UNISON: A Unified Sound Generation and Editing Framework via Deep LLM Fusion

Anonymous ACL submission

Abstract

We present **UNISON**, a latent diffusion framework that unifies speech generation, sound generation, and audio editing within a single model. A single model handles text-to-audio, text-to-speech, zero-shot speaker cloning, mixed speech-and-sound generation, scene-level audio editing, speech-in-scene editing, and timed temporal composition, all of which share a single set of weights. Our architecture features two core designs: (1) Layer-wise deep LLM fusion, which injects hidden states from uniformly sampled layers of a frozen MLLM into corresponding MM-DiT blocks via learned projections, providing depth-matched semantic conditioning that improves instruction following over single-layer baselines; and (2) a unified multi-task architecture where task identity is encoded solely by a channel-wise mask and source audio is provided through VAE-encoded channel concatenation. Training is stabilized by an online GPU-side multi-task data synthesis pipeline with task-homogeneous batching and a two-stage curriculum. With 621M–732M trainable parameters, UNISON achieves results competitive with or exceeding task-specialist models across evaluated domains, while being roughly $4\times$ smaller than comparable unified systems. Audio samples will be available at an anonymous repo: https://anonymous-for-conference22.github.io/UNISON-demo/

1 Introduction

A practical audio generation system should handle diverse tasks: generating sound effects from text descriptions, synthesizing intelligible speech in a target speaker's voice, inserting or removing specific acoustic events from recordings, and composing soundscapes with temporal structure. Currently, these tasks are typically addressed by specialized models trained on isolated datasets with disparate conditioning pipelines. This fragmentation increases deployment complexity and prevents cross-task knowledge transfer—particularly between generation and editing, which differ primarily in whether a source signal is present.

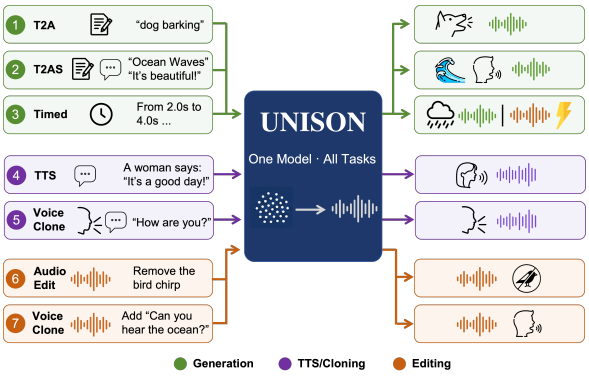

Figure 1: **Overview of UNISON.** A single flow-matching model handles text-to-audio generation, zero-shot TTS, gender control, audio-scene editing, and timed temporal composition. All tasks share the same architecture and weights, differentiated only by a task mask channel and optional source latent concatenation.

Recent work has moved toward unified systems. AudioBox (Vyas et al., 2023) unified speech and sound generation via flow-matching with in-context masking. MMAudio (Cheng et al., 2025) jointly trained on video-audio and text-audio data using an MM-DiT backbone. AudioOmni (Tian et al., 2026) expanded task coverage to include editing and multi-domain synthesis, while UniSonate (Qiang et al., 2026a) unified TTS, music, and sound effect generation through a phoneme-driven MM-DiT. Despite these advances, two fundamental limitations persist: **(1) Inconsistent latent spaces due to task-specific auxiliary modules.** Although these systems aim for unification, most still rely on heterogeneous components for different tasks—separate mel encoders for reference audio, dedicated phoneme front-ends for TTS, separate conditioning streams for editing versus generation, or specialized duration predictors. These

