable generation capabilities in complex, multi-layered scenes.

H.2 Human Evaluation Setting and Results

Figure A3 summarizes the subject-level mean human evaluation scores and standard error (SE) for different systems across all categories on both metrics (OVL and REL) in the MECAT benchmark. Figure A2 further presents the statistical significance and effect sizes of the human evaluation results across different categories.

We recruited 20 audio professionals to evaluate the outputs of different systems on MECAT. Each participant rated 35 test trials. These 35 trials were drawn from the 7 MECAT subcategories, with five trials sampled from each subcategory. Each test trial contained three audio clips paired with the same text description: one generated by Dasheng AudioGen, one generated by the Expert-Pipeline, and the ground truth recording. Participants were asked to rate the audio clips in terms of OVL and REL; the detailed instruction and the scoring criteria are provided in Appendix H.3. The evaluation items for each participant were randomly sampled from MECAT, and the final evaluation covered approximately 6% of the MECAT English subset.

For statistical analysis, we conduct all comparisons at the subject level: for each category and each system, we first average ratings within each annotator and then compare systems using these subject-level means. We use paired Wilcoxon signed-rank tests for significance testing and apply Holm correction within each metric-category block to account for the three pairwise system comparisons. In addition to significance, we report the paired effect size Cohen's d_z, computed from subject-wise score differences. Significance indicates whether a preference is statistically reliable under the current sample size, while effect size measures how large that preference is in practice.

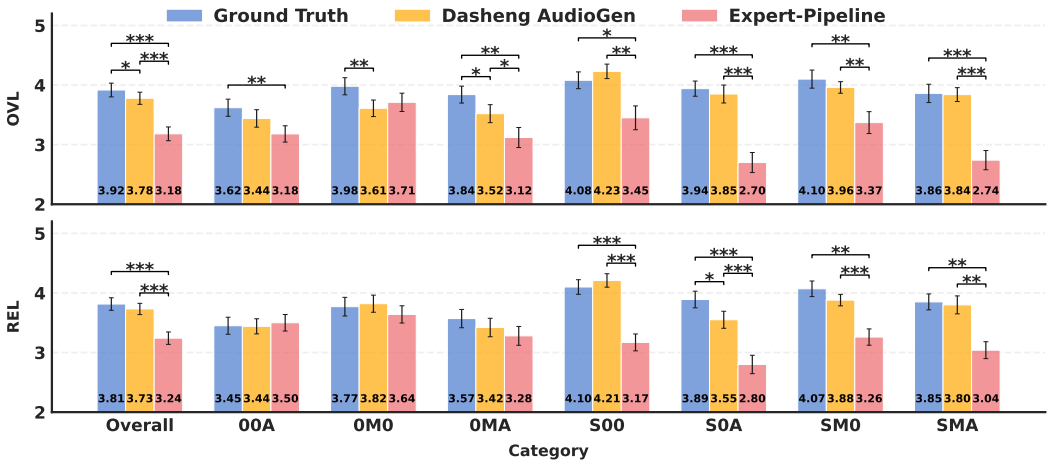

Figure A3: Human evaluation mean score and standard error (SE) on MECAT. We report subject-level mean ratings for overall quality (OVL) and relevance to the prompt (REL) on the overall set and each category. Error bars indicate variability across annotators after subject-level aggregation. Line over two bars indicates a statistically significant difference between the two systems. Significance levels: * ($p \leq 0.05$), ** ($p \leq 0.01$), and *** ($p \leq 0.001$).

H.3 Human Evaluation Instruction and Rating Criteria

Annotators were asked to rate each generated audio clip along two dimensions: overall quality (OVL) and relevance to the input prompt (REL). The instruction below was shown to emphasize that our target is not merely clean isolated sounds, but coherent audio scenes that sound plausibly recorded in the real world. In particular, realism was treated as a core criterion for OVL, while REL focused on whether the generated speech, sound effects, music, environment, and overall atmosphere matched the prompt.

Annotation Instructions

Read the text prompt and listen to the corresponding three audio samples. Please rate each audio on overall quality (OVL) and text relevance (REL) based on the following rating criteria. The main goal of the systems in this evaluation is to generate audio scenes with rich acoustic information that sound like they were recorded in the real world. Please treat **REALISM** as a core criterion, and give higher scores to audio that sounds more realistic and natural.

Rating Criteria for OVL

Please evaluate the overall quality and perceived realism of the audio. Consider both the intrinsic quality of each sound element (speech, sound effects, music) and, when multiple sounds are present, how natural and harmonious their mix is.

