ANYAUDIO: ACCURATE, VERSATILE, AND TRAINING-FREE AUDIO CONTROL VIA STEERING GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Text-to-audio generation systems have been capable of generating high-fidelity samples from natural language prompts with a scalable diffusion transformer (DiT). Recent advances have further allowed fine-grained controllability, which mainly spans precise timing control, long-form audio generation, as well as acoustic control (*e.g.*, pitch and loudness). However, existing methods usually leverage conditional learning, either from scratch or fine-tuning, or training-freely explore latent manipulation, both of which hinder them to achieve versatile audio control efficiently. In this work, we present a holistic audio control framework, AnyAudio, which precisely steers the sampling trajectory to desired target, fully exploiting the advantages of pre-trained audio DiT. Specifically, we first design Timing Attention Control in the condition space, aligning each sound event with their designated time intervals for precise event timing. Furthermore, we present Latent Spectral Decoupling in the diffusion intermediate representation space, fusing overlapping segments in the frequency domain for coherent long-form generation. Then, we develop Reference Attribute Control in waveform space, matching differentiable reference attributes for fine-grained physical control. These cross-space control signals share a unified paradigm of differentiable matching loss optimized via gradient descent, which can be integrated into a diffusion sampling trajectory in an accurate, composable, and training-free manner. Extensive experiments demonstrate that AnyAudio achieves state-of-the-art performance in control of timing, long-form generation, and acoustic characteristics, independently and jointly, without training resource. Demo Page: `https://audioguidance.github.io/`.

1 INTRODUCTION

Text-to-audio (TTA) generation systems have been advanced with scalable diffusion transformer (DiT) models in the latent space compressed from audio data space (Evans et al., 2024b; Lee et al., 2024; Hung et al., 2024; Valle et al., 2025), where the complex text-conditioned distribution of large-scale audio datasets can be faithfully reconstructed with iterative noise-to-data sampling steps. Recent research starts to explore fine-grained control in audio generation, which mainly includes *timing control* (Bar-Tal et al., 2023; Du et al., 2024; Wang et al., 2025b; Zhu et al., 2026), *e.g.*, "2s-5s, bird chirping", *long-form generation* (Dai et al., 2025; Jiang et al., 2025b; Guo et al., 2025), *e.g.*, minute-long generation with a TTA system pre-trained with 10-second short samples, and *reference control*, *e.g.*, pitch and loudness control from a reference sample (Novack et al., 2024b; García et al., 2025; Yuan et al., 2026; Novack et al., 2026).

Existing controllable audio generation methods can be categorized into two main classifications: conditional learning or training-free latent manipulation. In training-based methods, the controllable generation tasks are formulated as capturing various conditional distributions, where the model can be conditioned on a single or multiple control signals. This formulation inevitably requires a training stage for each conditional generation task, either from scratch (Wang et al., 2025b; Liu et al., 2026) or fine-tuning (Bar-Tal et al., 2023; Zhu et al., 2026), which increases training resources and may demand carefully designed conditioning techniques and multi-stage learning strategy (Jiang et al., 2025a; Dai et al., 2026). In training-free methods, the noisy latent representations along the diffusion sampling trajectory are usually manipulated in a pre-defined manner, which may manually craft the latent (Dai et al., 2025) or additionally leverage additional multi-modal large language models

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

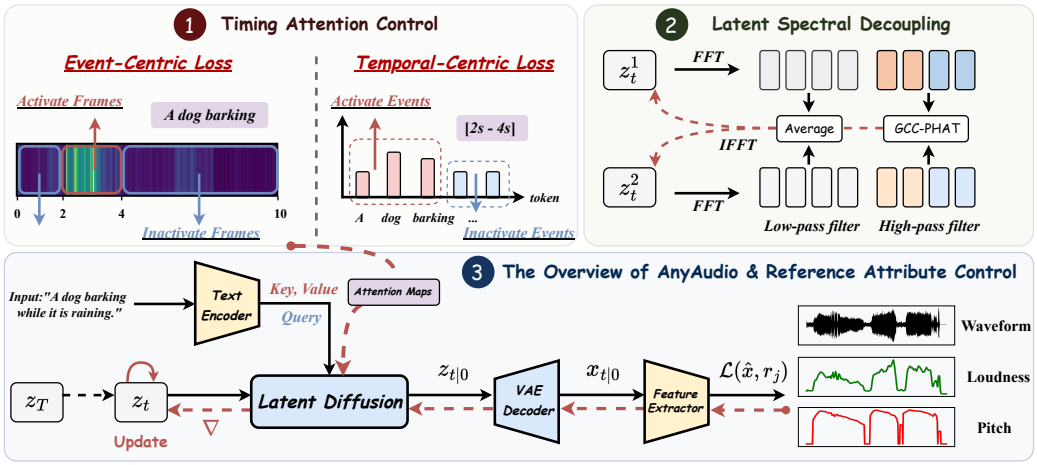


Figure 1: Overview of AnyAudio. The gradients from all modules are back-propagated to update z_t, supporting timing, long-form, and reference attribute control under a unified paradigm.

(MLLMs) (Jiang et al., 2025b) to achieve a specific control function, *e.g.*, timing control or long-form generation. Although these methods have greatly advanced controllable TTA generation, they still struggle to enable versatile audio control in an accurate, flexible, and training-efficient manner.

In this work, instead of designing a conditional learning algorithm or iteratively manipulating the synthesized latent, we present AnyAudio (Figure 1) to fully exploit the advantages of TTA generation systems that are established in the pre-training stage. Specifically, holistic audio control functions are achieved by precisely steering the sampling trajectory to the desired target in a training-free manner, through designed task-specific guidance terms cross spaces. First, the condition space of text-audio attention has established correspondences between text tokens and audio frames, lending itself to event-level timing control. Second, the latent space of audio waveform, which is usually encoded by a 1D VAE (Kingma & Welling, 2013; Evans et al., 2024a) along the temporal axis, has established correspondences between latent frames and time intervals, lending itself to long-horizon structural control. Finally, the data space of the decoded waveform demonstrates acoustic characteristics, where physical attributes can be controlled.

From the perspective of steering generation and given the analysis of the hierarchical space of audio DiT (Peebles & Xie, 2023), our proposed AnyAudio tackles precise timing control, long-form generation, and acoustic control as follows. Timing Attention Control (TAC) operates on text-audio attention to align sound events with their designated time intervals, while addressing the overlap and background dominance issues that arise along the audio temporal axis. Latent Spectral Decoupling (LSD) operates on the latent in the frequency domain, decoupling the fusion across overlapping segments into low-frequency sharing for global consistency and high-frequency alignment for local smoothness. Reference Attribute Control (RAC) operates on the decoded waveform, matching differentiable reference attributes such as loudness and pitch against user-provided references.

Extensive experiments and evaluation show that AnyAudio achieves state-of-the-art performance in the controllable generation of timing, long-forms, and acoustics, outperforming both training-based and training-free baselines. Beyond individual tasks, AnyAudio further supports the joint composition of multiple control objectives within a single inference pass, a capability that may require orchestrating multiple specialized models in previous work, evidenced by both qualitative and quantitative results. Our contributions are summarized as follows.

- We propose AnyAudio, a steering generation framework to enable accurate, versatile, and training-free audio control.
- We design three task-specific control methods, termed TAC, LSD, and RAC, in various spaces of latent audio diffusion, tackling timing control, long-form generation, and reference control.
- AnyAudio shows superior performance compared to training-based and training-free baselines in diverse tasks, and allows a joint composition of diverse controls within a single inference pass.

2 PRELIMINARIES

Latent Audio Diffusion. Recent TTA models design a latent diffusion framework that generates audio by reversing a predefined noising process in a compressed small space. Given a raw waveform x_0 of total duration D (in seconds), a 1D variational autoencoder (VAE) encodes it into a temporal latent $z_0 \in \mathbb{R}^{F\times d}$, where F is the number of latent frames along the temporal axis and d is the channel dimension; each latent frame corresponds to D/F seconds of waveform. A text prompt describing the audio scene is encoded by a frozen text encoder into a conditioning embedding c_{text}. The forward noising process progressively corrupts the latent according to

$$z_t = \sqrt{\bar{\alpha}_t}\, z_0 + \sqrt{1-\bar{\alpha}_t}\, \epsilon, \quad \epsilon \sim \mathcal{N}(0, I), \tag{1}$$

where $\{\bar{\alpha}_t\}_{t=1}^{T}$ is a predefined noise schedule. The reverse denoising process is parameterized by a neural network $v_\theta(z_t, c_{\text{text}}, t)$ that predicts the velocity $v_t = \sqrt{\bar{\alpha}_t}\,\epsilon - \sqrt{1-\bar{\alpha}_t}\, z_0$, conditioned on the text embedding c_{text}. At each denoising step t, the clean latent admits a differentiable estimate:

$$\hat{z}_0 = \sqrt{\bar{\alpha}_t}\, z_t - \sqrt{1-\bar{\alpha}_t}\cdot v_\theta(z_t, c_{\text{text}}, t), \tag{2}$$

which is decoded into a waveform estimate $\hat{x}_0 = \mathcal{D}(\hat{z}_0)$ via the VAE decoder $\mathcal{D}$. Unlike image or video latents that possess spatial dimensions, audio latents in recent works are usually 1D along the temporal axis, with each frame corresponding to a fixed-duration waveform segment.

Steering Generation. The sampling steps of diffusion-based TTA systems shown in Eq. equation 2 iteratively reconstruct the complex text-conditioned distribution of large-scale audio datasets, *i.e.*, $p_\theta(z_0|c_{\text{text}})$, via learned score functions $\nabla_{z_t}\log p_t(z_t|z_0)$, where z_t is constructed from z_0 with Eq. equation 1. For fine-grained controllable audio generation, AnyAudio explores steering the sampling trajectory toward desired target by adding an additional score, namely, the gradient of a predefined value function, $\nabla_{z_t}\mathcal{L}$, at each inference step, *i.e.*, $z_t \leftarrow z_t - \eta_t \nabla_{z_t}\mathcal{L}$. Specifically, we design three value functions across the space of latent audio DiT framework: cross-attention activations within v_θ that route text tokens to audio frames, the clean latent estimation $\hat{z}_0$ in Eq. equation 2, and the decoded waveform estimate $\hat{x}_0 = \mathcal{D}(\hat{z}_0)$, the details of which are introduced in Section 3. Under this formulation, diverse audio-controllable generation tasks can be realized with designed task-specific value functions, and holistic audio control can be achieved by employing these value functions flexibly and composably, *i.e.*, individually or jointly, without requiring any training resource.

3 METHOD

3.1 PROBLEM FORMULATION

Given a pretrained latent audio diffusion model, our goal is to enable training-free control over three complementary aspects of the generated audio: precise timing of sound events, long-horizon structural organization, and fine-grained physical attributes of the output waveform. The user provides a text prompt c together with one or more control objectives below, specified independently or jointly.

Timing Control. The input includes a set of event-level timing annotations $\mathcal{E} = \{(e_i, s_i, o_i)\}_{i=1}^{n}$, where e_i is a keyword or short phrase identifying the i-th sound event and $[s_i, o_i]$ is the desired time interval during which e_i should be audible. These annotations may overlap in time or leave uncovered gaps. Each seconds-level interval $[s_i, o_i]$ maps to a contiguous range of latent frame indices $[f_{s_i}, f_{o_i}]$ via the latent frame rate $f = F/D$.

Long-Form Control. The pretrained model generates audio of at most $D_{\max}$ seconds (typically 10 seconds), whereas practical scenarios often demand longer durations. The input includes a target duration $D_{\text{tgt}} > D_{\max}$, and the goal is to generate a single continuous recording of length D_{tgt} that remains coherent throughout the timeline.

Reference Control. The input includes reference signals $\{(\phi_j, r_j)\}_{j=1}^{J}$, where ϕ_j is a differentiable extractor and r_j is the reference target, specified as a target curve over the audio duration. We support loudness, pitch, and a customized waveform reference within a designated time interval.

3.2 TIMING-CONTROLLED AUDIO GENERATION

Prior timing-controllable TTA methods (Guo et al., 2024; Wang et al., 2025b; Jiang et al., 2025a) typically inject explicit timing signals through dedicated conditioning branches trained end-to-end,

which couples timing control to model retraining and a fixed control interface. We instead propose Timing Attention Control (TAC), a training-free alternative that exploits the text-audio attention within a frozen backbone, where a soft alignment between text tokens and audio frames already emerges and provides a natural handle for event-level timing. At each denoising step t, aggregating the text-audio attention weights across all layers and heads yields a routing matrix $A^t \in \mathbb{R}^{F \times N_p}$, with $A^t_{j,i}$ measuring how strongly audio frame j attends to text token i. Since audio latents are 1D along time, A^t directly couples text tokens to timeline positions, and its rows and columns expose two complementary handles for timing control: per-event temporal activation and per-frame event composition.

Event-Centric Loss. To ensure each event e_i is heard during its designated interval, we guide its temporal activation profile $A^t_{\cdot,i}$ to concentrate within the active frames $[f_{s_i}, f_{o_i}]$ and vanish outside:

$$\mathcal{L}_i^{\text{ec}} = \underbrace{-\frac{1}{|[f_{s_i}, f_{o_i}]|} \sum_{j \in [f_{s_i}, f_{o_i}]} A^t_{j,i}}_{\mathcal{L}_i^{\text{pos}}\text{: activate within } [s_i, o_i]} + \underbrace{\frac{1}{F - |[f_{s_i}, f_{o_i}]|} \sum_{j \notin [f_{s_i}, f_{o_i}]} A^t_{j,i}}_{\mathcal{L}_i^{\text{neg}}\text{: suppress outside } [s_i, o_i]}. \tag{3}$$

While this per-event formulation provides effective timing control, two audio-specific challenges remain. (1) *Overlap Competition*: when multiple events co-occur at the same frame, the fixed attention budget across tokens turns independent per-event optimization into a zero-sum redistribution rather than genuine co-activation. (2) *Background Dominance*: ambient events spanning the entire duration have no inactive frames, so $\mathcal{L}_i^{\text{neg}}$ vanishes and the unchecked $\mathcal{L}_i^{\text{pos}}$ drives background attention upward, crowding out foreground events. We address both with a complementary temporal-centric loss operating on the per-frame attention allocation.

Temporal-Centric Loss. The temporal-centric loss operates on the per-frame composition $A^t_{j,\cdot}$, asking at each frame which events should be in focus. For each frame j, let $\mathcal{A}(j) = \{i \mid j \in [f_{s_i}, f_{o_i}]\}$ denote the active events and $\bar{\mathcal{A}}(j)$ the inactive ones, with mean attention $\mu^{\text{pos}}(j)$ and $\mu^{\text{neg}}(j)$ averaged over the corresponding event sets. Aggregating μ^{pos} and μ^{neg} across frames by a simple mean would let a few poorly controlled frames be masked by the majority of well-behaved ones, leaving audible timing artifacts unpenalized. We therefore adopt LogSumExp aggregation, which concentrates gradient on the worst-performing frames:

$$\mathcal{L}^{\text{tc}} = \underbrace{\tau \log \sum_{j=1}^{F} \exp(-\mu^{\text{pos}}(j)/\tau)}_{\text{raise active-event attention at worst frame}} + \underbrace{\tau \log \sum_{j=1}^{F} \exp(\mu^{\text{neg}}(j)/\tau)}_{\text{suppress inactive-event attention at worst frame}}, \tag{4}$$

where τ controls the sharpness of the aggregation. By operating on the collective attention of all active events at each frame, $\mathcal{L}^{\text{tc}}$ handles overlapping events and scene-wide background sounds, complementing the former to deliver complete timing control.

Timing Attention Control. The effectiveness of the two losses hinges on each event e_i being identifiable as a token group in the caption, but user-provided keywords may not appear verbatim in c. We therefore prepend an LLM-based recaption step that minimally rewrites c into an extended caption c', preserving all original wording and weaving in only the missing events as contiguous spans $\text{span}(e_i)$; the step is skipped when c already covers all events. For each event e_i, we then define its attention profile $a_i^t \in \mathbb{R}^F$ as the average of the attention columns within $\text{span}(e_i)$, replacing the single-token profile $A^t_{\cdot,i}$ used previously. The two losses are then combined into the final timing attention control objective:

$$\mathcal{L}_{\text{TAC}} = \alpha \cdot \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}_i^{\text{ec}} + (1 - \alpha)\, \mathcal{L}^{\text{tc}}, \tag{5}$$

where $\alpha \in [0, 1]$ balances the two terms.