auxiliary modules fragment the latent space: each task operates in a subtly different representational regime, limiting cross-task knowledge transfer and complicating the training pipeline. A truly unified system should route all tasks through the *same* encoder, the *same* latent space, and the *same* forward pass, with task identity encoded minimally. **(2) Shallow text conditioning that discards hierarchical semantics.** A shared design choice across most existing systems is to condition the generative backbone on a single-layer text representation (typically the final hidden state of T5, CLAP, or an MLLM), which is fed identically to all DiT layers. Probing studies on transformer language models have shown that representations are organized hierarchically: lower layers primarily encode lexical and syntactic information, while higher layers capture more abstract semantic content (Tenney et al., 2019; Clark et al., 2019). Feeding only the final-layer embedding into a generative model discards this hierarchy, potentially limiting instruction-following capacity for compositionally complex audio prompts that simultaneously specify speaker attributes, acoustic events, and temporal structure.

To address these problems, we propose **UNISON** with the following contributions:

A unified generation-and-editing multi-task architecture with an efficient online training pipeline. We design an architecture where all tasks (including the generation and editing of both speech and sound) share the exact same VAE, DiT backbone, and forward pass. Task identity is encoded by a single mask channel concatenated with the audio latent; source/reference audio is provided through the same frozen VAE used for targets. We build an online GPU-side data synthesis pipeline that constructs all task variations on-the-fly with task-homogeneous batching and a two-stage curriculum, enabling stable joint training of generation and editing objectives within one model.

Layer-wise deep LLM fusion for enhanced instruction following. We inject hidden states from uniformly sampled layers of a frozen Qwen2.5-Omni-7B text backbone into the corresponding MM-DiT double-stream blocks via learned linear projections. This provides depth-matched conditioning. Specifically, early DiT blocks receive shallow LLM representations encoding lexical and phonetic structure, while later blocks process abstract semantic features. This hierarchical alignment improves text adherence across tasks (validated in ablations, §4.4).

Comprehensive evaluation across diverse audio tasks. We evaluate UNISON across multiple benchmarks spanning T2A, TTS, zero-shot cloning, mixed generation, audio editing, speech-in-scene editing, and timed composition, demonstrating that a single checkpoint achieves competitive or superior results compared to task-specialist models across all evaluated domains.

2 Related Work

2.1 Audio and Speech Generation

Text-conditioned sound generation has converged on latent diffusion and flow matching (Lipman et al., 2023; Liu et al., 2023b). AudioLDM (Liu et al., 2023a) and AudioGen (Kreuk et al., 2023) pioneered text-to-audio with latent diffusion and autoregressive token modeling, respectively; AudioLDM 2 (Liu et al., 2024), Make-An-Audio 2 (Huang et al., 2023), TangoFlux (Hung et al., 2026), GenAU (Haji-Ali et al., 2026), and MMAudio (Cheng et al., 2025) progressively improve quality through larger DiT (Peebles and Xie, 2023) models, preference optimization, or joint video-audio training, yet all condition the DiT on a *single* text layer (final T5/CLAP/LLM hidden state fed identically to every block). In TTS, neural codec language models such as VALL-E (Wang et al., 2023) demonstrated zero-shot cloning via in-context learning, inspiring modern flow-matching systems (F5-TTS (Chen et al., 2025), E2-TTS (Eskimez et al., 2024), MaskGCT (Wang et al., 2025), CosyVoice (Du et al., 2024a,b), ZipVoice (Zhu et al., 2025)) that achieve strong quality but still rely on task-specific text front-ends such as phoneme encoders, character encoders, or duration predictors. UniSonate (Qiang et al., 2026a) unifies TTS, T2A, and music with Qwen2.5-7B but still requires G2P phonemes and [SFX] tokens, and does not support cloning or editing.

UNISON differs in three ways: (i) Instead of phoneme/G2P pipelines, UNISON's transcripts are plain-text LLM instructions, with zero-shot speakers encoded by the *same* frozen VAE as targets; (ii) It feeds *per-block* projected LLM hidden states rather than a single-layer embedding, providing depth-matched semantic conditioning for compositional prompts. (iii) It unifies generation and editing: one checkpoint handles T2A, TTS, T2AS, zero-shot cloning, and scene editing via a task mask channel and VAE-encoded source latents, trained with an online multi-task pipeline, without separate heads or inversion stacks per task.