1. **Very poor.** The audio contains severe noise, clipping, or strong AI-generated artifacts and feels completely unrealistic. If multiple sounds are mixed, the result is extremely chaotic, heavily conflicting, and unbearable to hear.
2. **Poor.** Individual sounds have obvious artifacts, distortion, or unnaturalness and clearly sound generated. If multiple sounds are mixed, they feel crudely stitched together, interfere with each other, or are severely imbalanced in volume, making the result uncomfortable to hear.
3. **Average.** Individual sounds are basically clear, with occasional minor noise or AI artifacts, and only moderate realism. If multiple sounds are mixed, the parts are distinguishable and do not severely clash, but the layering is weak and the result feels ordinary.
4. **Good.** Individual sounds are clear and natural, and the listening experience is close to a real acoustic environment. If multiple sounds are mixed, they are blended fairly well, with a clear sense of foreground and background, resulting in a comfortable and harmonious sound.
5. **Excellent.** The audio quality is outstanding, with almost no AI artifacts and a highly realistic sound. If multiple sounds are mixed, all elements are integrated beautifully with rich layering, comparable to real-world field recordings or professional film and TV sound production, and highly immersive.

Rating Criteria for REL

Please evaluate how well the generated audio matches the input text, including whether it contains the requested dialogue, sound effects, music style, recording environment, and overall atmosphere.

1. **Completely irrelevant.** The generated audio has nothing to do with the input text, and all required elements are wrong or missing.
2. **Low relevance.** Most key descriptions are missing. For example, the dialogue may be completely incorrect, or the specified sound effects or music style may be absent.
3. **Basically relevant.** The audio matches the main intent of the text and includes the major sound elements, but some specific audio elements are still missing.
4. **Highly relevant.** Most of the text description is reproduced accurately. The speech, sound effects, music, and recording environment are basically correct, with only minor deviations in detail.
5. **Perfect match.** All detailed requirements of the text are reproduced accurately. The elements are complete, the pronunciation is correct, and the audio fully matches the atmosphere and emotion described in the text.