3.3 Long-Form Audio Generation

The most direct route to long-form generation is to train the backbone on long audio (Evans et al., 2024a), but the long-clip training is costly and public datasets are generally restricted to clips of up to

10 seconds (Kim et al., 2019; Gemmeke et al., 2017; Chen et al., 2020). A practical alternative is to keep the backbone fixed at $D_{\max}$ and stitch its outputs: the noise field of length D_{tgt} is partitioned into m overlapping segments of length $D_{\max}$ that are denoised in parallel, with the overlaps averaged at each step (Bar-Tal et al., 2023). This vanilla scheme suffers from two failure modes: (1) *Stylistic Drift*: each segment independently determines its own energy envelope and timbral character, breaking consistency across the timeline; (2) *Phase Cancellation*: small temporal misalignments leave adjacent segments partially anti-phase in the overlap, so direct averaging suppresses transients and fine textures. We propose Latent Spectral Decoupling (LSD), which decouples the fusion in the frequency domain and addresses the two failure modes with complementary operations: the low-frequency band is shared across all segments to anchor a common energy envelope and ambiance throughout the timeline, while within each overlap the high-frequency band is aligned in phase and fused by spectral energy to preserve fine details across the boundary.

Low-Frequency Sharing. Low-frequency components carry the slowly-varying global structure of the audio, including energy envelope, ambiance, and long-horizon timbre, and should remain consistent across all segments to prevent stylistic drift. At each denoising step $z_t \to z_{t-1}$, we apply a 1D FFT along the temporal axis of each segment latent z^i_{t-1} and extract its low-frequency band $\mathrm{FFT}(z^i_{t-1})_{\text{lp}}$ via a low-pass filter with cutoff ratio r. A single shared representation $L^\star$ is then obtained as the closed-form minimizer of the segment-wise reconstruction error:

$$L^\star = \arg\min_{L} \sum_{i=1}^{m} \left\| \mathrm{FFT}(z^i_{t-1})_{\text{lp}} - L \right\|^2 = \frac{1}{m} \sum_{i=1}^{m} \mathrm{FFT}(z^i_{t-1})_{\text{lp}}. \tag{6}$$

We replace each segment's low-frequency bins with $L^\star$ and transform back to the time domain via inverse FFT, leaving each segment to develop its own high-frequency content under its own denoising trajectory. This sharing is applied across all segments during the first T_{freq} denoising steps to establish the global structure, and is restricted to overlapping regions thereafter to maintain the low-frequency baseline required by the alignment step below.

High-Frequency Alignment. High-frequency components carry localized details such as transients, fine textures, and percussive onsets, which must transition smoothly across overlapping regions between adjacent segments. Let $H_a \coloneqq \mathrm{FFT}(z^a_{t-1})_{\text{hp}}$ and $H_b \coloneqq \mathrm{FFT}(z^b_{t-1})_{\text{hp}}$ denote the high-frequency spectra of two adjacent segments a and b within their overlapping region, obtained via a high-pass filter complementary to the low-pass filter above. Direct averaging $(H_a + H_b)/2$ would attenuate these details: a small temporal offset Δt between the two predictions makes their spectra approximately satisfy $H_b(\omega) \approx H_a(\omega) \cdot e^{-\mathrm{j}\omega\Delta t}$, and at high frequencies where $\omega\Delta t$ is no longer small, the two phasors approach anti-phase and average destructively. We therefore align the two spectra in time before fusion. A single global offset $\Delta t^\star$ is estimated via generalized cross-correlation with phase transform (GCC-PHAT) (Knapp & Carter, 1976):

$$\Delta t^\star = \arg\max_{\Delta t} \ \mathrm{Re}\left[\sum_{\omega} \frac{H_a(\omega)\, H_b^*(\omega)}{\left| H_a(\omega)\, H_b^*(\omega) \right|} \, e^{-\mathrm{j}\omega\, \Delta t} \right], \tag{7}$$

and applied to H_b as a linear-phase rotation $\tilde{H}_b(\omega) = H_b(\omega) \cdot e^{\mathrm{j}\omega\, \Delta t^\star}$, which rigidly shifts H_b to align with H_a while preserving its internal phase structure. The two phase-coherent spectra are then fused via an amplitude-weighted least-squares objective with closed-form solution:

$$H^\star = \arg\min_{H} \ \left\| |H_a| \odot (H - H_a) \right\|^2 + \left\| |\tilde{H}_b| \odot (H - \tilde{H}_b) \right\|^2 \ = \ \frac{|H_a|^2 \odot H_a + |\tilde{H}_b|^2 \odot \tilde{H}_b}{|H_a|^2 + |\tilde{H}_b|^2}. \tag{8}$$

Each frequency bin is dominated by whichever segment exhibits larger spectral energy at that bin, mirroring SNR-weighted combining in array signal processing. The fused spectrum $H^\star$ is mapped back to the time domain and written into the overlap region.

3.4 Reference Audio Generation

Existing methods for fine-grained acoustic control inject reference attribute curves as auxiliary conditioning signals trained end-to-end with the diffusion backbone (Zhu et al., 2026; Guo et al., 2024). We instead propose Reference Attribute Control (RAC), which matches user-provided

references against attributes extracted from the decoded waveform $\hat{x}_0 = \mathcal{D}(\hat{z}_0)$ in a training-free manner. Reference attributes are pointwise functionals of the waveform, so their target curves can be matched directly via gradient descent on z_t without auxiliary conditioning networks.

Time-Varying Reference Attributes. Each reference attribute is defined by a differentiable extractor ϕ that maps $\hat{x}_0$ to a continuous curve over the audio duration. We instantiate two attributes in this work: loudness, computed as the frame-level RMS energy on the decibel scale, and pitch, estimated using a differentiable harmonic-comb response followed by soft-argmax proxy in Appendix B.1. The 1D temporal VAE encodes the waveform with strict temporal locality, allowing ϕ to support both localized attribute control over arbitrary intervals and long-horizon control spanning the entire audio duration. Given a target curve r, the attribute contributes a matching loss:

$$\mathcal{L}(\phi, r) = \left\| \phi(\hat{x}_0) - r \right\|^2. \tag{9}$$

The loss is restricted to the specified interval via a binary mask. A user-provided waveform reference is accommodated under the same form by setting ϕ to the identity.

Reference Attribute Control. The RAC objective $\mathcal{L}_{\mathrm{RAC}}$ collects the matching losses of all active references, with its gradient back-propagated through $\mathcal{D}$ and ϕ to update z_t at each denoising step. This requires neither a pretrained surrogate nor a dedicated conditioning branch, and accommodates any new attribute by plugging in its differentiable extractor.

3.5 Overall AnyAudio Inference

Algorithm 1 summarizes the inference procedure of AnyAudio. Given a target duration D_{tgt}, the timeline is partitioned into m overlapping segments determined by D_{tgt} and the backbone capacity $D_{\max}$, and the user-provided event annotations $\mathcal{E}$ and reference attributes $\mathcal{R}$ are projected onto each segment by interval intersection. At each denoising step, TAC and RAC compute their matching losses on the active annotations and references for every segment, and the latent z_t is refined for K_t inner iterations under the time-travel trick (Yu et al., 2023; Shen et al., 2024; Jang et al., 2025), which interleaves guidance updates with renoising to strengthen guidance without additional denoising steps. After the solver advances z_t to z_{t-1}, LSD shares its low-frequency band across all segments and aligns the high-frequency band over their overlaps when $m > 1$. The two operations are alternated until z_0, with absent controls contributing nothing to the update, allowing any subset of timing, long-form, and reference attribute controls to be specified jointly within a single inference pass.

Algorithm 1 AnyAudio Inference

Require: c, D_{tgt}, $\mathcal{E}$, $\mathcal{R}$, ϕ, T_{freq}, total steps T, repeat step K_t, step size η_t, model $v_\theta(\cdot,\cdot,\cdot)$

Split $[0, D_{\mathrm{tgt}}]$ into m overlapping segments; restrict $\mathcal{E}, \mathcal{R}$ accordingly
$z_T \sim \mathcal{N}(0, I)$
for $t = T, \ldots, 1$ **do**
  for $k = 1, \ldots, K_t$ **do**
    $v_t, A^t \leftarrow v_\theta(z_t, c, t)$
    $\hat{z}_0 \leftarrow \sqrt{\bar{\alpha}_t} z_t - \sqrt{1-\bar{\alpha}_t}\, v_t$
    $\mathcal{L}_t \leftarrow \mathcal{L}_{\mathrm{TAC}}(A^t, \mathcal{E}, t)$
    $\mathcal{L}_t \mathrel{+}= \mathcal{L}_{\mathrm{RAC}}(\phi(\mathcal{D}(\hat{z}_0)), \mathcal{R}, t)$
    $z_t \leftarrow \text{Time-Travel}(z_t, \hat{z}_0, \nabla_{z_t}\mathcal{L}_t)$
  end for
  $z_{t-1} \leftarrow \text{DPM-Solver}(z_t, c, t)$
  if $m > 1$ **then**
    $z_{t-1} \leftarrow \text{LowFreq-Share}(z_{t-1}, t)$
    $z_{t-1} \leftarrow \text{HighFreq-Align}(z_{t-1})$
  end if
end for
return $\hat{x}_0 = \mathcal{D}(z_0)$

4 Experiments

4.1 Experimental Settings

Model Configurations. We adopt a pretrained DiT-based latent audio diffusion model following the Stable Audio architecture (Evans et al., 2024a;b; 2025), conditioned on text via cross-attention from a frozen Flan-T5 large (Chung et al., 2024) text encoder. The model generates 10-second audio segments at 16 kHz, encoded by a DAC-VAE (Evans et al., 2024a) with a latent temporal resolution of 25 Hz. The denoising network is trained from noised ground-truth latents under the v-objective (Salimans & Ho, 2022). At inference, we use DPM-Solver++ (Lu et al., 2025) for 100 sampling steps. Further model details are provided in Appendix A, and task-specific implementation details for each control module are described in Appendix B.1.

Table 1: Objective and subjective evaluation results on the AudioCondition test set. For each metric, the best result is bold and the second-best is underlined. AnyAudio (base) denotes generation from the text caption only, without any control modules.

Method	Timing			Loudness		Pitch		Quality			
	Eb↑	At↑	MOS↑	MAE↓	MOS↑	MAE↓	MOS↑	FAD↓	KL↓	CLAP↑	OVL↑
Ground Truth	43.37	67.53	$4.56_{\pm 0.08}$	-	$3.21_{\pm 0.14}$	-	$3.05_{\pm 0.16}$	-	-	0.377	$4.33_{\pm 0.09}$
AudioLDM	9.21	49.35	$1.58_{\pm 0.24}$	11.51	$2.01_{\pm 0.21}$	270.38	$1.83_{\pm 0.19}$	4.14	2.18	0.378	$2.54_{\pm 0.27}$
AudioLDM 2	6.33	47.67	$2.04_{\pm 0.20}$	9.87	$1.87_{\pm 0.18}$	363.05	$1.94_{\pm 0.22}$	3.00	<u>1.77</u>	0.423	$2.66_{\pm 0.23}$
Tango	7.76	45.52	$1.93_{\pm 0.22}$	9.68	$1.92_{\pm 0.20}$	247.26	$1.90_{\pm 0.21}$	2.16	2.11	0.364	$2.79_{\pm 0.19}$
Stable Audio Open	3.37	45.21	$1.66_{\pm 0.25}$	13.81	$1.95_{\pm 0.17}$	339.61	$1.75_{\pm 0.24}$	3.12	2.87	0.280	$3.57_{\pm 0.15}$
TangoFlux	5.93	53.97	$1.45_{\pm 0.28}$	18.52	$1.83_{\pm 0.22}$	281.79	$1.62_{\pm 0.20}$	2.58	1.86	**0.494**	$3.42_{\pm 0.18}$
AnyAudio (base)	9.64	66.59	$1.75_{\pm 0.19}$	18.00	$1.75_{\pm 0.24}$	320.19	$1.64_{\pm 0.23}$	1.90	1.84	<u>0.430</u>	**3.83**$_{\pm 0.11}$
AudioComposer-S	43.51	60.83	$3.37_{\pm 0.16}$	26.57	$2.12_{\pm 0.20}$	198.20	$2.41_{\pm 0.17}$	4.57	1.99	0.284	$2.67_{\pm 0.21}$
T2A-ControlNet	47.98	67.92	-	<u>4.08</u>	-	**119.28**	-	-	-	-	-
T2A-Adapter	**54.36**	<u>68.26</u>	$3.65_{\pm 0.13}$	**1.40**	$2.94_{\pm 0.12}$	148.02	$2.63_{\pm 0.15}$	<u>1.87</u>	**1.53**	0.307	$3.59_{\pm 0.14}$
AnyAudio	<u>48.62</u>	**75.12**	**4.12**$_{\pm 0.10}$	5.22	**3.07**$_{\pm 0.11}$	<u>137.98</u>	**2.86**$_{\pm 0.12}$	**1.78**	1.80	0.427	$3.81_{\pm 0.13}$

Evaluation Datasets. We evaluate AnyAudio on three publicly available datasets, each targeting a distinct aspect of our framework. For multi-condition controllable generation, we adopt the AudioCondition (Guo et al., 2024) test set. The test split contains 1110 clips covering the event vocabulary supported by the PB-SED (Mesaros et al., 2016) evaluator, and provides simultaneous supervision for timing, loudness, and pitch evaluation within a single benchmark. To further broaden the evaluation, we complement this benchmark with additional music-domain reference-control experiments in Appendix D.1 and D.2. For long-form audio generation, we follow prior work and evaluate on AudioCaps (Kim et al., 2019) for general sound effects and MusicCaps (Agostinelli et al., 2023) for music; both provide natural-language captions that we use as text prompts at three target durations of 10, 26, and 90 seconds. Further dataset details are provided in Appendix B.2.

Evaluation Metrics. We conduct a comprehensive evaluation covering four aspects: temporal control, reference attribute control, audio quality, and long-form generation. For temporal control, we follow prior work Guo et al. (2024); Wang et al. (2025b) and report two metrics computed by a sound event detection (SED) system Mesaros et al. (2016): event-based measures (Eb) and clip-level macro F1 score (At). For reference attribute control, we follow Zhu et al. (2026) and report the Mean Absolute Error (MAE) between the attribute curves extracted from the generated audio and the reference targets, separately for loudness and pitch. For audio quality, we employ a suite of standard metrics including Fréchet Audio Distance (FAD), Kullback–Leibler (KL) divergence, Fréchet Distance (FD), Inception Score (IS) (Liu et al., 2023), and CLAP (Wu et al., 2023). We further conduct Mean Opinion Score (MOS) tests where 20 participants rate four aspects on a five-point scale: control fidelity for timing, loudness, and pitch, and overall audio quality (OVL). For long-form audio generation, we additionally report subjective MOS scores on quality, consistency and coherence across the extended duration. Further details are provided in the Appendix B.3.

4.2 Multi-Condition Controllable Generation

Baselines. For training-based controllable TTA models, we include MC-Diffusion (Guo et al., 2024) together with its CCTA variant that conditions solely on control signals without text, AudioComposer (Wang et al., 2025b) and Tango + LControl that specify temporal control through natural language, T2A-ControlNet and T2A-Adapter (Zhu et al., 2026), and the recent ControlAudio (Jiang et al., 2025a). We further report results from a set of pretrained TTA backbones, including AudioLDM-large (Liu et al., 2023), AudioLDM 2-large (Liu et al., 2024), Tango (Ghosal et al., 2023), Stable Audio Open (Evans et al., 2025), and TangoFlux (Hung et al., 2024), which generate audio from text prompts alone. We further compare with training-free methods including TG-Diff (Du et al., 2024), FreeAu-

Table 2: Timing control results with controllable TTA on the AudioCondition test set.

Method	Eb↑	At↑
Ground Truth	43.37	67.53
CCTA	14.57	18.27
MC-Diffusion	29.07	47.11
Tango + LControl	21.46	55.15
AudioComposer-S	43.51	60.83
AudioComposer-L	44.40	63.30
ControlAudio	**55.58**	**79.52**
T2A-ControlNet	47.98	67.92
T2A-Adapter	<u>54.36</u>	68.26
TG-Diff‡	26.70	60.06
FreeAudio	44.34	68.50
TAG	47.21	74.77
AnyAudio	48.62	<u>75.12</u>

Table 3: Objective evaluation on AudioCaps and MusicCaps test sets at different output lengths. † indicates publicly reported results. Other results are evaluated using publicly available checkpoints.