2.2 Audio Editing

Auido editing ranges from word-region speech tools (FluentSpeech (Jiang et al., 2023), EdiTTS (Tae et al., 2022)) to scene-level manipulation of mixed audio. UNISON focuses on the latter. ZETA (Manor and Michaeli, 2024) and SDEdit (Meng et al., 2022) edit via DDPM inversion or noise–denoise schedules. MMEDIT (Tao et al., 2025) trains an MM-DiT on synthetic pairs with a separate Qwen2-Audio (Chu et al., 2023) encoder. Audio-Omni (Tian et al., 2026) uses hybrid MLLM cross-attention plus a mel channel for editing.

UNISON treats editing as conditional generation: The source audio is encoded with the same VAE as a channel-concatenated input latent specified by a condition mask. This avoids inversion, auxiliary mel encoders, and task-specific decoders while preserving spectral detail in the latent domain.

2.3 Unified Architectures and Representation Fusion

UniAudio (Yang et al., 2024) and AudioBox (Vyas et al., 2023) unify multiple tasks via next-token prediction and flow-matching infilling, respectively. The closest concurrent systems are Audio-Omni (Tian et al., 2026) and UniSonate (Qiang et al., 2026a). Audio-Omni (3.05B DiT) feeds only the *penultimate* MLLM layer through cross-attention and routes mel/video through a second stream, which separates semantics from low-level cues but duplicates conditioning paths; its TTS is primarily evaluated on English. UniSonate (1.30B) uses last-layer Qwen features with phoneme-driven MM-DiT and omits editing and reference-based cloning. As summarized in Table 9 (Appendix A), UNISON (621M–732M) is, to our knowledge, the first to jointly offer layer-wise deep fusion, plain-text bilingual zero-shot TTS, scene-level editing, and timed composition in one MM-DiT without phoneme or mel side-encoders.

On the representation side, routing frozen LLM hidden states layer-by-layer into a DiT improves text–image alignment (Tang et al., 2025) and scales to large visual generators (Cai et al., 2025); BAGEL (Deng et al., 2025) further interleaves language and visual tokens. Audio instructions are often more compositional than image captions (speaker + lexicon + background + timestamps), making depth-matched fusion particularly beneficial. UNISON is the first to apply this principle inside a unified audio model covering generation, cloning, and editing.

3 Method

Figure 2 and the following subsections describe the architecture for generation, editing, and TTS. Frozen modules (VAE, Qwen) provide latents and text features; the trainable DeepFusion MM-DiT predicts a flow-matching velocity field.

3.1 Overview and notation

End-to-end pipeline. For each training sample we (i) build a text instruction and optional source waveform, (ii) encode waveforms with the frozen VAE to obtain $\mathbf{z}$ (target) and $\mathbf{z}_s$ (source/reference), (iii) sample flow time t and form the noisy target $\mathbf{z}_t$, (iv) concatenate $[\mathbf{z}_t \,\|\, \mathbf{z}_s \,\|\, \mathbf{m}]$ and embed it to audio tokens $\mathbf{h}_0$, (v) run the frozen LLM once and feed per-block text conditions $\tilde{\mathbf{E}}_k$ into the trainable MM-DiT, and (vi) predict the velocity v_θ and back-propagate a flow-matching loss. Inference repeats step (vi) with an ODE solver starting from noise, then VAE-decodes the denoised latent.

Symbols. C and T' denote VAE latent channels and frames; d=1024 is the DiT token dimension; D is the number of double-stream blocks; L=28 is the number of Qwen layers; N is the instruction length in tokens.