H.4 Human Evaluation Interface

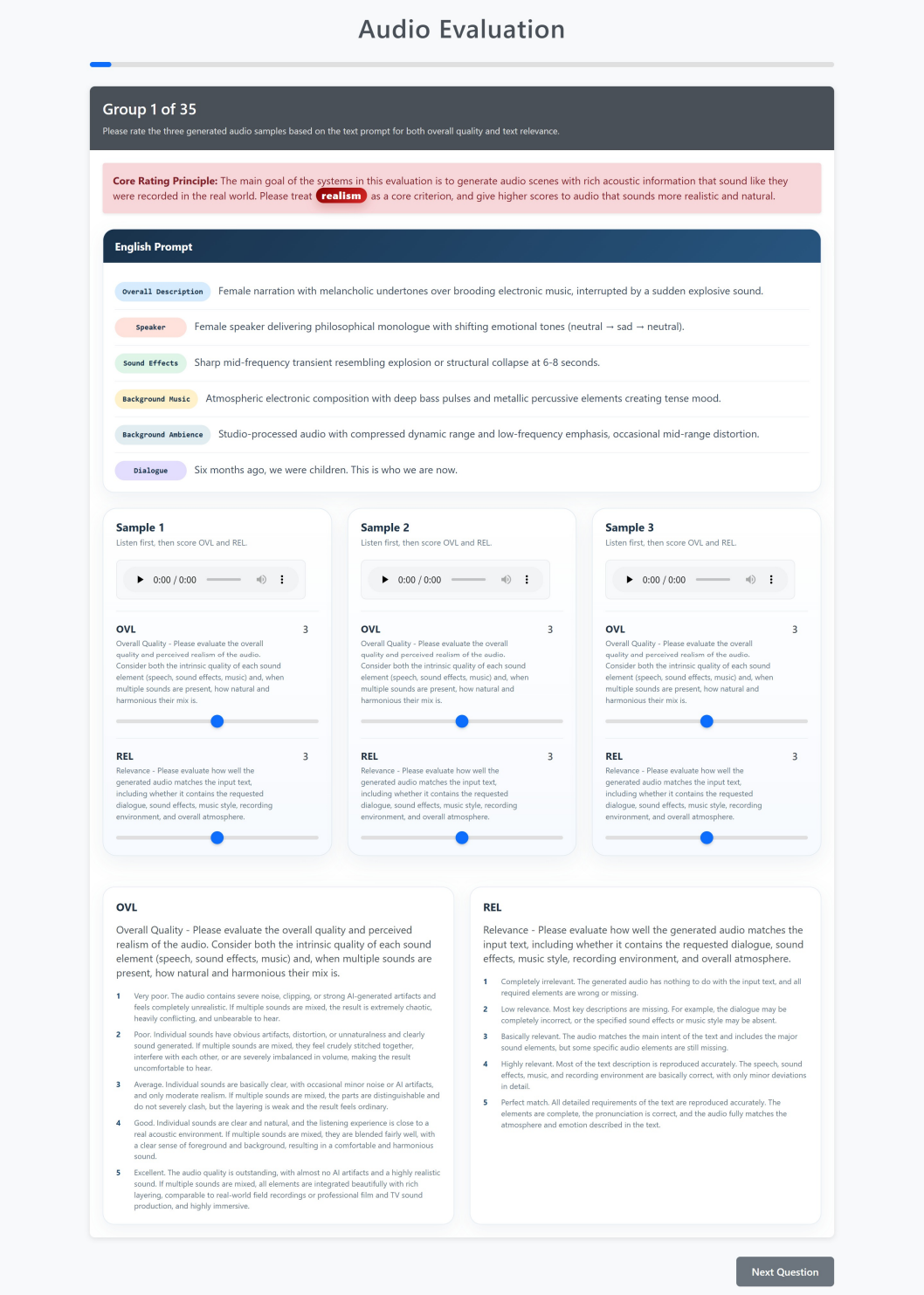

Figure A4: Human evaluation interface screenshot. Annotators rated Overall Quality (OVL) and Text Relevance (REL) for each audio sample.

I LLM Evaluation

I.1 PAFI Setting and Results

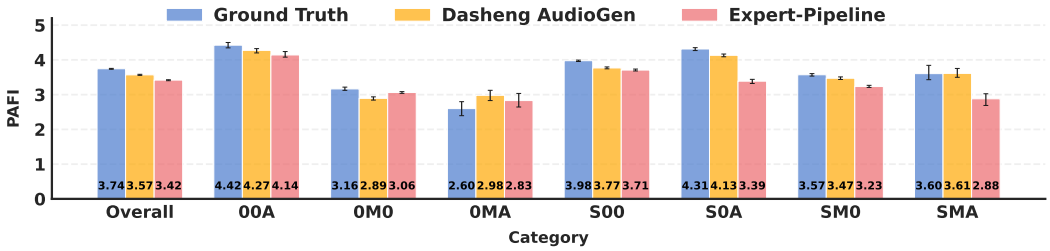

Figure A5: Mean PAFI scores and bootstrap 95% confidence intervals on MECAT. Higher is better.

Since human evaluation covers only a subset of MECAT, we introduce Physical Acoustic Fidelity Index (PAFI) as a complementary metric for human evaluation. PAFI is an LLM-as-a-judge metric powered by Gemini-3.1-Pro, which focuses on physical acoustic fidelity, including spatial consistency, reverberation coherence, and physically plausible source interaction. The prompt used for PAFI scoring is provided in Section I.3. We report PAFI results for Ground Truth, Dasheng AudioGen, and the Expert-Pipeline on the overall set and on each MECAT category.

Because PAFI produces a single score for each audio sample, all statistical analyses are conducted at the sample level. For each category and each system pair, we perform paired Wilcoxon signed-rank tests, and apply Holm correction across the three pairwise comparisons within the same category at sample-level. We also report the paired effect size Cohen's d_z computed from the aligned score differences. Figure A5 reports the mean scores together with bootstrap 95% confidence intervals, and Figure A6 summarizes the corresponding pairwise significance and effect sizes.

The results show a clear overall ordering. On the full evaluation set, the mean PAFI scores are 3.74 for Ground Truth, 3.57 for Dasheng AudioGen, and 3.42 for the Expert-Pipeline. All three overall pairwise comparisons remain significant after Holm correction, with Dasheng AudioGen ranking between Ground Truth and the Expert-Pipeline. The overall advantage of Dasheng AudioGen over the Expert-Pipeline has $d_z = 0.117$ and $p < 0.001$. The largest advantages over the Expert-Pipeline appear in speech-containing mixed categories such as S0A and SMA. In particular, on SMA, Dasheng AudioGen (3.611) is nearly tied with Ground Truth (3.604) while remaining far above the Expert-Pipeline (2.880). The difference from Ground Truth is not statistically significant ($p > 0.05$), whereas the advantage over the Expert-Pipeline has $d_z = 0.45$ and $p < 0.001$, indicating better preservation of physically coherent interactions among overlapping audio elements. Overall, these results show that PAFI captures meaningful system-level differences, and Section I.2 further shows that its system preferences are strongly aligned with human judgment.