Method	Sec.	AudioCaps					MusicCaps				
		FAD↓	KL↓	FD↓	IS↑	CLAP↑	FAD↓	KL↓	FD↓	IS↑	CLAP↑
AudioLDM†	10	8.37	2.00	37.27	6.51	0.200	6.34	1.45	34.44	2.49	0.220
AudioGen† / MusicGen†	10	1.72	**1.39**	**13.29**	10.22	0.270	4.55	1.43	25.40	2.24	0.180
AudioLDM 2†	10	1.97	1.49	26.34	8.46	0.220	2.80	**1.26**	15.61	2.84	0.230
Stable Audio Open†	10	3.15	2.01	29.01	10.37	0.210	3.23	1.51	36.33	2.94	0.230
AnyAudio	10	**1.48**	1.47	19.83	**14.30**	**0.536**	**1.67**	1.44	**10.14**	**3.96**	**0.392**
AudioLDM	26	5.99	1.89	35.31	7.15	0.409	3.64	1.41	20.06	2.79	**0.427**
AudioGen / MusicGen	26	3.38	2.15	19.07	7.82	0.338	5.69	1.76	25.93	2.43	0.318
AudioLDM 2	26	4.26	1.77	28.71	11.71	0.406	3.16	1.48	20.73	3.29	0.361
Stable Audio Open	26	4.17	2.30	28.82	10.17	0.281	2.94	1.47	19.92	2.99	0.410
SaFa	26	3.98	**1.36**	23.30	9.43	0.494	3.08	**1.07**	15.03	3.00	0.407
FreeAudio†	26	1.85	1.49	**16.21**	-	0.527	2.46	1.65	-	2.17	0.359
AnyAudio	26	**1.59**	1.48	20.98	**13.76**	**0.531**	**1.81**	1.58	**11.58**	**3.77**	0.409
AudioLDM	90	4.81	1.79	28.49	7.99	0.430	3.74	1.73	21.78	2.74	**0.428**
AudioGen / MusicGen	90	3.38	2.45	20.37	7.42	0.314	6.12	1.85	28.37	2.27	0.306
AudioLDM 2	90	7.75	1.98	35.94	8.54	0.356	5.09	1.66	29.90	3.34	0.327
SaFa	90	4.08	**1.45**	24.69	10.26	0.485	2.34	**1.25**	14.62	3.09	0.410
FreeAudio†	90	2.19	1.68	**19.95**	-	0.495	2.57	1.76	-	2.29	0.342
AnyAudio	90	**1.74**	1.67	22.41	**13.94**	**0.515**	**1.93**	1.56	**11.46**	**3.90**	0.391

dio (Jiang et al., 2025b), and TAG (Zhang et al., 2026). TG-Diff‡ is evaluated under a different SED model (Turpault et al., 2019) and we report its results separately.

Main Results. Tables 1 and 2 report objective and subjective results on the AudioCondition test set. AnyAudio achieves competitive or superior performance across timing, loudness, pitch, and audio quality on both objective metrics and MOS ratings, despite operating without any task-specific training. On timing control, AnyAudio surpasses all training-free baselines and matches the strongest training-based methods. On pitch and loudness control, AnyAudio achieves accurate alignment to the reference curves, demonstrating that loss-guided optimization over the decoded waveform is a viable training-free alternative to dedicated conditional fine-tuning. Compared with the AnyAudio base model in which all guidance modules are disabled, the full AnyAudio yields consistent gains on every controlled axis while preserving the backbone audio quality. Additional quantitative results in Appendix E further validate the effectiveness of combining multiple control objectives within a single inference process.

4.3 Long-Form Audio Generation

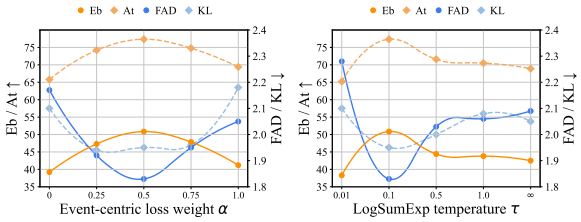
For long-form audio generation models, AudioLDM-large extends to longer durations via an outpainting-style scheme, while AudioGen-medium (Kreuk et al., 2022), MusicGen-large (Copet et al., 2023), and AudioLDM 2-large natively support long-form generation, with AudioGen and MusicGen used for general sounds and music respectively. Stable Audio Open supports up to 47-second generation. We further compare with training-free methods including FreeAudio and SaFa-DiT (Dai et al., 2025). Table 3 reports objective metrics on AudioCaps and MusicCaps at 10, 26, and 90 seconds. At 26 and 90 seconds, AnyAudio achieves the best results on most metrics across both datasets, indicating that LSD sustains generation quality as the audio duration grows. The subjective MOS in Table 4 corroborates these findings, with AnyAudio receiving the highest ratings on Quality, Consistency, and Coherence on both datasets. Additional results on more pretrained backbones are provided in Appendix D.3.

Table 4: Subjective evaluation results for AudioCaps and MusicCaps.

Method	Second	AudioCaps			MusicCaps		
		Quality↑	Consistency↑	Coherence↑	Quality↑	Consistency↑	Coherence↑
AudioGen / MusicGen	26	$2.87_{\pm 0.18}$	$3.45_{\pm 0.21}$	$3.71_{\pm 0.16}$	$3.35_{\pm 0.24}$	$3.67_{\pm 0.19}$	$3.58_{\pm 0.22}$
AudioLDM 2	26	$3.05_{\pm 0.20}$	$3.52_{\pm 0.17}$	$3.44_{\pm 0.23}$	$3.28_{\pm 0.18}$	$3.31_{\pm 0.25}$	$3.46_{\pm 0.21}$
Stable Audio Open	26	$3.42_{\pm 0.14}$	$3.58_{\pm 0.20}$	$3.55_{\pm 0.18}$	$3.74_{\pm 0.15}$	$3.62_{\pm 0.17}$	$3.87_{\pm 0.13}$
SaFa	26	$3.65_{\pm 0.12}$	$3.78_{\pm 0.16}$	$3.86_{\pm 0.11}$	$3.51_{\pm 0.22}$	$3.61_{\pm 0.19}$	$3.62_{\pm 0.18}$
AnyAudio	26	$\mathbf{3.72}_{\pm 0.10}$	$\mathbf{4.05}_{\pm 0.12}$	$\mathbf{4.12}_{\pm 0.11}$	$\mathbf{4.08}_{\pm 0.14}$	$\mathbf{3.96}_{\pm 0.15}$	$\mathbf{3.92}_{\pm 0.13}$

4.4 Ablation Studies

Ablation on TAC Strategies. We ablate the core design choices of TAC on the AudioCondition test set, with results reported in Figure 2. We first vary the balancing weight α between the event-centric and temporal-centric losses, where $\alpha = 1$ reduces TAC to event-centric only and $\alpha = 0$ to temporal-centric only. Both ends fall behind the combined objective, confirming that the two losses address complementary failure modes and are jointly necessary. We then vary the temperature τ in the LogSumExp aggregation of $\mathcal{L}^{\text{tc}}$, with $\tau \to \infty$ (denoted as Mean) corresponding to mean aggregation. Performance degrades at both small τ (close to hard max, unstable optimization) and large τ (close to mean, losing the worst-frame focus that motivates LogSumExp). We further apply TAC to alternative pretrained backbones and observe consistent timing gains, demonstrating that the attention guidance generalizes across different latent audio diffusion models. Finally, removing LLM recaption (i.e., directly using the original sound event annotations from the dataset) leads to a performance drop, as the raw annotations often mismatch the caption tokens semantically and cannot reliably anchor the attention activations. More ablation analyses of TAC are provided in Appendix C.1.



Method	Eb↑	At↑	FAD↓	KL↓	CLAP↑
SAO	3.37	45.21	3.12	2.87	0.280
TangoFlux	5.93	53.97	2.58	1.86	**0.494**
SAO + TAC	40.27	65.27	3.54	2.61	0.291
TangoFlux + TAC	38.35	61.96	2.66	1.93	0.452
AnyAudio	**48.62**	**75.12**	**1.78**	**1.80**	0.427
w/o recaption	41.45	69.21	1.86	1.97	0.392

Figure 2: Hyperparameter sensitivity and cross-backbone comparison.

Ablation on LSD Strategies. We compare LSD against latent average from MultiDiffusion (Bar-Tal et al., 2023) and latent swap from Swap Forward (Dai et al., 2025) under the same TTA backbone, and further remove the two components of LSD in turn. To better characterize cross-segment coherence, we additionally report Intra-LPIPS (I-LPIPS) and Intra-CLAP (I-CLAP) following (Dai et al., 2025), which measure the internal similarity among 10-second sliding-window crops from each generated audio. LSD outperforms both alternatives on AudioCaps and MusicCaps, and both components contribute to the final performance, with each addressing a complementary failure mode. More qualitative analysis of LSD and the choice of the low-pass cutoff ratio is provided in Appendix C.2.

Table 5: Ablation on the overlap fusion strategy at 26 seconds.

Method	AudioCaps						MusicCaps					
	FAD↓	KL↓	IS↑	CLAP↑	I-LPIPS↓	I-CLAP↑	FAD↓	KL↓	IS↑	CLAP↑	I-LPIPS↓	I-CLAP↑
MultiDiffusion	1.84	1.56	11.71	0.504	0.392	0.831	2.42	1.44	3.62	0.387	0.373	0.845
Swap Forward	1.70	1.51	13.18	0.523	0.363	0.874	1.87	1.38	3.91	0.394	0.360	0.861
LSD (full)	1.59	1.48	13.76	0.531	0.324	0.905	1.81	1.58	3.77	0.409	0.348	0.867
w/o HF align	1.82	1.57	11.90	0.506	0.326	0.907	2.05	1.46	3.58	0.382	0.350	0.870
w/o LF share	1.94	1.86	13.27	0.481	0.391	0.850	2.18	1.62	3.74	0.378	0.384	0.843

5 Conclusion

In this paper, we presented AnyAudio, a training-free framework for holistic control in TTA diffusion. AnyAudio integrates three task-specific control modules within a single inference loop: TAC steers the text-audio cross-attention to align sound events with their target time intervals, LSD fuses overlapping latent segments in the frequency domain to extend generation beyond the backbone's training duration, and RAC matches differentiable reference attributes extracted from the decoded waveform against user-provided references. The three modules operate on a frozen pretrained backbone without any task-specific retraining, and any subset of them can be activated jointly within a single inference pass. Across timing, long-form, and reference attribute control tasks, AnyAudio achieves state-of-the-art performance against both training-based and training-free baselines, while uniquely supporting the composition of all three control types under a unified inference framework.

AI USE STATEMENT

We used generative AI tools to draft and edit parts of the text for readability and to check the manuscript for internal consistency. All AI-assisted text was reviewed and revised by the authors, who take full responsibility for the content of the paper. Generative AI tools were not used to generate data, to develop the method, to design or run the experiments, or to analyze the results.

REFERENCES

Andrea Agostinelli, Timo I Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, et al. Musiclm: Generating music from text. *arXiv preprint arXiv:2301.11325*, 2023.

Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: Fusing diffusion paths for controlled image generation. 2023.

Thierry Bertin-Mahieux, Daniel PW Ellis, Brian Whitman, and Paul Lamere. The million song dataset. 2011.

Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 721–725. IEEE, 2020.

Minghao Chen, Iro Laina, and Andrea Vedaldi. Training-free layout control with cross-attention guidance. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 5343–5353, 2024.

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. *Advances in Neural Information Processing Systems*, 36:47704–47720, 2023.

Yusheng Dai, Chenxi Wang, Chang Li, Chen Wang, Jun Du, Kewei Li, Ruoyu Wang, Jiefeng Ma, Lei Sun, and Jianqing Gao. Latent swap joint diffusion for long-form audio generation. *arXiv preprint arXiv:2502.05130*, 2025.

Yusheng Dai, Zehua Chen, Yuxuan Jiang, Baolong Gao, Qiuhong Ke, Jun Zhu, and Jianfei Cai. Omni2sound: Towards unified video-text-to-audio generation. *arXiv preprint arXiv:2601.02731*, 2026.

Michaël Defferrard, Kirell Benzi, Pierre Vandergheynst, and Xavier Bresson. Fma: A dataset for music analysis. *arXiv preprint arXiv:1612.01840*, 2016.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

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

Tianjiao Du, Jun Chen, Jiasheng Lu, Qinmei Xu, Huan Liao, Yupeng Chen, and Zhiyong Wu. Controllable text-to-audio generation with training-free temporal guidance diffusion. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6. IEEE, 2024.

Janek Ebbers, Reinhold Haeb-Umbach, and Romain Serizel. Threshold independent evaluation of sound event detection scores. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1021–1025. IEEE, 2022.

Zach Evans, CJ Carr, Josiah Taylor, Scott H Hawley, and Jordi Pons. Fast timing-conditioned latent audio diffusion. In *Forty-first International Conference on Machine Learning*, 2024a.

Zach Evans, Julian D Parker, CJ Carr, Zack Zukowski, Josiah Taylor, and Jordi Pons. Long-form music generation with latent diffusion. *arXiv preprint arXiv:2404.10301*, 2024b.

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

Eduardo Fonseca, Xavier Favory, Jordi Pons, Frederic Font, and Xavier Serra. Fsd50k: an open dataset of human-labeled sound events. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:829–852, 2021.

Yao Fu, Dong-Ki Kim, Jaekyeom Kim, Sungryull Sohn, Lajanugen Logeswaran, Kyunghoon Bae, and Honglak Lee. Autoguide: Automated generation and selection of context-aware guidelines for large language model agents. *Advances in Neural Information Processing Systems*, 37:119919–119948, 2024.

Hugo Flores García, Oriol Nieto, Justin Salamon, Bryan Pardo, and Prem Seetharaman. Sketch2sound: Controllable audio generation via time-varying signals and sonic imitations. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 776–780. IEEE, 2017.

Deepanway Ghosal, Navonil Majumder, Ambuj Mehrish, and Soujanya Poria. Text-to-audio generation using instruction guided latent diffusion model. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 3590–3598, 2023.

Yuxin Guo, Teng Wang, Yuying Ge, Shijie Ma, Yixiao Ge, Wei Zou, and Ying Shan. Audiostory: Generating long-form narrative audio with large language models. *arXiv preprint arXiv:2508.20088*, 2025.

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

Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pp. 131–135. IEEE, 2017.

Shawn Hershey, Daniel PW Ellis, Eduardo Fonseca, Aren Jansen, Caroline Liu, R Channing Moore, and Manoj Plakal. The benefit of temporally-strong labels in audio event classification. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 366–370. IEEE, 2021.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

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

Rongjie Huang, Jiawei Huang, Dongchao Yang, Yi Ren, Luping Liu, Mingze Li, Zhenhui Ye, Jinglin Liu, Xiang Yin, and Zhou Zhao. Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models. In *International Conference on Machine Learning*, pp. 13916–13932. PMLR, 2023b.

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.

Sangwon Jang, Taekyung Ki, Jaehyeong Jo, Jaehong Yoon, Soo Ye Kim, Zhe Lin, and Sung Ju Hwang. Frame guidance: Training-free guidance for frame-level control in video diffusion models. *arXiv preprint arXiv:2506.07177*, 2025.

Yuxuan Jiang, Zehua Chen, Zeqian Ju, Yusheng Dai, Weibei Dou, and Jun Zhu. Controlaudio: Tackling text-guided, timing-indicated and intelligible audio generation via progressive diffusion modeling. *arXiv preprint arXiv:2510.08878*, 2025a.

Yuxuan Jiang, Zehua Chen, Zeqian Ju, Chang Li, Weibei Dou, and Jun Zhu. Freeaudio: Training-free timing planning for controllable long-form text-to-audio generation. *arXiv preprint arXiv:2507.08557*, 2025b.

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

Heeseung Kim, Sungwon Kim, and Sungroh Yoon. Guided-tts: A diffusion model for text-to-speech via classifier guidance. In *International Conference on Machine Learning*, pp. 11119–11133. PMLR, 2022.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

C. Knapp and G. Carter. The generalized correlation method for estimation of time delay. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 24(4):320–327, 1976. doi: 10.1109/TASSP.1976.1162830.

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.

Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi. Audiogen: Textually guided audio generation. *arXiv preprint arXiv:2209.15352*, 2022.

Edith Law, Kris West, Michael I Mandel, Mert Bay, and J Stephen Downie. Evaluation of algorithms using games: The case of music tagging. In *ISMIR*, pp. 387–392, 2009.

Sang-gil Lee, Zhifeng Kong, Arushi Goel, Sungwon Kim, Rafael Valle, and Bryan Catanzaro. Etta: Elucidating the design space of text-to-audio models. *arXiv preprint arXiv:2412.19351*, 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. *arXiv preprint arXiv:2301.12503*, 2023.

Haohe Liu, Yi Yuan, Xubo Liu, Xinhao Mei, Qiuqiang Kong, Qiao Tian, Yuping Wang, Wenwu Wang, Yuxuan Wang, and Mark D Plumbley. Audioldm 2: Learning holistic audio generation with self-supervised pretraining. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

Yisu Liu, Chenxing Li, Wanqian Zhang, Wenfu Wang, Meng Yu, Ruibo Fu, Zheng Lin, Weiping Wang, and Dong Yu. Degdit: Controllable audio generation with dynamic event graph guided diffusion transformer. *IEEE Transactions on Audio, Speech and Language Processing*, 2026.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *Machine Intelligence Research*, 22 (4):730–751, 2025.

Navonil Majumder, Chia-Yu Hung, Deepanway Ghosal, Wei-Ning Hsu, Rada Mihalcea, and Soujanya Poria. Tango 2: Aligning diffusion-based text-to-audio generations through direct preference optimization. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 564–572, 2024.

Ilaria Manco, Benno Weck, Seungheon Doh, Minz Won, Yixiao Zhang, Dmitry Bogdanov, Yusong Wu, Ke Chen, Philip Tovstogan, Emmanouil Benetos, et al. The song describer dataset: a corpus of audio captions for music-and-language evaluation. *arXiv preprint arXiv:2311.10057*, 2023.