Symbol	Shape	Meaning
$\mathbf{z}$	$C \times T'$	Clean **target** latent (audio to generate)
$\mathbf{z}_s$	$C \times T'$	**Source** latent (zeros for pure generation)
$\mathbf{m}$	T'	Per-frame **task tag** in $\{0, 1, 2\}$
$\boldsymbol{\epsilon}$	$C \times T'$	Gaussian noise for flow matching
$\mathbf{z}_t$	$C \times T'$	**Noised** target at flow time t
$\mathbf{X}$	$(2C{+}1) \times T'$	DiT channel input $[\mathbf{z}_t \,\|\, \mathbf{z}_s \,\|\, \mathbf{m}]$
$\mathbf{h}_k$	$T' \times d$	Audio tokens after block k
$\mathbf{E}^{(l)}$	$N \times 3584$	Qwen hidden states at layer l
$\tilde{\mathbf{E}}_k$	$N \times d$	Text tokens fed to DiT block k
v_θ	$C \times T'$	Predicted velocity (on target channels only)

3.2 Audio VAE

We adopt the MMAudio continuous VAE (Cheng et al., 2025). A waveform is converted to a mel spectrogram (STFT), then encoded to $\mathbf{z}$ or $\mathbf{z}_s \in \mathbb{R}^{C \times T'}$ (C=40 at 44.1 kHz, C=20 at 16 kHz). The same VAE encodes targets, edit sources, and speaker references, and decodes the denoised latent at inference.

3.3 Multi-task inputs

All tasks share one network; only $(\mathbf{z}, \mathbf{z}_s, \mathbf{m})$ and the text instruction change.

Task tag $\mathbf{m}$. Each latent frame carries a scalar tag (broadcast as one channel in $\mathbf{X}$):

- $\mathbf{m}$=0: *Generation* (T2A, TTS, T2AS, timed composition). $\mathbf{z}_s$=$\mathbf{0}$.
- $\mathbf{m}$=1: *Editing*. $\mathbf{z}_s$ is the VAE latent of the pre-edit mix; $\mathbf{z}$ is the post-edit target.

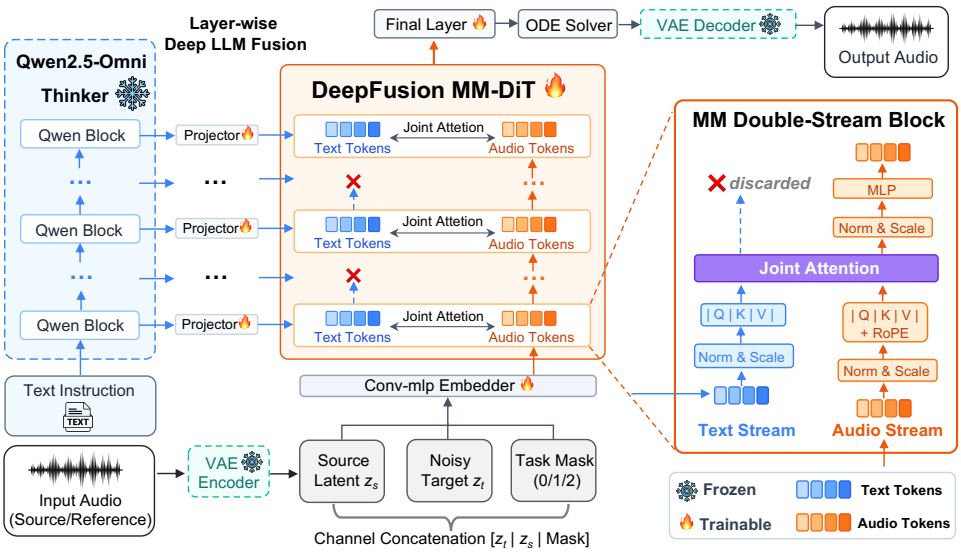

Figure 2: **UNISON Architecture. Left:** Layer-wise deep LLM fusion injects per-layer Qwen hidden states into corresponding DiT blocks via learned projectors. **Middle:** Each double-stream block performs joint attention; text tokens are refreshed per block (✗) while audio tokens pass through the MLP. **Bottom:** $[\mathbf{z}_t \,\|\, \mathbf{z}_s \,\|\, \mathbf{m}]$ are channel-concatenated and embedded; the ODE solver denoises the latent, which is VAE-decoded to waveform. See §3.4.