I.2 PAFI-Human Evaluation Consistency

To evaluate whether the proposed PAFI metric aligns with human preference, we further examine consistency at the level of *relative system preference*. Human evaluation covers only a subset of MECAT, whereas PAFI is computed on the full benchmark. To make the two sources directly comparable, we select the intersection between the human-rated subset and the PAFI-scored set, and compare them at the sample level. For each category and each system pair, we compute paired Cohen's d_z for both human OVL and PAFI. The sign follows the order of the pair label: for a pair A vs. B, we compute the signed difference as A minus B, so a positive effect size favors A and a negative one favors B. For example, a positive effect size for GT-EP (GT vs. Expert-Pipeline) on the OVL metric in the SM0 category indicates that human raters preferred the ground-truth system over the Expert-Pipeline.

Figure A7 visualizes the distribution of paired effect sizes derived from human OVL and PAFI across categories and system pairs. We evaluate consistency from two perspectives: whether the two effect sizes have the same sign, and how strongly they correlate across all category and system pairs.

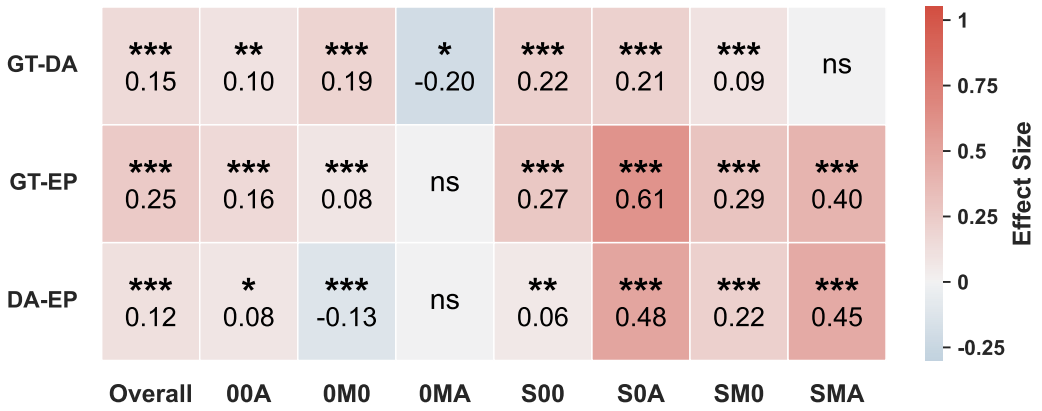

Figure A6: Pairwise significance and effect size summary on PAFI. Each cell compares a pair of systems within one category. GT = Ground Truth, DA = Dasheng AudioGen, EP = Expert-Pipeline. Significance levels: ns ($p > 0.05$), * ($p \leq 0.05$), ** ($p \leq 0.01$), and *** ($p \leq 0.001$).