Matthias Mauch and Simon Dixon. pyin: A fundamental frequency estimator using probabilistic threshold distributions. In *2014 ieee international conference on acoustics, speech and signal processing (icassp)*, pp. 659–663. IEEE, 2014.

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*, 2024.

Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen. Metrics for polyphonic sound event detection. *Applied Sciences*, 6(6):162, 2016.

Zachary Novack, Julian McAuley, Taylor Berg-Kirkpatrick, and Nicholas Bryan. Ditto-2: Distilled diffusion inference-time t-optimization for music generation. *arXiv preprint arXiv:2405.20289*, 2024a.

Zachary Novack, Julian McAuley, Taylor Berg-Kirkpatrick, and Nicholas J Bryan. Ditto: Diffusion inference-time t-optimization for music generation. *arXiv preprint arXiv:2401.12179*, 2024b.

Zachary Novack, Zack Zukowski, CJ Carr, Julian Parker, Zach Evans, Josiah Taylor, Taylor Berg-Kirkpatrick, Julian McAuley, and Jordi Pons. Low-resource guidance for controllable latent audio diffusion. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 15072–15076. IEEE, 2026.

Zakaria Patel and Kirill Serkh. Enhancing image layout control with loss-guided diffusion models. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 3916–3924. IEEE, 2025.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

Karol J Piczak. Esc: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM international conference on Multimedia*, pp. 1015–1018, 2015.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022.

Abraham Savitzky and Marcel JE Golay. Smoothing and differentiation of data by simplified least squares procedures. *Analytical chemistry*, 36(8):1627–1639, 1964.

Shira Schiber, Ofir Lindenbaum, and Idan Schwartz. Tempocontrol: Temporal attention guidance for text-to-video models. *arXiv preprint arXiv:2510.02226*, 2025.

Yifei Shen, Xinyang Jiang, Yifan Yang, Yezhen Wang, Dongqi Han, and Dongsheng Li. Understanding and improving training-free loss-based diffusion guidance. *Advances in Neural Information Processing Systems*, 37:108974–109002, 2024.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Fang-Duo Tsai, Shih-Lun Wu, Weijaw Lee, Sheng-Ping Yang, Bo-Rui Chen, Hao-Chung Cheng, and Yi-Hsuan Yang. Musecontrollite: Multifunctional music generation with lightweight conditioners. *arXiv preprint arXiv:2506.18729*, 2025.

Nicolas Turpault, Romain Serizel, Ankit Parag Shah, and Justin Salamon. Sound event detection in domestic environments with weakly labeled data and soundscape synthesis. In *Workshop on Detection and Classification of Acoustic Scenes and Events*, 2019.

Rafael Valle, Rohan Badlani, Zhifeng Kong, Sang-gil Lee, Arushi Goel, Sungwon Kim, Joao Felipe Santos, Shuqi Dai, Siddharth Gururani, Aya Aljafari, et al. Fugatto 1: Foundational generative audio transformer opus 1. In *The Thirteenth International Conference on Learning Representations*, 2025.

Apoorv Vyas, Bowen Shi, Matthew Le, Andros Tjandra, Yi-Chiao Wu, Baishan Guo, Jiemin Zhang, Xinyue Zhang, Robert Adkins, William Ngan, et al. Audiobox: Unified audio generation with natural language prompts. *arXiv preprint arXiv:2312.15821*, 2023.

Junyou Wang, Zehua Chen, Binjie Yuan, Kaiwen Zheng, Chang Li, Yuxuan Jiang, and Jun Zhu. Audiomog: Guiding audio generation with mixture-of-guidance. *arXiv preprint arXiv:2509.23727*, 2025a.

Yuanyuan Wang, Hangting Chen, Dongchao Yang, Zhiyong Wu, and Xixin Wu. Audiocomposer: Towards fine-grained audio generation with natural language descriptions. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025b.

Shih-Lun Wu, Chris Donahue, Shinji Watanabe, and Nicholas J Bryan. Music controlnet: Multiple time-varying controls for music generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:2692–2703, 2024.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Jiwen Yu, Yinhuai Wang, Chen Zhao, Bernard Ghanem, and Jian Zhang. Freedom: Training-free energy-guided conditional diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 23174–23184, 2023.

Yi Yuan, Xubo Liu, Haohe Liu, Xiyuan Kang, Zhuo Chen, Yuxuan Wang, Mark D Plumbley, and Wenwu Wang. Dreamaudio: Customized text-to-audio generation with diffusion models. *IEEE Transactions on Audio, Speech and Language Processing*, 2026.

Hanwen Zhang, Jinshen Zhang, Cong Zhang, Shuhui Wang, and Wei Yang. Tag: Structured temporal audio generation via llm-guided manual scription and control. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 17752–17756. IEEE, 2026.

Haina Zhu, Yao Xiao, Xiquan Li, Ziyang Ma, Jianwei Yu, Bowen Zhang, Mingqi Yang, and Xie Chen. Audio controlnet for fine-grained audio generation and editing. *arXiv preprint arXiv:2602.04680*, 2026.

A Model Details

A.1 Model Architecture

Our framework builds on a pretrained latent audio diffusion model that follows the Stable Audio architecture (Evans et al., 2024b; 2025), with three main components: a 1D audio autoencoder, a text encoder, and a diffusion transformer (DiT).

Audio Autoencoder. The autoencoder operates on monaural waveforms at 16kHz with a maximum input length of 160,000 samples (10 seconds). The encoder follows the DAC-VAE (Evans et al., 2024a) design with stride configuration $[4, 4, 4, 10]$, yielding a total downsampling ratio of 640 and a latent temporal resolution of 25Hz. The encoder produces a 128-dimensional intermediate representation, which is mapped through a VAE bottleneck to a 64-dimensional latent. The decoder mirrors the encoder with rate configuration $[10, 4, 4, 4]$ and a hidden dimension of 1,536.

Text Encoder. We use a frozen Flan-T5 large (Chung et al., 2024) encoder to embed the input text prompt, with a maximum sequence length of 256 tokens. The output token embeddings have a dimension of 768 and are passed to the diffusion model via cross-attention.

Diffusion Transformer. The denoising network is a 24-layer continuous DiT operating in the 64-dimensional latent space. Each block uses 24 attention heads with an embedding dimension of 1,536. Cross-attention conditions on three signals concatenated as a token sequence: the Flan-T5 text embedding (*prompt*), the segment start time (*seconds_start*), and the segment total duration (*seconds_total*), with a shared conditioning dimension of 768. The two scalar conditioning signals are additionally injected as global conditioning of dimension 1,536.

A.2 Base Model Pretraining Data

Our base TTA model is pretrained on a large-scale mixture of publicly available audio-text corpora, summarized in Table 6. The mixture combines datasets with descriptive captions, including AudioCaps (Kim et al., 2019), WavCaps (Mei et al., 2024), and Clotho v2 (Drossos et al., 2020), with corpora annotated by event labels such as AudioSet (Gemmeke et al., 2017), FSD50k (Fonseca et al., 2021), ESC-50 (Piczak, 2015), and VGG-Sound (Chen et al., 2020). We additionally include music-oriented datasets MTT (Law et al., 2009), MSD (Bertin-Mahieux et al., 2011), and FMA (Defferrard et al., 2016) to broaden coverage of musical content. This combination of detailed descriptions, sound event vocabulary, and music captions enables the model to learn robust and versatile audio-text representations.

All audio samples are resampled to 16kHz and converted to mono. Each sample is processed into a uniform 10-second clip: shorter samples are right-padded with silence, while longer samples are randomly cropped. We emphasize that this pretraining data is used only to obtain the base TTA model and is not used by any of our guidance modules.

Table 6: Audio-text source datasets used for pretraining the base TTA model.

Dataset	Hours (h)	Samples	Text Type
AudioCaps	109	44K	Caption
WavCaps	7,090	400K	Caption
Clotho v2	152	7K	Caption
AudioSet	5,800	2M	Label
FSD50K	108	51K	Label
ESC-50	2.8	2K	Label
VGG-Sound	550	210K	Label
MTT	200	24K	Caption
MSD	7,333	880K	Caption
FMA	900	11K	Caption

A.3 Training Strategy

We pretrain the base TTA model with a v-prediction (Salimans & Ho, 2022) objective on the dataset mixture described in Section A.2. The model is optimized using AdamW (Loshchilov & Hutter, 2017) with learning rate 5×10^{-5}, betas $(0.9, 0.999)$, and weight decay 10^{-3}. We adopt an inverse learning rate schedule with $\gamma_{\text{inv}} = 10^6$, power 0.5, and a warmup ratio of 0.99. An exponential moving average (EMA) of the model weights is maintained throughout training with $\beta = 0.9999$ and power $3/4$, updated every step starting from step 1. Training runs for 1M steps on 8 NVIDIA A800 GPUs with a per-device batch size of 8, yielding a total batch size of 64.

B Experimental Details

B.1 Implementation Details

Table 7: Hyperparameter settings used in our experiments.

Setting	Hyperparameter	Value
Multi-Condition Controllable Generation (AudioCondition test set)		
TAC	event-centric weight α	0.5
	LogSumExp temperature τ	0.1
	guidance step range	first 15
	η_{TAC} / repeat step K_t, steps 1–10	0.3 / 10
	η_{TAC} / repeat step K_t, steps 11–15	0.15 / 5
RAC (loudness + pitch)	guidance step range	first 75
	loudness weight λ_{loud}	1.0
	pitch weight λ_{pitch}	1.0
	η_{RAC} / repeat step K_t, steps 1–10	3 / 3
	η_{RAC} / repeat step K_t, steps 11–30	2 / 2
	η_{RAC} / repeat step K_t, steps 31–75	1 / 1
Other	LLM for recaption	Gemini 2.5 Pro
	classifier-free guidance scale	3.5
Long-Form Audio Generation (AudioCaps test set / MusicCaps)		
LSD	low-pass sharing bins	10
	full-segment sharing steps T_{freq}	50
	transition band bins	5
	segment count m at 26s / 90s	3 / 11
	overlap length (latent frames)	50
Other	classifier-free guidance scale	7.0

Table 7 summarizes the task-specific hyperparameters used in our experiments. All experiments are conducted on a single NVIDIA A800 GPU with 80 GB memory. At each guided denoising step, we apply a time-travel trick (Yu et al., 2023; Shen et al., 2024; Jang et al., 2025) that performs multiple inner-loop updates on z_t before advancing to z_{t-1}, which strengthens the effect of guidance without increasing the number of denoising steps.

Cross-Attention Extraction. The routing matrix $A^t \in \mathbb{R}^{F\times N_p}$ used in TAC is computed at each denoising step by aggregating the cross-attention activations across all L transformer layers and H heads. We take the post-softmax attention probabilities, which are normalized along the text-token axis at each layer-head pair, and average them uniformly across layers and heads to produce A^t. No smoothing is applied across denoising steps, so A^t is recomputed from scratch at each step.

Differentiable Pitch Extraction. Standard pitch evaluation protocols rely on dynamic-programming-based decoding (e.g., Viterbi smoothing in PYIN (Mauch & Dixon, 2014)), which is non-differentiable and cannot be used as a guidance objective. We instead derive a differentiable proxy by scoring candidate frequencies through harmonic-comb matching against the STFT magnitude and applying a soft-argmax over candidates, yielding a continuous f_0 estimate that is back-propagatable to

the latent. The reference f_0 contour for evaluation is still extracted with PYIN, so the guidance signal and the evaluation metric are computed under different estimators. We acknowledge this estimator mismatch as a limitation of the proxy: while empirically optimizing the proxy loss improves the PYIN-evaluated pitch MAE, the proxy is not a perfect substitute for the evaluation pipeline and may introduce a systematic gap in finer-grained pitch behavior.

B.2 Evaluation Datasets

AudioCondition Test Set. We use the publicly released test split of AudioCondition (Guo et al., 2024), which is built upon AudioSet-Strong (Hershey et al., 2021) and contains 1110 audio clips covering the event vocabulary supported by the PB-SED (Mesaros et al., 2016) evaluator. Each clip is paired with a WavCaps-style (Mei et al., 2024) scene caption describing the overall acoustic scene, frame-level timestamps for sound events drawn from a vocabulary of 456 classes, and per-frame loudness and pitch contours extracted by standard signal-processing tools. Due to the limited class coverage of the PB-SED evaluator, our timing evaluation on AudioCondition is restricted to the 10 sound classes that PB-SED supports, while the scene caption and other annotations remain unchanged. For prompt construction, we start from the original scene caption and check whether each annotated event keyword is already present. When an event keyword is missing, we insert it into the caption through an LLM-based recaption step (Figure 3) that preserves all original wording and weaves in only the missing keywords as a contiguous token span, producing an extended caption c' in which every event is grounded to a token span (Section 3.2). For clips whose original caption already covers all events, the recaption step is skipped and the caption is used as is. Reference targets for the timing, loudness, and pitch metrics are taken directly from the dataset's frame-level event timestamps and per-frame loudness and pitch contours, ensuring that all evaluations on AudioCondition are performed against a single, internally consistent set of references.

AudioCaps Test Set. We adopt the standard test split of AudioCaps (Kim et al., 2019) for evaluating long-form generation on general sounds, following prior work (Dai et al., 2025; Jiang et al., 2025b). The test split contains natural-language captions describing the audio content, which we use directly as text prompts without any additional processing. On the same batch of test captions, we generate audio at three target durations of 10, 26, and 90 seconds, so that the three duration settings share an identical prompt distribution and any difference in evaluation metrics across durations reflects the long-form behavior of each method rather than caption-level variation. Reference statistics for FAD, KL, FD, IS, and CLAP at each duration are computed against the corresponding ground-truth audio in the test split.

MusicCaps. We adopt the MusicCaps (Agostinelli et al., 2023) test set for evaluating long-form generation on music, following the same protocol as on AudioCaps. Each clip is paired with a free-form natural-language caption describing musical attributes such as genre, instrumentation, mood, and rhythmic feel, which we use directly as text prompts. On the same batch of test captions, we generate music at three target durations of 10, 26, and 90 seconds, sharing prompts across durations to isolate the effect of long-form generation from prompt-level variation. Reference statistics at each duration are computed against the corresponding ground-truth music in the test split.

B.3 Evaluation Metrics

Audio Quality Metrics. We follow standard practice and report Fréchet Audio Distance (FAD), Kullback–Leibler (KL) divergence, Fréchet Distance (FD), Inception Score (IS) (Liu et al., 2023), and the CLAP score (Wu et al., 2023), all computed against the corresponding reference audio at 16 kHz. FAD is computed in the VGGish (Hershey et al., 2017) feature space, FD and IS are computed in the PANN (Kong et al., 2020) feature space, KL is computed over PANN tag probabilities, and the CLAP score is computed as the cosine similarity between the text embedding of the prompt and the audio embedding of the generated waveform.

Long-Form Metrics. The above metrics are originally designed for short audio at the seconds scale, while our long-form evaluation operates at 26 and 90 seconds. To handle this length mismatch without retraining the evaluators, we adapt each metric using overlapping analysis windows that match the input length the corresponding evaluator was trained on. Concretely, each generated or reference clip is segmented into 10-second windows with 5-second overlap, and the per-window

You are an audio-text alignment expert.

Inputs:
- caption: a natural language description of an audio clip.
- labels: a list of event labels with onset/offset times. Labels may repeat at different times; each unique label is grounded only once.

Goal:
Produce an extended caption in which every unique event label is grounded to a contiguous phrase, and report the matched phrase for each label.

Procedure:
Case 1. If the caption already covers all labels, set the extended caption equal to the original caption, and locate the phrase in the caption that semantically corresponds to each label.
Case 2. If some labels are absent, produce a minimally extended caption by inserting only the missing events into the original caption, then locate the matched phrase for every label.

Phrase rules:
- The phrase must be copied verbatim from the (extended) caption as a contiguous span of 2-5 lowercase words.
- Prefer descriptive subject-verb chunks (e.g., "a man is talking") over generic single nouns (e.g., "speech"); fall back to a generic noun only when no more specific phrase exists.
- When multiple labels could map to the same phrase, assign each label a distinct, label-specific phrase.

Recaption rules (Case 2 only):
- Preserve every word of the original caption verbatim.
- Insert missing events with natural connectives (e.g., "while", "with", "alongside").
- Do not introduce qualifiers that alter event semantics (e.g., do not turn "breathing" into "heavy breathing").

Output a JSON object with two fields:
- extended_caption: the original or minimally extended caption.
- mapping: { label: { phrase: <matched phrase> }, ... }, keyed by unique label name.

Figure 3: Prompt used for grounding event labels to caption phrases on the AudioCondition test set.

Table 8: Examples of LLM recaption on the AudioCondition test set. Tokens **inserted** by the LLM into the original caption are highlighted in bold.