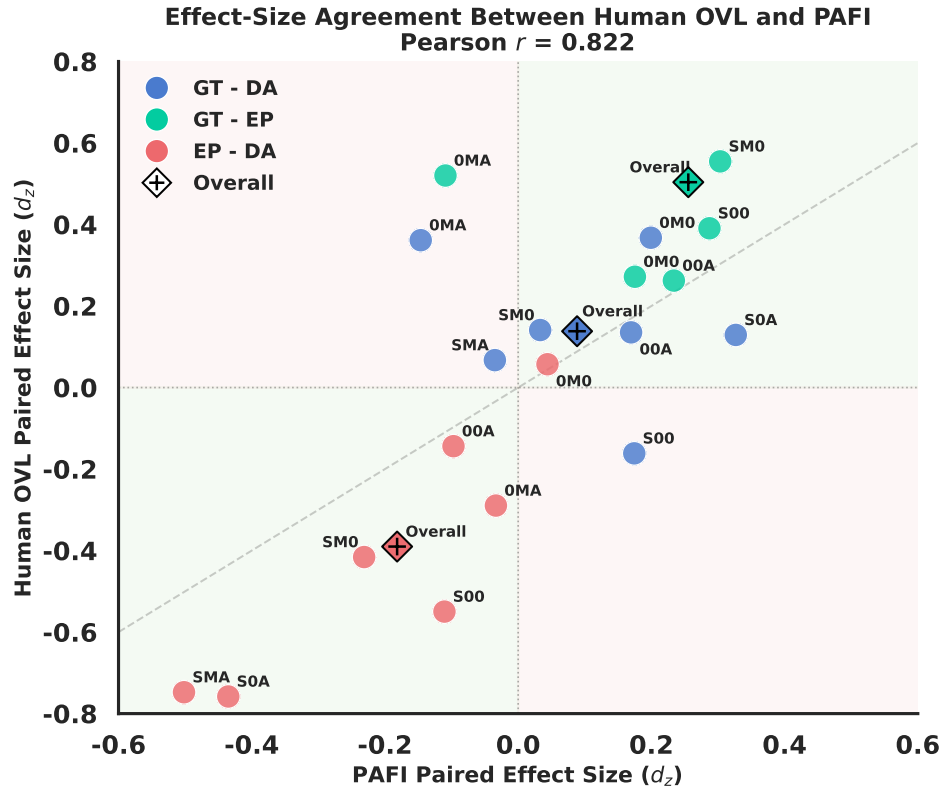

Figure A7: Agreement between human OVL and PAFI in paired effect sizes. Each point corresponds to one system pair in one category. The horizontal axis shows the paired effect size computed from PAFI, and the vertical axis shows the corresponding paired effect size computed from human OVL. The diagonal dashed line indicates perfect agreement in both sign and magnitude. Green-shaded quadrants indicate sign agreement, and red-shaded quadrants indicate sign disagreement. GT-DA means Ground Truth vs. Dasheng AudioGen, GT-EP means Ground Truth vs. Expert-Pipeline, and EP-DA means Expert-Pipeline vs. Dasheng AudioGen.

The results show that PAFI captures human preference reasonably well at the level of relative system comparison. Across the 21 category-level pairwise comparisons, the effect-size direction agrees in 17

cases, corresponding to an agreement ratio of 81.0%. In addition, the effect sizes derived from human OVL and PAFI are positively correlated, with Pearson $r = 0.8224$ and $p < 0.001$. Notably, the three overall comparisons all lie close to the ideal-fit diagonal, further indicating that PAFI is reliable for overall system-level ranking.

Most disagreements occur in comparisons where the effect size is already close to zero, suggesting borderline cases rather than systematic contradictions. Moreover, 3 disagreements out of 4 occur in the Dasheng AudioGen vs. Ground Truth comparisons, which suggests that Dasheng AudioGen is already close to ground truth in these categories and therefore harder to distinguish consistently.

Overall, these results support the use of PAFI as a useful automatic metric for system ranking, while also indicating that human evaluation remains necessary for resolving fine-grained differences in more challenging cases.

I.3 Physical Acoustic Fidelity Index (PAFI) Prompt

Role

You are a Senior Acoustic Engineer and an expert in Computational Auditory Scene Analysis (CASA). You possess profound knowledge of physical acoustics, psychoacoustics, and spatial audio coding. You are highly sensitive to physical acoustic anomalies, such as unnatural reverberation, phase cancellations, or spectral gaps, and can accurately judge the physical authenticity of an audio signal.

Task

Perform a deep physical-modeling analysis of the provided audio, or of its audio description. Your objective is to determine whether the audio follows the acoustic laws of the real physical world, and whether it should be classified as an authentic capture, spliced tracks, or artifact-heavy AI audio.

Evaluation Dimensions

1. **Spatial Consistency**: Determine whether all sound sources are situated within a unified acoustic soundfield. Consider whether early reflections, distance cues, HRTF patterns, and high-frequency attenuation match the implied space.
2. **Reverberation Coherence**: Determine whether the primary subject and background sounds share a consistent reverb profile. Strong mismatch in dryness or RT60 across sources should be treated as evidence of splicing or artificial composition.
3. **Dynamic Layering & Masking**: Determine whether simultaneous sources interact naturally. Consider masking effects, depth ordering, and whether overlaps sound physically plausible rather than phase-conflicting.
4. **Environmental Immersion**: Determine whether the audio contains a credible noise floor and a coherent sense of place, including room tone, low-frequency disturbances, reflections, and diffusion.
5. **Physical & Kinematic Logic**: Determine whether source motion obeys acoustic physics, including Doppler shift and inverse-square SPL changes.

Scoring Rubric (0–5)

- **0**: Severe violation. Obvious phase cancellations, fractured spectra, mismatched reverberation, or clearly broken splicing / AI artifacts.
- **1**: Barely acceptable. The scene is recognizable, but the acoustics are highly unstable or disconnected.
- **2**: Flawed. Basic acoustic logic is present, but key physical cues are missing or artifacted.
- **3**: Adequate. The soundfield is mostly coherent, with only mild artificial traces.

- **4**: Highly simulated. Physical consistency is strong and difficult to distinguish from professional production.
- **5**: Authentic perfection. The audio is fully coherent and indistinguishable from a high-fidelity real-world recording.

Output Format

```
{
  "score": <int>,
  "reason": "<string>"
}
```

Now, please process the following user input:

```
{{user_input_audio}}
```

J Prompt of Agentic Prompt Refiner

Role

You are an expert Audio Scene Architect and Foley Designer. Your task is to deeply understand the user's natural language audio description, and translate it into a highly structured JSON format for a **10-second audio clip**, based on real-world acoustic logic and scene realism.

Task

Process the user's natural language input (which may be in any language), extract the audio elements, and output a JSON object containing the following specific keys.

JSON Schema Definition

- `Caption`: (String, Required) The overall, comprehensive description of the 10-second audio scene.
- `Speech`: (String, Optional) Speaker identity (e.g., middle-aged man, energetic girl) and speaking style (e.g., deep voice, anxious, echoing).
- `ASR`: (String, Optional) The actual transcript / spoken dialogue. **Must ONLY contain the spoken words. NO speaker labels (e.g., do not use "Man A:", "Speaker 1 says:").**
- `SFX`: (String, Optional) Specific sound effects present in the audio (e.g., footsteps, doorbell, dog barking).
- `Music`: (String, Optional) Description of background music (e.g., soft jazz, tense orchestral).
- `ENV`: (String, Optional) Environmental or ambient background noise (e.g., city bustle, forest wind and crickets).

Crucial Generation Rules

1. **Absolute English-Only Output**: ALL fields in the output JSON MUST be generated in English, regardless of the language of the user's input. The `ASR` field must also be entirely in English. Do NOT output any other languages.
2. **10-Second Constraint**: The entire audio scene is exactly 10 seconds long. If generating `ASR`, keep the dialogue concise and realistic for a 10-second window (usually 1-2 short sentences).
3. **Strict ASR Formatting**: The `ASR` field must ONLY contain the raw spoken text. Never include names, character tags, or action descriptions within the ASR string.

(Correct: "Watch out for that car!" | Incorrect: "Man: Watch out for that car!"). If there are multiple speakers, just combine their dialogue naturally without labels.

4. **Scene Enrichment**: Act as a sound designer and logically enrich the scene. If the user says "at a train station," automatically add train horn (SFX) and crowd murmurs (ENV).
5. **Logical Nulls**: Except for `Caption` (mandatory), if a field is not mentioned by the user AND makes no logical sense in the scene, set its value to `null`.
6. **Strict Output**: Output ONLY valid JSON. Do not include markdown blocks like ```` ```json ```` or any explanatory text.

Examples

User Input: "A man complaining about the weather on a rainy street."
Output:

```
{
  "Caption": "A man complaining about the rainy weather on a wet city street
    .",
  "Speech": "A frustrated middle-aged male voice speaking loudly.",
  "ASR": "Damn it! I can't believe I forgot my umbrella again!",
  "SFX": "Heavy raindrops hitting the pavement, a car splashing water.",
  "Music": null,
  "ENV": "Urban street background noise, distant traffic."
}
```

User Input: " 两个女孩在游乐场里兴奋地指着过山车"
Output:

```
{
  "Caption": "Two young girls expressing excitement over a roller coaster at a
     busy amusement park.",
  "Speech": "Two young female voices, cheerful and energetic, speaking over
    each other slightly.",
  "ASR": "Look at that one! It goes completely upside down! Let's go line up
    right now!",
  "SFX": "The mechanical clanking of a roller coaster climbing, followed by a
    loud whoosh.",
  "Music": "Upbeat, faint carnival music playing from nearby speakers.",
  "ENV": "Crowd chatter, distant screams of thrill, general theme park
    ambiance."
}
```

User Input: "A calm ambient electronic music track for studying."
Output:

```
{
  "Caption": "A relaxing and continuous ambient electronic music track
    designed for deep focus.",
  "Speech": null,
  "ASR": null,
  "SFX": null,
  "Music": "Slow, pulsing ambient synthesizer pads with a very soft, steady
    rhythmic beat.",
  "ENV": null
}
```

Now, please process the following user input:

`{{user_input}}`