Original Caption	Event Labels	Extended Caption	Matched Phrase Span
A dog barks loudly outside.	Dog, Footsteps	A dog barks loudly outside**, with footsteps approaching from the path**.	Dog → a dog barks Footsteps → footsteps approaching
Breathing and female speech can be heard with mechanisms and a sliding door.	Female speech, Male speech	Breathing, female speech, and **male speech** can be heard with mechanisms and a sliding door.	Female speech → female speech Male speech → male speech
Water is boiling and a faucet is running.	Water tap, faucet	Water is boiling and a faucet is running.	Water tap, faucet → a faucet is running
Children speak and shout while cutlery clinks.	Child speech, Male speech, Cutlery, Female speech	Children speak and shout while cutlery clinks**, a man is speaking, and a woman is speaking**.	Child speech → children speak and shout Male speech → a man is speaking Cutlery → cutlery clinks Female speech → a woman is speaking
People are talking, laughing, shouting, and a fire crackles.	Conversation, Male speech, Female speech	People are talking, laughing, shouting, and a fire crackles**, with male speech and female speech**.	Conversation → people are talking Male speech → male speech Female speech → female speech
Dogs howl over music and television sounds.	Dog, Male speech	Dogs howl over music and television sounds**, with a man speaking**.	Dog → dogs howl Male speech → a man speaking
A telephone rings and crunching sounds occur intermittently.	Female speech, Telephone bell ringing	A telephone rings and crunching sounds occur intermittently**, while a woman speaks**.	Female speech → a woman speaks Telephone bell ringing → a telephone rings
A crowd cheers, whoops and a basketball is bounced in a game.	Male speech	A crowd cheers, whoops and a basketball is bounced in a game**, while a man is speaking**.	Male speech → a man is speaking
Telephones ring and people speak.	Speech, Male speech, Telephone bell ringing	Telephones ring and people speak**, including a man speaking**.	Speech → people speak Male speech → a man speaking Telephone bell ringing → telephones ring
A bathtub is being filled or washed.	Bathtub (filling or washing)	A bathtub is being filled or washed.	Bathtub (filling or washing) → a bathtub is being filled

outputs are aggregated into a single clip-level statistic before metric computation. For FAD, FD, and IS, the per-window embeddings are pooled into a single distribution before computing the corresponding distance or score. For KL, the per-window logits are averaged across windows and then passed through a softmax to obtain the clip-level tag distribution, following the protocol of Stable Audio (Evans et al., 2024b). For the CLAP score, the per-window audio embeddings are averaged before computing the cosine similarity with the prompt embedding. This adaptation preserves the original training distribution of each evaluator and avoids artifacts that would arise from feeding lengths beyond what the evaluators were trained on. Beyond reference-based metrics, we further evaluate cross-segment coherence within a single generated clip via Intra-LPIPS and Intra-CLAP, following (Dai et al., 2025; Jiang et al., 2025b). Both metrics measure the internal similarity among 10-second sliding-window crops from the generated audio: Intra-LPIPS aggregates the pairwise LPIPS distances between window-level mel-spectrogram representations, where lower values indicate more consistent timbral and textural characteristics across the clip; Intra-CLAP aggregates the pairwise cosine similarities between window-level CLAP audio embeddings, where higher values indicate more consistent semantic content across the clip.

Timing Control Metrics. We follow prior work (Guo et al., 2024; Wang et al., 2025b) and report two timing metrics computed by the PB-SED sound event detection system (Ebbers et al., 2022; Mesaros et al., 2016): the event-based F1 score (Eb), which measures whether each predicted event matches a reference event in both type and time interval, and the clip-level macro F1 score (At), which measures whether each event class is correctly predicted at the clip level. Both metrics are computed against the frame-level event timestamps provided by the AudioCondition test set, and are restricted to the 10 sound classes supported by PB-SED.

Attribute Control Metrics. For loudness and pitch control, we follow (Zhu et al., 2026) and report the Mean Absolute Error (MAE) between the attribute curves extracted from the generated audio and the reference targets. Loudness curves are extracted as the frame-level root-mean-square energy of the waveform on the decibel scale, followed by Savitzky-Golay smoothing (Savitzky & Golay, 1964) to suppress short-term fluctuations. Pitch curves are extracted by first estimating the fundamental frequency f_0 of the waveform, converting it to the log-frequency domain, and then applying a continuous wavelet transform (CWT) to capture multi-scale variations. Both extractors operate at the latent frame rate of the backbone, so that reference and generated curves are aligned along the temporal axis without resampling. The same extractors are used for both reference target preparation and evaluation, which eliminates extractor-induced bias between the two sides of the MAE computation.

Subjective Evaluation. We further conduct Mean Opinion Score (MOS) tests with 20 participants, each rating audio clips on a five-point scale. For multi-condition controllable generation, participants rate four aspects: control fidelity for timing, loudness, and pitch, and overall audio quality (OVL). For long-form generation, participants rate three aspects: overall quality, internal consistency across the extended duration, and coherence of transitions. Each participant is presented with the text prompt, the reference targets when applicable, and the generated audio clip, and is asked to rate each aspect independently. Audio clips are randomly shuffled across methods to avoid ordering bias.

B.4 Evaluation Results

Detailed Timing Results. Tables 9 and 10 report category-wise Eb and At scores on the 10 sound classes supported by the SED evaluator, complementing the aggregated results in Table 1. The numbers for ControlAudio (Jiang et al., 2025a) and FreeAudio (Jiang et al., 2025b) are taken directly from their original papers, while AudioComposer-S (Wang et al., 2025b) and all other methods are evaluated by ourselves using their publicly available checkpoints under our SED evaluator, ensuring identical event-class coverage and identical decoding configuration across training-free methods.

C Ablation Studies

C.1 Timing Attention Control

Per-Category Analysis. We provide additional analyses on the AudioCondition test set. The category-wise gains in Tables 9 and 10 reveal that TAC delivers its largest improvements on background-style events such as Frying, Toothbrush, Water, and Cleaner, where the temporal-centric loss explicitly suppresses the background dominance that would otherwise crowd out foreground attention; the gains are more moderate on short transient events such as Cat, Dog, and Dishes, where the per-event attention signal is inherently sparse and harder to stabilize through guidance. We also observe that the At score on the Speech category drops from 96.41 in the AnyAudio base to 86.83 in the full AnyAudio, reflecting an inherent trade-off in attention-based timing control. This drop arises from the dominant presence of speech in the AudioCondition test set: among 1110 clips, 973 (87.7%) contain speech-related labels, with Male/Female speech occurrences reaching 2381 and 1236 respectively, far exceeding any other event class. The same dominance has motivated prior work (Jiang et al., 2025a) to introduce a dedicated phoneme branch for speech-aware generation, while in our training-free setting the temporal-centric loss strongly suppresses the broad contextual presence of speech that the backbone naturally produces, exchanging clip-level recall for finer event-level localization.

Table 9: Category-wise Eb scores on the AudioCondition test set.

Method	Alarm	Blender	Cat	Dishes	Dog	Toothbrush	Frying	Water	Speech	Cleaner	Avg
AudioComposer-S	52.08	33.77	18.60	30.62	32.63	35.62	30.34	50.43	67.86	42.17	39.41
T2A-Adapter	45.19	35.16	25.00	17.86	37.89	30.48	33.48	27.40	53.75	22.22	32.84
ControlAudio	53.08	48.57	27.06	40.55	50.41	80.51	61.15	52.01	64.23	78.26	55.58
AudioLDM	0.92	10.53	1.11	3.17	1.87	25.64	19.19	13.89	2.41	13.33	9.21
AudioLDM 2	1.48	3.28	0.99	3.16	4.19	3.57	25.64	17.95	3.06	0.00	6.33
Tango	0.40	6.78	0.86	0.55	4.14	5.26	28.21	19.47	3.26	8.70	7.76
Stable Audio Open	0.72	0.00	0.00	2.01	3.14	8.70	9.26	0.00	2.20	7.69	3.37
TangoFlux	1.87	0.00	0.00	1.98	5.29	16.95	16.28	3.15	3.77	10.00	5.93
FreeAudio	40.00	47.00	14.00	20.00	23.00	78.00	60.00	51.00	39.00	72.00	44.34
AnyAudio (base)	1.62	9.68	0.00	3.53	5.78	7.14	35.29	10.34	4.22	18.75	9.64
AnyAudio	46.94	50.78	22.62	23.76	26.12	80.05	63.92	53.48	45.09	73.44	48.62

Table 10: Category-wise At scores on the AudioCondition test set.

Method	Alarm	Blender	Cat	Dishes	Dog	Toothbrush	Frying	Water	Speech	Cleaner	Avg
AudioComposer-S	70.14	50.85	28.57	42.37	60.38	52.83	47.37	64.00	98.72	60.23	57.55
T2A-Adapter	72.25	44.12	50.98	39.45	82.44	48.19	54.08	32.26	97.97	26.67	54.84
ControlAudio	87.22	69.39	41.25	67.48	85.50	91.67	91.38	82.69	88.65	90.00	79.52
AudioLDM	52.98	51.43	39.34	44.03	61.63	46.75	41.81	50.00	74.76	30.77	49.35
AudioLDM 2	76.92	13.64	42.50	43.69	66.29	31.58	57.49	40.58	90.70	13.33	47.67
Tango	56.80	9.52	33.33	23.90	68.79	23.53	66.15	65.98	97.20	10.00	45.52
Stable Audio Open	68.37	16.00	25.81	59.26	69.80	46.15	18.39	25.45	70.72	52.17	45.21
TangoFlux	61.76	0.00	54.84	36.11	71.59	66.67	53.33	53.57	97.35	44.44	53.97
FreeAudio	73.00	67.00	38.00	42.00	59.00	88.00	85.00	81.00	76.00	76.00	68.50
AnyAudio (base)	74.71	50.00	55.17	48.57	71.17	72.00	75.34	65.42	96.41	57.14	66.59
AnyAudio	80.29	67.54	53.12	55.46	84.76	84.71	79.41	76.45	86.83	82.63	75.12

Generalization Across Backbones. The cross-backbone results in Figure 2 further demonstrate that TAC generalizes beyond the specific architecture of our base model. While AnyAudio and Stable Audio Open both adopt a DiT backbone with cross-attention between text tokens and audio latent tokens, TangoFlux (Hung et al., 2024) is built on an MM-DiT architecture combining double-stream and single-stream blocks. We adapt TAC to TangoFlux by extracting the text-to-audio cross-attention from the double-stream blocks as the routing matrix A^t, and applying the same event-centric and temporal-centric losses on top of it. The consistent timing gains observed across both DiT and MM-DiT backbones indicate that TAC is agnostic to the specific architectural realization of text-audio interaction, as long as the routing can be read out from the attention activations.

C.2 LATENT SPECTRAL DECOUPLING

Choice of Low-Frequency Sharing Bandwidth. LSD shares a low-frequency band across all segments to maintain global consistency, with the bandwidth controlled by the low-pass sharing bins. We sweep this bandwidth and report the results on the AudioCaps test set at 26 seconds in Table 11. As the bandwidth grows, the cross-segment coherence metrics (I-LPIPS and I-CLAP) improve monotonically, while the reference-based quality metrics (FAD, FD, CLAP) peak around 10 bins and degrade beyond it. A small bandwidth leaves the low-frequency baseline insufficiently coupled across segments and admits stylistic drift, whereas an excessively large bandwidth imposes too much uniformity on the shared spectrum and suppresses each segment's freedom to develop its own content. The chosen value of 10 bins lies at the elbow of this trade-off, attaining strong coherence without sacrificing quality.

Table 11: Effect of the low-pass sharing bandwidth on cross-segment coherence and generation quality. Results are reported on the AudioCaps test set at 26 seconds.

Sharing bins	FAD↓	KL↓	FD↓	IS↑	CLAP↑	I-LPIPS↓	I-CLAP↑
0	1.94	1.86	22.54	13.27	0.481	0.391	0.850
5	1.72	1.62	21.65	13.55	0.510	0.355	0.880
10	1.59	1.48	20.98	13.76	0.531	0.324	0.905
15	1.81	1.55	23.34	13.20	0.522	0.318	0.911
20	2.17	1.58	23.85	13.19	0.510	0.310	0.918
30	2.84	1.60	21.78	13.22	0.493	0.302	0.924

Choice of Intra-Overlap Frequency Cutoff. Within each overlap region, LSD averages the bins below an intra-overlap frequency cutoff and phase-aligns the bins above it. We sweep this cutoff and report the results on the AudioCaps test set at 26 seconds in Table 12. A small cutoff sends most overlap bins through phase alignment, preserving fine details but leaving the low-frequency baseline insufficiently anchored within the overlap. A large cutoff sends most bins through averaging, extending the baseline into mid-frequencies where small misalignments between adjacent segments trigger phase cancellation and erode transient details. The chosen value of 5 bins anchors only the lowest bins through averaging and leaves the remaining high-frequency content to phase alignment, which generally performs best compared with others.

Table 12: Effect of the intra-overlap frequency cutoff. Results are reported on the AudioCaps test set at 26 seconds.

Overlap cutoff bins	FAD↓	KL↓	FD↓	IS↑	CLAP↑	I-LPIPS↓	I-CLAP↑
0	1.74	1.52	22.45	13.50	0.524	0.336	0.892
5	1.59	1.48	20.98	13.76	0.531	0.324	0.905
10	1.65	1.54	21.32	13.64	0.526	0.325	0.903
15	1.71	1.55	21.85	13.20	0.518	0.328	0.901
20	1.78	1.53	22.43	12.45	0.512	0.327	0.904
26	1.82	1.57	22.84	11.90	0.506	0.326	0.907

Choice of Full-Segment Sharing Schedule. LSD enforces low-frequency sharing across all segments during the first T_{freq} denoising steps, after which the sharing is restricted to overlapping regions only. This schedule controls how long the global structure is anchored before each segment is released to develop its own content. We sweep T_{freq} over $\{0, 25, 50, 75, 100\}$ on the AudioCaps test set at 26 seconds and report the results in Table 13. A small T_{freq} leaves the global structure insufficiently established before sharing is released, allowing stylistic drift to develop in subsequent steps; a large T_{freq} extends the cross-segment averaging into the late denoising stage, where the fine-grained content of each segment is being formed, and over-constrains the segments toward a homogenized output. The chosen value of $T_{\text{freq}} = 50$ aligns with the natural division between coarse structure formation and fine detail refinement in diffusion sampling, achieving the best reference-based quality (FAD, FD, CLAP) while preserving competitive cross-segment coherence. Note that I-LPIPS and I-CLAP improve monotonically with larger T_{freq} as expected, since prolonged sharing pushes segments toward identical low-frequency content; however, this improvement comes at the cost of audio fidelity beyond $T_{\text{freq}} = 50$.

Table 13: Effect of the full-segment sharing schedule T_{freq}. Results are reported on the AudioCaps test set at 26 seconds.

T_{freq}	FAD↓	KL↓	FD↓	IS↑	CLAP↑	I-LPIPS↓	I-CLAP↑
0	1.94	1.86	22.54	13.27	0.481	0.391	0.850
25	1.78	1.60	22.32	13.42	0.515	0.342	0.871
50	1.59	1.48	20.98	13.76	0.531	0.324	0.905
75	2.20	1.56	21.27	13.26	0.523	0.318	0.911
100	2.80	1.55	22.74	12.54	0.510	0.303	0.926

D Additional Comparison

D.1 Comparison with DITTO

DITTO (Novack et al., 2024b) is closely related to RAC, as both enable training-free control through differentiable objectives at inference time. However, DITTO primarily targets controllable music within an approximately 6-second native window, whereas AnyAudio is a general TTA framework for both sound effects and music, jointly supporting event-level timing, long-form generation, and time-varying acoustic control such as loudness and pitch. Technically, DITTO optimizes the initial noise through the full sampling trajectory, while RAC directly updates the current latent at selected denoising steps, enabling its integration with TAC and LSD in a unified sampling process. Below, we compare their overlapping capabilities on music editing and time-varying intensity control.

Music editing. We further compare AnyAudio with DITTO (Novack et al., 2024b) on music editing, including outpainting, inpainting, and looping. Following DITTO, we perform editing using temporally masked waveform-space RAC without task-specific training, adopt its caption template and CLAP-music FAD evaluation protocol, and match its absolute overlap/gap settings. Since the checkpoint and experimental implementation used in DITTO are not publicly available, the DITTO and MultiDiffusion (Bar-Tal et al., 2023) (MD) results associated with DITTO are paper-reported, whereas all AnyAudio and MD results on our backbone are evaluated by us. For our 10-second backbone, we additionally evaluate a 5-second overlap/5-second generation setting. For looping, the first and last 1 second of the generated audio are matched. Because DITTO does not specify the source of editing references, we use held-out MusicCaps (Agostinelli et al., 2023) and Song Describer Dataset (Manco et al., 2023) clips that are not included in backbone pretraining.

Table 14: Comparison with DITTO on music outpainting and looping. For the paper-reported DITTO setting, the native generation window is approximately 6 seconds, such that overlaps of 1, 2, and 3 seconds correspond to 5, 4, and 3 seconds of newly generated audio, respectively. For our 10-second backbone, overlaps of 1, 2, 3, and 5 seconds correspond to 9, 8, 7, and 5 seconds of newly generated audio. Lower FAD is better.

Dataset / Setting	Method	o=1	o=2	o=3	o=5	Looping
DITTO (paper-reported)	MD (DITTO)	0.733	0.716	0.710	–	0.749
	DITTO	**0.716**	**0.703**	**0.698**	–	**0.746**
MusicCaps	MD (Ours)	0.452	0.423	0.401	0.371	0.767
	AnyAudio	**0.439**	**0.411**	**0.368**	**0.350**	**0.502**
Song Describer Dataset	MD (Ours)	0.804	0.742	0.713	0.692	0.879
	AnyAudio	**0.680**	**0.647**	**0.642**	**0.635**	**0.721**

For outpainting, AnyAudio consistently improves over MD across all tested overlap lengths, including the challenging settings with long newly generated regions, suggesting that waveform-space RAC can effectively preserve and exploit the observed context during music continuation. The looping results show a similar trend, indicating that RAC is also effective when enforcing consistency between distant temporal boundaries.

AnyAudio also consistently improves over MD for inpainting. Importantly, the advantage remains substantial for long missing intervals, e.g., on MusicCaps the FAD decreases from 0.767 to 0.502 for an 8-second gap. Together with the outpainting and looping results, these experiments demonstrate

Table 15: Comparison with DITTO on music inpainting. For the paper-reported DITTO setting, gaps of 2, 3, and 4 seconds are conditioned on 4, 3, and 2 seconds of observed audio, respectively. For our 10-second backbone, gaps of 2, 3, 4, 6, and 8 seconds are conditioned on 8, 7, 6, 4, and 2 seconds of observed audio, respectively. Lower FAD is better.

Dataset / Setting	Method	g=2s	g=3s	g=4s	g=6s	g=8s
DITTO (paper-reported)	MD (DITTO)	0.690	0.694	0.701	–	–
	DITTO	**0.686**	**0.688**	**0.690**	–	–
MusicCaps	MD (Ours)	0.390	0.472	0.628	0.764	0.767
	AnyAudio	**0.339**	**0.371**	**0.437**	**0.469**	**0.502**
Song Describer Dataset	MD (Ours)	0.648	0.675	0.793	0.851	0.879
	AnyAudio	**0.610**	**0.616**	**0.665**	**0.704**	**0.721**

that RAC extends beyond the speech-heavy examples in our original evaluation and provides effective waveform-reference control for music editing.

Intensity control. We evaluate RAC on time-varying musical intensity control. DITTO follows the Music ControlNet (Wu et al., 2024) protocol but does not fully specify the construction of its control set. We therefore instantiate the public protocol using 24 control curves obtained from three curve shapes, two orientations, and four amplitudes, and combine them with 10 genres and 10 moods, resulting in approximately 2.4K control examples. We evaluate both these constructed controls and held-out intensity trajectories from the Song Describer Dataset using MSE, FAD, and CLAP.

Table 16: Comparison with DITTO on time-varying music intensity control. The DITTO rows are paper-reported. We additionally evaluate AnyAudio on held-out Song Describer Dataset trajectories and on a constructed control set following the public Music ControlNet protocol. Lower MSE and FAD are better, while higher CLAP is better.

Dataset / Setting	Method	MSE↓	FAD↓	CLAP↑
DITTO (paper-reported)	Default TTM (DITTO)	40.843	0.707	0.373
	DITTO	**4.758**	**0.682**	**0.433**
Song Describer Dataset	Default TTA (Ours)	57.051	0.367	0.367
	AnyAudio	**8.202**	**0.361**	**0.375**
Created Controls	Default TTA (Ours)	63.044	**0.336**	0.323
	AnyAudio	**5.880**	0.348	**0.353**

RAC substantially reduces intensity-tracking error on both real and constructed control trajectories. On the Song Describer Dataset, the MSE decreases from 57.051 to 8.202 while FAD and CLAP remain comparable. On the constructed control set, RAC reduces the MSE from 63.044 to 5.880 while improving CLAP from 0.323 to 0.353, with only a small increase in FAD. These results show that waveform-space RAC can accurately follow time-varying musical dynamics while largely preserving generation quality.

D.2 Comparison with MusicGen and MuseControlLite

Long-form dynamics control. We further compare AnyAudio with the training-based MuseControlLite on long-form music continuation with dynamics control. The paper-reported MusicGen (Copet et al., 2023) and MuseControlLite (Tsai et al., 2025) results are evaluated on 50 Song Describer Dataset clips with 47-second outputs, where the first 24 seconds serve as reference and the final 23 seconds are used for evaluation. We evaluate AnyAudio on 50 randomly sampled test clips following the same evaluation protocol. Since the exact clip identities used by MuseControlLite are unavailable, this comparison is protocol-matched rather than sample-matched. The Base and RAC variants operate on 10-second audio, whereas MD, LSD, and LSD+RAC evaluate 23-second continuations. We use `stable-audio-metrics`[1] together with the dynamics-correlation and smoothness metrics defined by MuseControlLite.

[1] `https://github.com/Stability-AI/stable-audio-metrics`

Table 17: Comparison with MuseControlLite on music continuation and dynamics control. MusicGen and MuseControlLite results are paper-reported, while all AnyAudio results are evaluated by us. $\mathrm{FD}_{\mathrm{OpenL3}}$ and $\mathrm{KL}_{\mathrm{PaSST}}$ measure distributional quality, while dynamics correlation and smoothness evaluate the fidelity of time-varying dynamics control.

Method	$\mathrm{FD}_{\mathrm{OpenL3}}\downarrow$	$\mathrm{KL}_{\mathrm{PaSST}}\downarrow$	CLAP↑	Dyn. Corr.↑	Smoothness↑
MusicGen	207.48	0.380	0.320	0.130	-0.180
MuseControlLite (Text+Audio)	184.55	0.370	0.340	0.140	-0.160
MuseControlLite (Text+Audio+Dynamics)	189.15	0.320	0.350	0.770	0.100
AnyAudio (Base, 10s)	206.38	0.639	0.361	0.021	-0.134
AnyAudio (RAC, 10s)	205.25	0.520	0.352	0.812	0.092
AnyAudio (MD, 23s)	268.67	0.790	0.311	0.090	-0.112
AnyAudio (LSD, 23s)	214.59	0.612	0.355	0.082	-0.108
AnyAudio (LSD+RAC, 23s)	203.05	0.583	0.356	0.808	0.087

RAC provides strong time-varying dynamics control in the music domain. For 10-second generation, RAC increases dynamics correlation from 0.021 to 0.812 while maintaining comparable distributional and semantic quality. More importantly, combining RAC with LSD enables joint long-form continuation and dynamics control: LSD+RAC achieves a dynamics correlation of 0.808 and a smoothness score of 0.087 on 23-second continuations. These values are comparable to the training-based MuseControlLite with explicit dynamics conditioning (0.770 and 0.100, respectively), although MuseControlLite remains stronger on some distributional metrics. This comparison demonstrates that AnyAudio can jointly support long-form music generation and fine-grained dynamics control without task-specific training.

D.3 Comparison on More Backbones

Same-backbone comparison with FreeAudio. The comparisons in main paper are primarily end-to-end system comparisons, where different methods may rely on different pretrained backbones. Therefore, the absolute performance gap cannot be attributed solely to the proposed control modules. To further disentangle the contribution of our long-form generation strategy from backbone quality, we conduct a controlled same-backbone comparison with FreeAudio (Jiang et al., 2025b). Since FreeAudio does not provide official code, we reimplement its long-form components, including Contextual Latent Composition and Reference Guidance, on the same backbone used by AnyAudio and evaluate under the protocol of Table 5. All model weights, sampling configurations, prompts, and evaluation settings are kept identical, with only the long-form composition strategy changed.

Table 18: Same-backbone comparison with FreeAudio on 26-second long-form generation. We reimplement the long-form components of FreeAudio on the same frozen backbone as AnyAudio and compare them with our Latent Spectral Decoupling (LSD) under identical settings. Results are reported on AudioCaps and MusicCaps.

Dataset	Method	FAD↓	KL↓	IS↑	CLAP↑	I-LPIPS↓	I-CLAP↑
AudioCaps	FreeAudio	1.81	1.54	12.16	0.510	0.374	0.862
	LSD	**1.59**	**1.48**	**13.76**	**0.531**	**0.324**	**0.905**
MusicCaps	FreeAudio	2.32	1.65	3.40	0.371	0.365	0.851
	LSD	**1.81**	**1.58**	**3.77**	**0.409**	**0.348**	**0.867**

Under the same backbone and evaluation protocol, LSD consistently improves over the reimplemented FreeAudio baseline across all reported metrics on both datasets. On AudioCaps (Kim et al., 2019), LSD reduces FAD from 1.81 to 1.59 and KL from 1.54 to 1.48, while improving IS from 12.16 to 13.76 and CLAP from 0.510 to 0.531. The intra-clip consistency metrics also improve substantially, with I-LPIPS decreasing from 0.374 to 0.324 and I-CLAP increasing from 0.862 to 0.905. A similar trend is observed on MusicCaps (Agostinelli et al., 2023), where LSD reduces FAD from 2.32 to 1.81 and improves CLAP from 0.371 to 0.409.

These controlled results complement the end-to-end comparisons in the main paper and show that the long-form generation improvements of LSD do not merely arise from differences in backbone quality. Instead, when evaluated on an identical frozen generator, LSD provides consistently better generation quality and intra-clip coherence than the reimplemented FreeAudio long-form strategy.

Public-backbone comparison. To further disentangle the contribution of LSD from the quality of our self-pretrained backbone, we evaluate LSD on two publicly available pretrained models, TangoFlux (Hung et al., 2024) and Stable Audio Open (SAO) (Evans et al., 2025), using the same prompt sets and evaluation protocol. For both backbones, we report 10- and 26-second single-window generation results, and further compare different strategies for extending generation to 90 seconds, including an AudioLDM (Liu et al., 2023)-style sequential outpainting baseline, MultiDiffusion (Bar-Tal et al., 2023) (MD), and our LSD.

For sequential outpainting, approximately half of each native generation window is retained as context while the remaining part is generated repeatedly until reaching the target duration. For the controlled comparison between MD and LSD, all settings are kept identical except for the overlap-fusion strategy. Specifically, SAO uses two 47-second windows with a 4-second overlap, while TangoFlux uses four 30-second windows with a 2-second overlap, producing 114 seconds of audio from which the first 90 seconds are evaluated. Both backbones are evaluated on AudioCaps, while SAO is additionally evaluated on MusicCaps since the public TangoFlux checkpoint was not trained with a dedicated text-to-music corpus.

Table 19: Long-form generation results on AudioCaps using public pretrained backbones. For 90-second generation, we compare sequential outpainting, MD, and our LSD under the same backbone. The best result among the 90-second methods for each backbone is highlighted in bold.

Backbone	Method	Sec.	FAD↓	KL↓	FD↓	IS↑	CLAP↑
TangoFlux	Single-window	10	2.43	1.24	19.69	12.38	0.483
TangoFlux	Single-window	26	2.54	1.48	21.41	10.39	0.477
TangoFlux	Seq. Outpainting	90	3.25	1.47	21.52	10.06	0.452
TangoFlux	MD	90	3.14	1.39	20.95	10.50	0.469
TangoFlux	LSD	90	**2.98**	**1.35**	**20.37**	**11.02**	**0.480**
SAO	Single-window	10	3.15	2.01	29.01	10.37	0.210
SAO	Single-window	26	4.17	2.30	28.82	10.17	0.281
SAO	Seq. Outpainting	90	4.37	2.53	47.55	10.16	0.252
SAO	MD	90	4.28	**2.42**	27.45	10.09	0.260
SAO	LSD	90	**4.25**	2.44	**25.71**	**10.18**	**0.263**

As shown in Table 19, LSD consistently improves over sequential outpainting on both public backbones. On TangoFlux, LSD further outperforms MD across all reported metrics at 90 seconds, reducing FAD from 3.14 to 2.98 and FD from 20.95 to 20.37, while improving IS from 10.50 to 11.02 and CLAP from 0.469 to 0.480. On SAO, LSD improves FAD, FD, IS, and CLAP over MD, with only a marginally higher KL divergence. These results demonstrate that the effectiveness of LSD is not specific to our self-pretrained backbone and transfers to different publicly available audio generators.

Table 20: Long-form generation results on MusicCaps using Stable Audio Open (SAO). The best result among the 90-second methods is highlighted in bold.

Method	Sec.	FAD↓	KL↓	FD↓	IS↑	CLAP↑
SAO	10	3.23	1.51	36.33	2.94	0.230
SAO	26	2.94	1.47	19.92	2.99	0.410
SAO (Seq. Outpainting)	90	3.35	1.74	24.53	2.82	0.368
SAO (MD)	90	**3.21**	1.64	24.48	2.85	**0.374**
SAO (LSD)	90	3.24	**1.52**	**22.10**	**2.87**	0.371

On MusicCaps, LSD achieves the best KL, FD, and IS among the 90-second methods, while MD obtains slightly better FAD and CLAP. Together with the AudioCaps results, these experiments show that LSD generalizes across both backbone architectures and audio domains, and that its long-form generation gains are not solely attributable to the quality of our self-pretrained generator.

E Composability Analysis of Joint Control

A central property of AnyAudio is that different control objectives can be activated jointly within a single sampling process. To quantify the interaction among different control dimensions, we conduct a single- versus joint-control decomposition at both the native 10-second duration and the extended

26-second duration. We define composability as the ability to satisfy multiple control objectives simultaneously with bounded interference, rather than requiring the joint configuration to match the optimum of every single-objective configuration.

10-Second Single- and Joint-Control Decomposition. We first analyze cross-control interference at the native 10-second generation duration. Timing control, loudness control, and pitch control are individually activated and progressively composed within the same denoising trajectory. All timing, attribute, and generation-quality metrics are computed from the same generated samples. This decomposition isolates the effect of each control objective and reveals how one control dimension affects the others when multiple guidance objectives are jointly optimized.

Table 21: Single- versus joint-control decomposition at 10 seconds. All control objectives operate within the same sampling process. The comparison evaluates cross-control interference among timing, loudness, and pitch control.

Method	Eb↑	At↑	$\mathrm{MAE}_{\mathrm{loudness}}$ ↓	$\mathrm{MAE}_{\mathrm{pitch}}$ ↓	FAD↓	KL↓	CLAP↑
Text	9.64	66.59	18.00	320.19	1.90	1.84	0.430
Text + Timing	50.14	76.23	16.29	297.20	1.95	1.92	0.411
Text + Loudness	10.38	65.15	3.80	280.09	1.87	1.78	0.416
Text + Pitch	11.37	63.30	14.02	115.71	1.76	1.76	0.405
Text + Loudness + Pitch	8.34	68.27	5.06	133.49	1.81	1.74	0.416
Text + Timing + Loudness + Pitch	48.62	75.12	5.22	137.98	1.78	1.80	0.427

As shown in Table 21, each individual control substantially improves its corresponding target metric. When loudness and pitch are optimized together, a moderate interaction appears between the two attribute objectives: the loudness MAE increases from 3.80 to 5.06 relative to loudness-only control, while the pitch MAE increases from 115.71 to 133.49 relative to pitch-only control. Adding timing control introduces only a small additional trade-off, resulting in loudness and pitch MAEs of 5.22 and 137.98, respectively. Meanwhile, timing remains strongly controlled, with Eb and At reaching 48.62 and 75.12 compared with 50.14 and 76.23 under timing-only control. Generation quality also remains comparable across the single- and joint-control configurations. These results indicate that the three control objectives can be jointly optimized with bounded cross-control interference.

26-Second Joint-Control Evaluation Set. The 10-second analysis evaluates the interaction among timing and waveform-level attributes, but does not involve long-form generation. To further evaluate full composability among TAC, LSD, and RAC, we construct a dedicated 26-second joint-control evaluation set containing 100 clips. The evaluation set consists of two complementary subsets: 50 real long-form clips and 50 simulated soundscapes.

For the *real subset*, we select 50 long-form recordings from the Clotho v2 evaluation set and the FSD50K evaluation set. Each selected recording is at least 26 seconds long and contains at least two event classes supported by PB-SED. We extract a 26-second segment from each recording and manually annotate the onset and offset of every target event. The captions of the real subset are manually verified to ensure that all annotated target events are explicitly described.

For the *simulated subset*, we construct 50 controlled soundscapes using long-form recordings from Clotho v2 (Drossos et al., 2020) evaluation set and FSD50K (Fonseca et al., 2021) evaluation set as environmental backgrounds. Following the class distribution of the AudioCondition (Guo et al., 2024) test set, foreground sounds with a single target-class label are sampled from VGG-Sound (Chen et al., 2020) test split, held-out ESC-50 (Piczak, 2015) folds, FSD50K evaluation split, and mixed with the background recordings. None of these evaluation partitions is used for backbone pretraining.

Since a target sound may occupy only a portion of its original foreground recording, we manually segment and annotate the actual audible interval of each target event before mixing. We additionally apply simple SNR balancing during mixing. The captions of the simulated subset are constructed from the corresponding background and foreground labels.

Together, the two subsets contain 100 clips of 26 seconds each with precise event-level onset and offset annotations. The evaluation set covers overlapping events, events spanning generation-window overlap regions, and events with diverse temporal durations, while covering all ten event classes supported by PB-SED (Mesaros et al., 2016). For each clip, loudness and pitch reference trajectories

are extracted directly from the same 26-second waveform using the same waveform-based evaluation protocol as in the main experiments.

Evaluation Protocol. PB-SED operates on 10-second inputs, whereas each joint-control example is 26 seconds long. We therefore evaluate four overlapping temporal windows, $[0, 10]$, $[5, 15]$, $[10, 20]$, and $[16, 26]$ seconds. Frame-level class scores from each window are mapped back to the global 26-second timeline. For time–class locations covered by multiple windows, the corresponding scores are averaged. We then apply the same smoothing, thresholding, and event-decoding procedure to compute Eb and At over the complete 26-second sequence. Loudness and pitch are evaluated using the same waveform-based protocols as in the 10-second experiments. Unless explicitly marked as MD, all 26-second controlled configurations use LSD for long-form generation.

Real Long-Form Audio. Table 22 reports the decomposition on the 50 real 26-second clips. In addition to the full TAC+LSD+RAC configuration, we evaluate the corresponding long-form baseline, timing-only, attribute-only, and two-attribute configurations under matched conditions.

Table 22: Single- versus joint-control decomposition on the 26-second real joint-control set (50 clips). All controlled configurations use LSD for long-form generation unless explicitly marked as MD.

Method	Eb↑	At↑	$\mathrm{MAE}_{\mathrm{loudness}}$ ↓	$\mathrm{MAE}_{\mathrm{pitch}}$ ↓	FAD↓	KL↓	CLAP↑
Ground Truth	42.20	57.76	–	–	–	–	0.437
Text (MD)	9.48	50.14	14.19	285.32	2.90	2.52	0.383
Text (LSD)	10.40	50.87	14.15	291.08	2.72	2.23	0.392
Text + Timing	43.25	64.37	15.27	273.26	2.66	2.29	0.374
Text + Loudness	10.37	51.56	2.54	263.22	2.65	2.21	0.375
Text + Pitch	9.31	49.79	13.04	134.16	2.68	2.31	0.363
Text + Loudness + Pitch	9.83	50.47	2.63	149.25	2.74	2.29	0.401
Text + Timing + Loudness + Pitch	42.92	65.88	2.61	150.89	2.84	2.30	0.414

On the real subset, full joint control preserves the effectiveness of the individual control dimensions. Compared with timing-only control, the full configuration obtains a nearly identical Eb score (43.25 vs. 42.92), while At slightly improves from 64.37 to 65.88. Loudness control is also largely preserved, with MAE changing only from 2.54 under loudness-only control to 2.61 under full joint control. The most visible degradation occurs for pitch, where MAE increases from 134.16 under pitch-only control to 150.89 under full joint control. Importantly, the loudness+pitch configuration already yields a pitch MAE of 149.25, indicating that most of this degradation arises from the interaction between the two waveform-level attributes rather than from additionally introducing timing control. The generation-quality metrics remain comparable to the LSD-only setting, indicating that simultaneous control does not cause a substantial quality collapse.

Simulated Long-Form Audio. We further evaluate the same configurations on the 50 simulated 26-second soundscapes. The simulated subset provides complementary controlled cases with explicit event placement, overlapping events, and diverse event durations.

Table 23: Single- versus joint-control decomposition on 26-second simulated joint-control set (50 clips). All controlled configurations use LSD for long-form generation unless explicitly marked as MD.

Method	Eb↑	At↑	$\mathrm{MAE}_{\mathrm{loudness}}$ ↓	$\mathrm{MAE}_{\mathrm{pitch}}$ ↓	FAD↓	KL↓	CLAP↑
Ground Truth	45.61	65.39	–	–	–	–	0.358
Text (MD)	7.42	42.55	11.89	290.14	3.86	2.79	0.340
Text (LSD)	8.91	51.22	10.94	287.62	3.54	2.48	0.356
Text + Timing	52.51	71.30	12.01	272.96	3.63	2.52	0.348
Text + Loudness	6.15	42.80	1.50	283.40	3.68	2.58	0.356
Text + Pitch	7.16	44.89	10.87	141.56	3.51	2.40	0.354
Text + Loudness + Pitch	9.32	50.43	1.86	153.20	3.65	2.46	0.350
Text + Timing + Loudness + Pitch	51.58	70.18	1.90	155.51	3.57	2.44	0.377

The simulated subset exhibits the same overall behavior. Timing remains largely preserved when all controls are activated: Eb changes from 52.51 under timing-only control to 51.58 under full joint control, and At changes from 71.30 to 70.18. Loudness MAE increases only moderately from 1.50 to

1.90. For pitch, the MAE increases from 141.56 under pitch-only control to 155.51 under full joint control, while the loudness+pitch configuration already reaches 153.20. This again shows that the dominant interaction occurs between loudness and pitch, with only a small additional effect from introducing timing control. The FAD, KL, and CLAP scores remain comparable to the LSD-only configuration.

Discussion. Across both real and simulated subsets, the full TAC+LSD+RAC configuration retains timing accuracy close to the timing-only configuration, loudness and pitch fidelity close to their corresponding attribute-control configurations, and generation quality comparable to long-form LSD generation. The consistent behavior across the two subsets provides quantitative evidence that timing control, long-form generation, and waveform-level reference control can operate simultaneously within a single frozen generator. The main observed trade-off is the interaction between loudness and pitch when both waveform-level objectives are optimized together. We therefore characterize AnyAudio's composability as simultaneous multi-dimensional control with bounded interference, rather than uniform preservation of every single-objective optimum.

F Efficiency Analysis

We report latency, RTF, and peak GPU memory on one 80GB A800 in Tab. 24. DITTO (Novack et al., 2024b)'s paper-reported 40GB A100 result is included only as context; its multi-objective runtime and memory are unreported, and its RTF is derived from its approximately 6-second output.

The additional latency and memory of TAC and RAC mainly arise from their gradient-based inner updates. When jointly activated, TAC and RAC share a single denoising process rather than requiring two independent generations. Under the same window configuration, LSD introduces no additional denoiser evaluations relative to MultiDiffusion (Bar-Tal et al., 2023). The results show that LSD scales nearly linearly in runtime with output duration while requiring only moderate additional memory. In this work, we focus on unlocking fine-grained controllability of a pre-trained TTA model. With highquality controllable generation results and the corresponding guided sampling trajectory, we are able to leverage distillation techniques to compress the number of inference steps.

Table 24: Inference efficiency comparison across different methods.

Method	Duration (s)	Latency (s)↓	RTF↓	Peak Memory (GiB)↓
DITTO	6	82.192	13.8	-
AnyAudio (base)	10	3.820	0.38	5.71
TAC	10	40.723	4.07	9.07
RAC	10	40.181	4.02	9.75
TAC+RAC	10	67.207	6.72	12.01
LSD	26	5.164	0.20	6.16
LSD	90	17.003	0.19	8.02

G Related Work

G.1 Text-to-Audio Generation

Text-to-audio (TTA) generation has rapidly progressed from early latent diffusion systems such as AudioLDM (Liu et al., 2023) and AudioLDM 2 (Liu et al., 2024) to stronger instruction-conditioned and preference-aligned models including Tango (Ghosal et al., 2023), Tango 2 (Majumder et al., 2024), and Make-An-Audio series (Huang et al., 2023b;a). Recently, more modern and scalable backbones such as Stable Audio Open (Evans et al., 2025), Audiobox (Vyas et al., 2023), TangoFlux (Hung et al., 2024), ETTA (Lee et al., 2024), and Fugatto 1 (Valle et al., 2025) have substantially improved fidelity, duration, generation speed, and instruction-following ability. Despite these advances, most TTA models still expose only coarse prompt-level control, making it difficult to precisely regulate when events occur, how long structure unfolds, or which low-level reference attributes should be followed.

G.2 Controllable Text-to-Audio Generation

Controllable TTA generation augments standard TTA with extra conditions such as event timing, reference attributes, and structural constraints. Existing methods often inject such controls through task-specific training, including MC-Diffusion and its CCTA variant (Guo et al., 2024), AudioComposer (Wang et al., 2025b), ControlAudio (Jiang et al., 2025a), and Tango with LControl (Ghosal et al., 2023). Recent works further adapt vision-style conditional modules to model acoustic or multi-type controls, such as T2A-ControlNet and T2A-Adapter (Zhu et al., 2026). In contrast, training-free methods such as TG-Diff (Du et al., 2024), TAG (Zhang et al., 2026), and FreeAudio (Jiang et al., 2025b) are designed for zero-shot and lightweight inference-time control.

Related ideas have also been explored in image and video generation, where training-free attention guidance enables structural control at inference time. Prior studies investigate layout control in text-to-image generation through cross-attention guidance and loss-guided diffusion (Chen et al., 2024; Patel & Serkh, 2025), and extend similar attention-based mechanisms to temporal control in text-to-video models (Schiber et al., 2025). However, most existing approaches remain tied to specific control interfaces or limited control types, leaving unified inference-time support for timing, structural, and reference-level control underexplored.

G.3 Inference-Time Guidance

Guidance is a key mechanism for steering generative models toward desired semantics or constraints. In diffusion models, classifier guidance (Dhariwal & Nichol, 2021) uses gradients from an external classifier, while classifier-free guidance (Ho & Salimans, 2022) combines conditional and unconditional predictions without requiring an auxiliary classifier. Early work applied classifier guidance to diffusion-based speech generation, as exemplified by Guided-TTS (Kim et al., 2022). Recent extensions further broaden inference-time control, including AutoGuidance (Fu et al., 2024), AudioMoG (Wang et al., 2025a), and optimization-based methods for music generation such as DITTO and DITTO-2 (Novack et al., 2024b;a). Overall, guidance provides a flexible inference-time framework for aligning generative sampling with downstream objectives across diffusion and related probabilistic generative models (Song et al., 2020; Saharia et al., 2022).

H Potential Broader Impact and Limitations

Broader Impact. AnyAudio enables training-free, fine-grained control over text-to-audio generation, lowering the cost of producing audio content with precise event timing, long-form structure, and controllable reference attributes. Fine-grained controllability makes audio generation more accessible to creators who require specific timing, duration, or acoustic characteristics, such as filmmakers producing sound effects, accessibility tools generating descriptive audio, and researchers studying auditory perception. The training-free design also reduces the computational cost of supporting new control objectives, lowering the barrier for downstream applications and enabling rapid integration with existing pretrained backbones. Like other generative audio systems, AnyAudio could be misused to fabricate misleading sound effects or environmental audio that imitate real events, potentially supporting misinformation or audio-based deception.

Limitations. AnyAudio inherits the audio fidelity of its backbone and introduces additional inference overhead depending on active control modules. The differentiable pitch extractor in RAC also introduces a systematic gap between the guidance objective and the standard f_0 evaluation metric.

I Case Study

We provide qualitative comparisons across timing-controlled, long-form, reference-attribute, and holistic generation to complement the quantitative results and illustrate temporal alignment, long-horizon stability, reference-attribute preservation, and joint controllability.

I.1 TIMING-CONTROLLED AUDIO GENERATION

Prompt. A man speaks while motor vehicles pass by.

Timing condition (10.0s). a man speaks: [0.000, 6.150], [6.945, 7.512], [8.874, 9.181] · motor vehicles pass by: [0.000, 10.000]

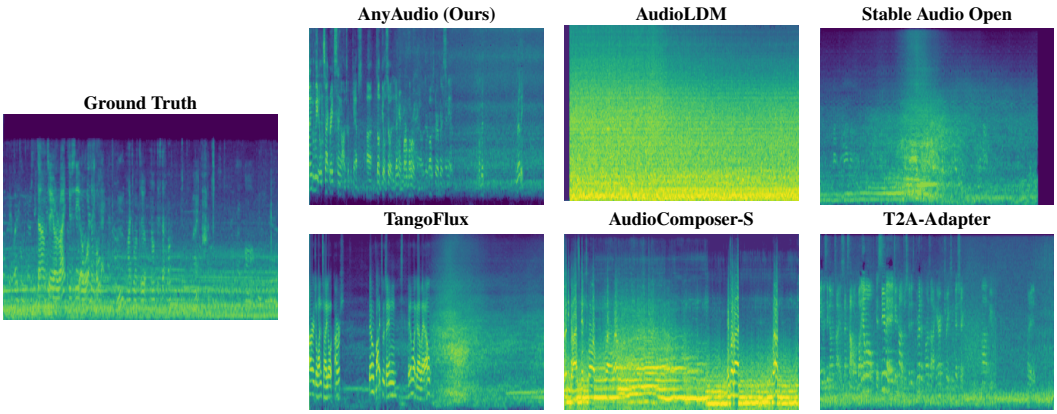


Figure 4: Timing-controlled audio generation, Case 1.

Prompt. Coins drop and a man speaks before a dog barks.

Timing condition (10.0s). Coins drop: [0.008, 1.197], [1.591, 2.016], [2.803, 3.693], [4.008, 5.606] · A man speaks: [6.913, 7.409] · A dog barks: [8.803, 9.417]

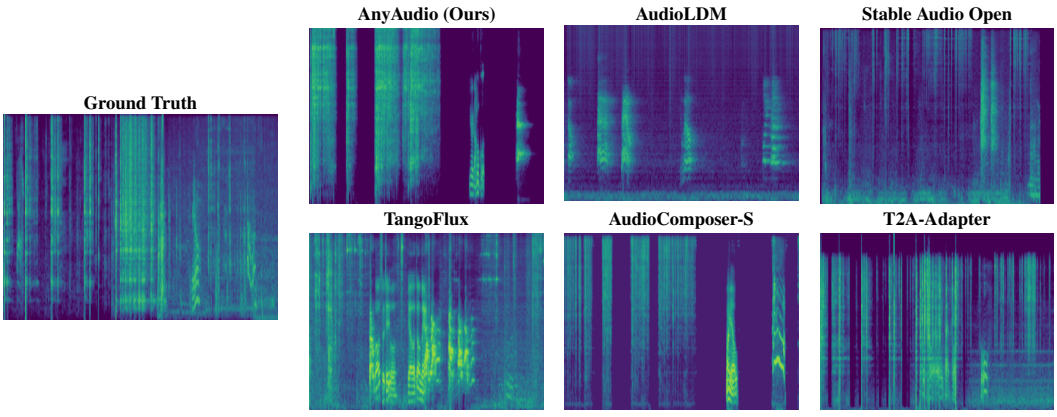


Figure 5: Timing-controlled audio generation, Case 2.

Prompt. A police car's siren and an alarm sound.

Timing condition (10.0s). A police car's siren: [0.000, 10.000] · an alarm sound: [3.764, 5.378]

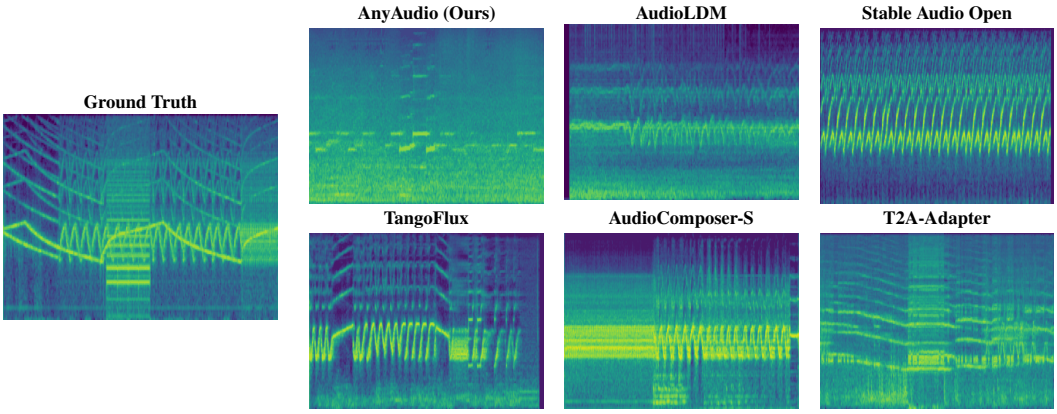


Figure 6: Timing-controlled audio generation, Case 3.

Prompt. A bathtub is being filled or washed.

Timing condition (10.0s). A bathtub is being filled or washed: [0.000, 6.598], [6.748, 10.000]

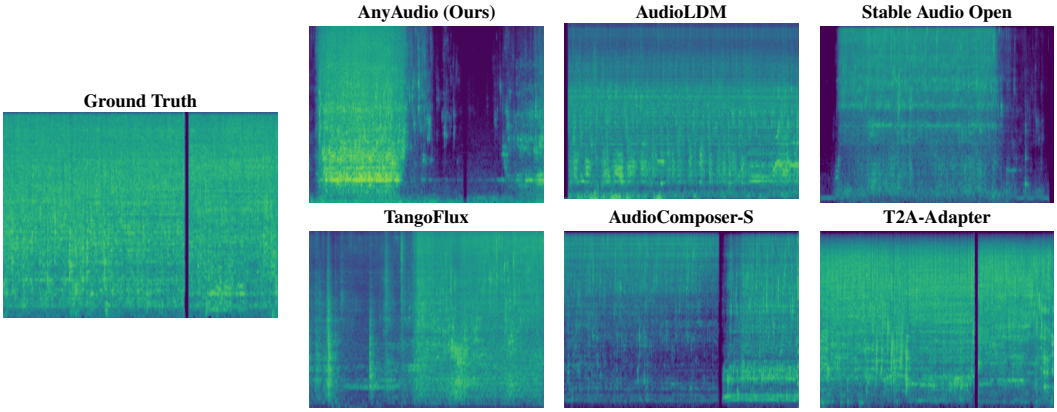


Figure 7: Timing-controlled audio generation, Case 4.

I.2 Long-Form Audio Generation

Prompt. Rain falling and wind blowing hard with rustling leaves.

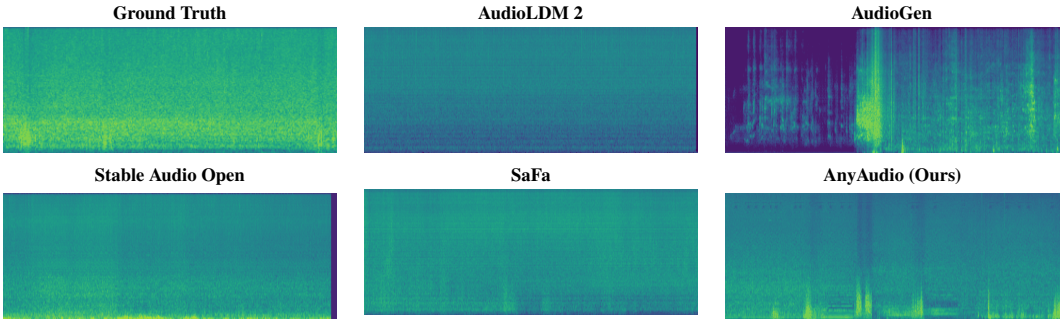


Figure 8: Long-form audio generation, Case 1.

Prompt. As music plays in the foreground, faint machine gunfire occurs in the background.

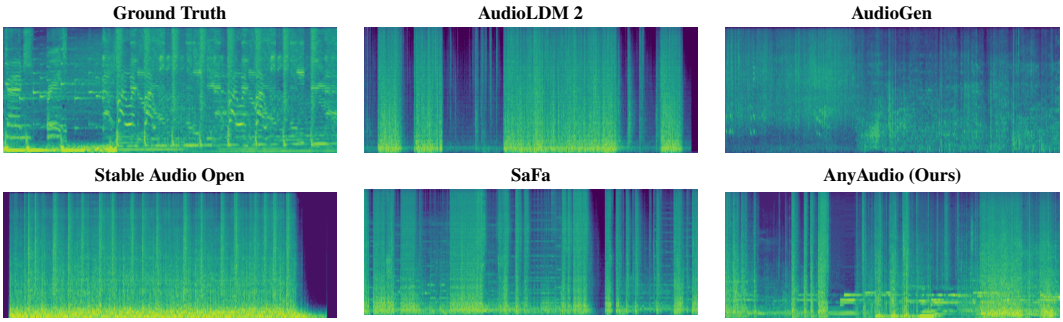


Figure 9: Long-form audio generation, Case 2.

Prompt. A male vocalist sings this energetic Punjabi folk song. The tempo is medium fast with an infectious tabla and Dhol percussive beat, ektara rhythm and a funky keyboard accompaniment. The song is lively, spirited, cheerful, simple, happy, playful, enthusiastic, vivacious with a festive, celebratory vibe and dance groove. This song is a Festive Punjabi Folk song.

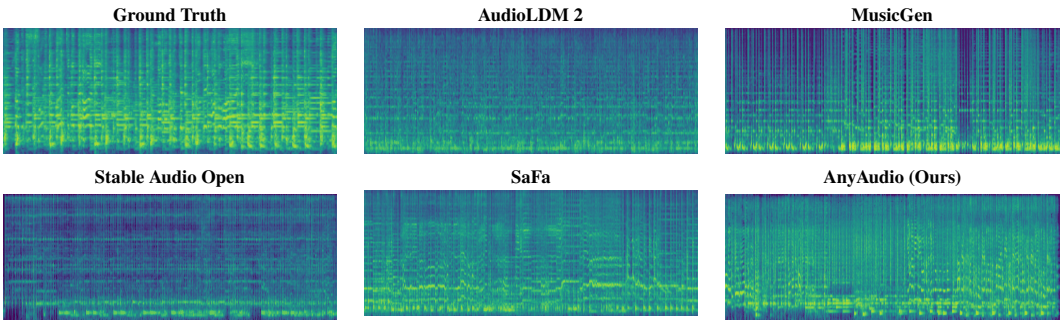


Figure 10: Long-form audio generation, Case 3.

I.3 Reference-Attribute Audio Generation

Prompt. A woman speaks repeatedly followed by beeping sounds.

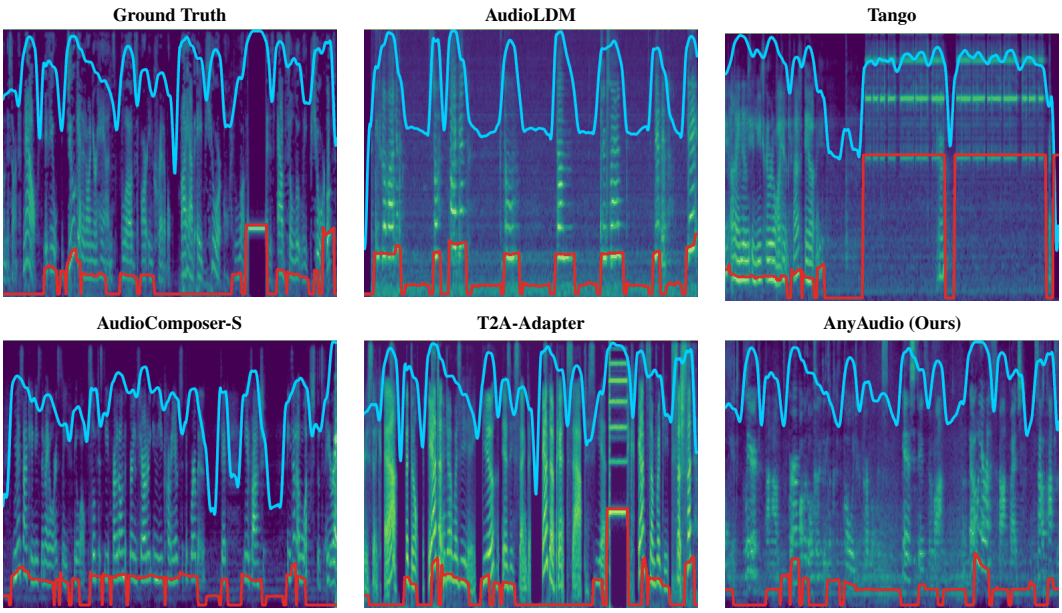


Figure 11: Reference-attribute audio generation, Case 1. The red curve denotes pitch, and the blue curve denotes loudness.

Prompt. Men sing and speak to music.

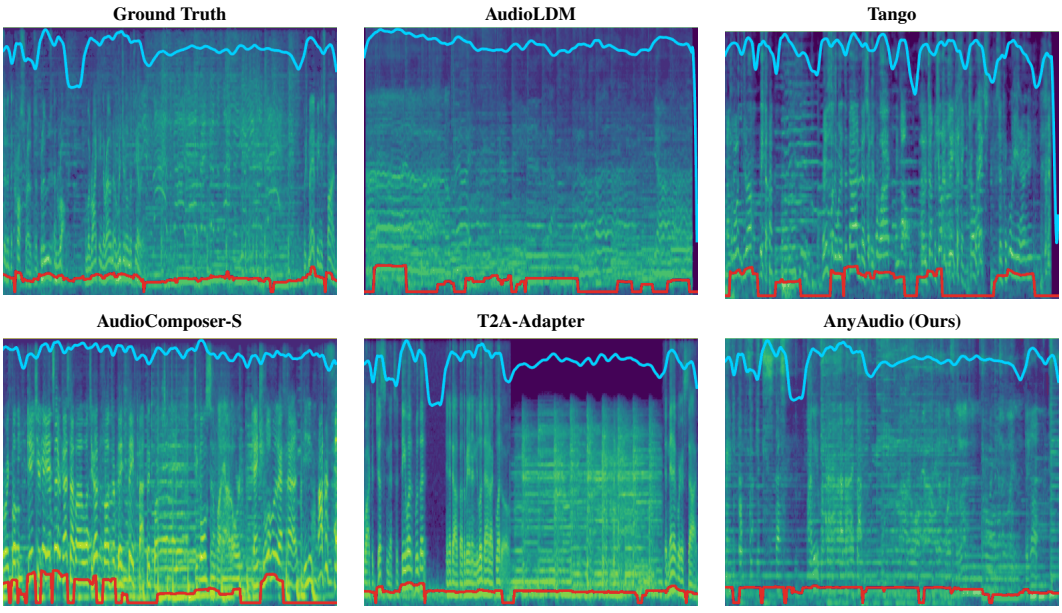


Figure 12: Reference-attribute audio generation, Case 2.

Prompt. Humans walk and speak while various sounds are heard in the background such as wind and field recordings.

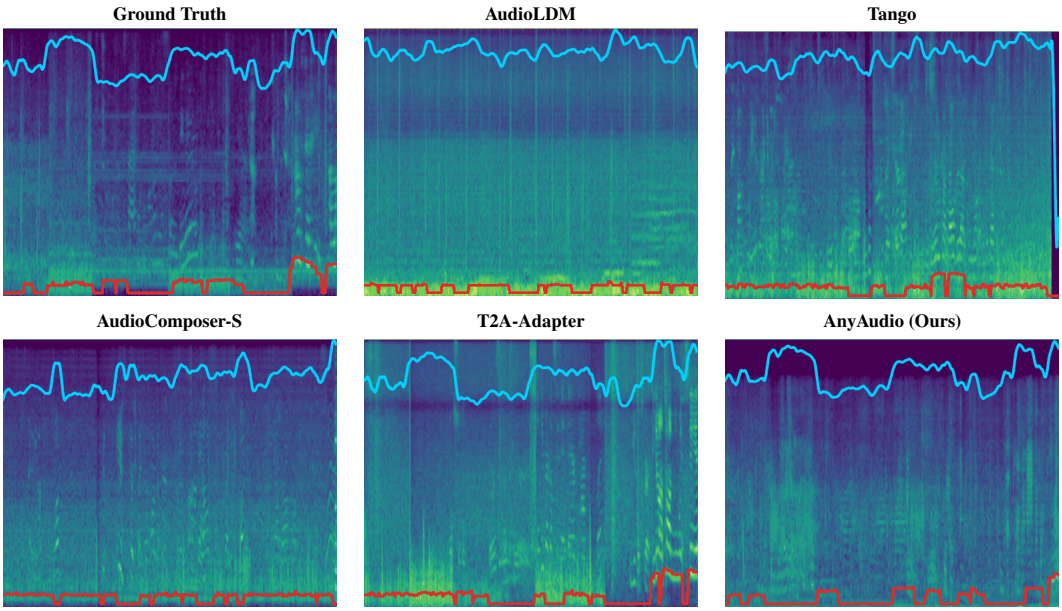


Figure 13: Reference-attribute audio generation, Case 3.

Prompt. A woman is speaking, with music and bird vocalization in the background.

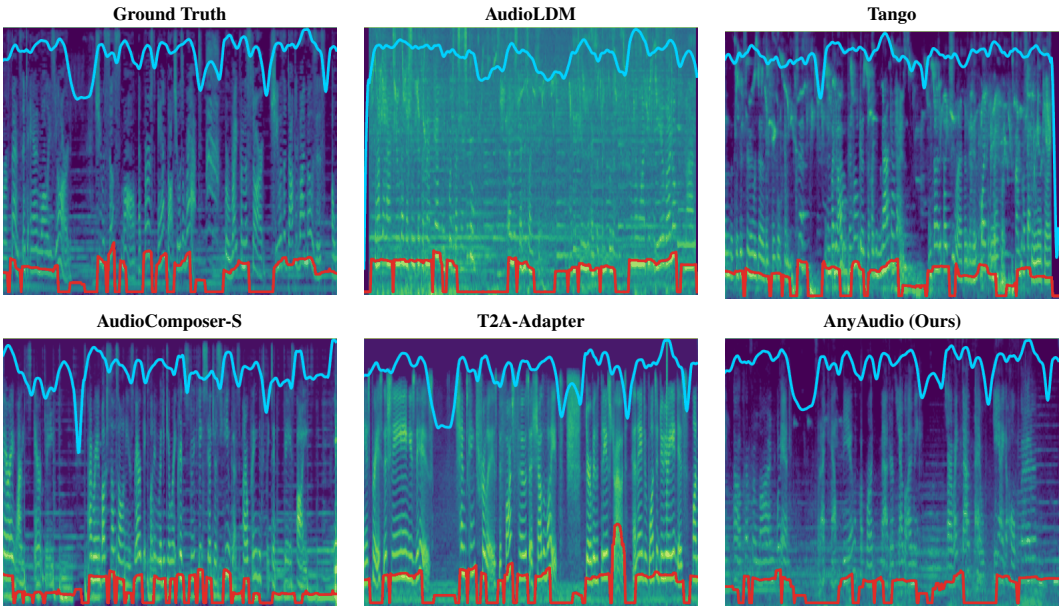


Figure 14: Reference-attribute audio generation, Case 4.

I.4 HOLISTIC AUDIO GENERATION

Prompt. A woman is speaking, humming, and babbling with music playing in the background.

Timing condition (26.0s). music: [0.000, 26.000] · hum: [0.000, 6.000] · babbling: [6.000, 16.000] · female speech: [16.000, 26.000]

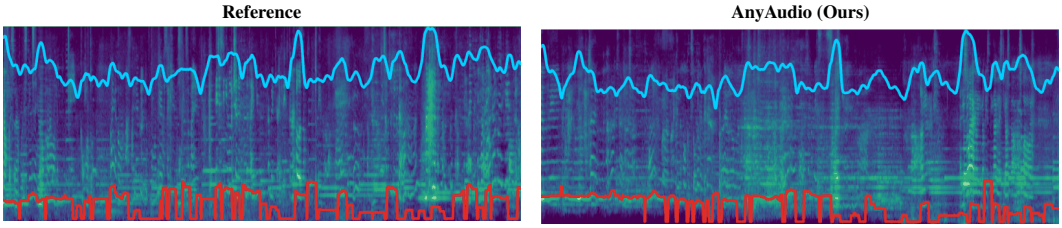


Figure 15: Holistic audio generation, Case 1.

Prompt. A sink is filled with water and music plays.

Timing condition (26.0s). music: [0.000, 26.000] · sink (filling or washing): [0.000, 26.000]

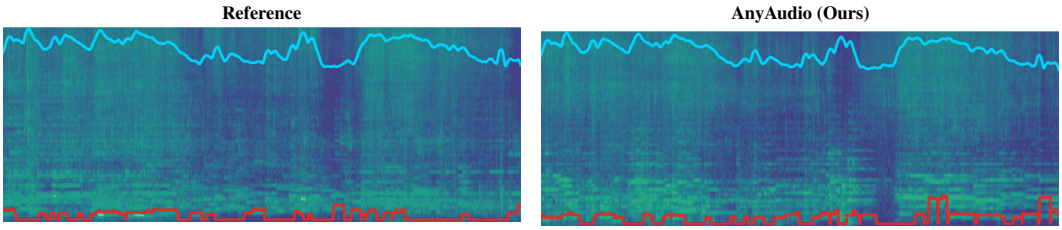


Figure 16: Holistic audio generation, Case 2.

Prompt. Men and women are speaking and honking horns in a busy, noisy setting.

Timing condition (26.0s). a busy, noisy setting: [0.000, 26.000] · men and women are speaking: [0.000, 26.000] · vehicle horn: [16.000, 17.000]

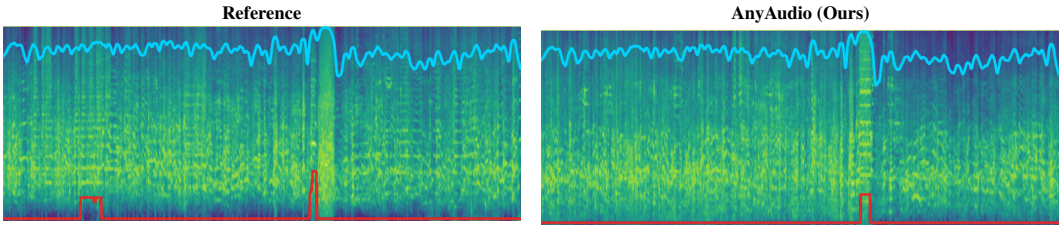


Figure 17: Holistic audio generation, Case 3.