- $\mathbf{m}$=2: *Zero-shot TTS*. $\mathbf{z}_s$ encodes the reference prefix; tags distinguish the reference region from frames to synthesize.

Per-task construction of $(\mathbf{z}, \mathbf{z}_s)$ and instruction templates is summarized in Table 10 (Appendix B).

Duration. We use fixed-length padded latents without a separate duration head; trailing silence is learned implicitly. Timed events are specified in the text and parsed by the LLM.

3.4 DeepFusion MM-DiT

Building on the channel input $\mathbf{X}$ from §3.3, the trainable backbone is a flow-matching (Lipman et al., 2023; Liu et al., 2023b) MM-DiT (Peebles and Xie, 2023; Cheng et al., 2025) with **layer-wise deep LLM fusion** (Tang et al., 2025). It maps $(\mathbf{X}, \{\tilde{\mathbf{E}}_k\}, t)$ to a velocity v_θ on the target channels. A default config: D=20 blocks (denoted as D20), d=1024, 8 heads.

Noising and channel input. Given clean $\mathbf{z}$ and noise $\boldsymbol{\epsilon}$, we sample σ_t and set $\mathbf{z}_t = (1-\sigma_t)\mathbf{z} + \sigma_t\boldsymbol{\epsilon}$. The DiT sees

$$\mathbf{X} = [\mathbf{z}_t \,\|\, \mathbf{z}_s \,\|\, \mathbf{m}] \in \mathbb{R}^{(2C+1)\times T'}. \tag{1}$$

A Conv-MLP embedder $\mathcal{E}$ maps $\mathbf{X}$ to $\mathbf{h}_0 \in \mathbb{R}^{T'\times d}$ (one token per frame). We *zero-initialize the weights* in $\mathcal{E}$ that connect $\mathbf{z}_s$ and $\mathbf{m}$ to the token space, while $\mathbf{z}_t$ uses a standard initialization—so early training behaves like denoising-only, and the model gradually learns to use source and task channels.

Text conditioning (left branch in Fig. 2). A frozen Qwen2.5-Omni-7B (Xu et al., 2025) Thinker runs once on the instruction, returning $\mathbf{E}^{(l)}$ for $l = 1, \ldots, L$. Since the number of DiT blocks D may differ from L, we uniformly pick $i_k = \left\lfloor 1 + k \cdot \frac{L-1}{D-1} \right\rfloor$ to get

$$\tilde{\mathbf{E}}_k = \mathbf{E}^{(i_k)}\mathbf{W}_k, \tag{2}$$

where $\mathbf{W}_k \in \mathbb{R}^{3584\times d}$ is the corresponding Linear projector. This ensures shallow DiT blocks see shallow Qwen layers (lexical/syntax), deep blocks see deeper semantics (Tenney et al., 2019).

Double-stream block (right branch in Fig. 2). Block k receives $(\mathbf{h}_k, \tilde{\mathbf{E}}_k)$. AdaLN injects t; *joint attention* lets all audio and text tokens attend to each other. Only $\mathbf{h}_k$ is updated by the MLP to form $\mathbf{h}_{k+1}$. We do **not** pass $\tilde{\mathbf{E}}_k$ to the next block; instead, each depth receives a fresh $\tilde{\mathbf{E}}_k$ from Qwen. Because $\tilde{\mathbf{E}}_k$ already encodes rich semantics from the frozen LLM, the DiT injects language information without relearning the full language structure. Skipping a text MLP also saves compute. RoPE (Su et al., 2023) applies to $\mathbf{h}_k$ only (indices $0, \ldots, T'-1$); QK-norm stabilizes attention.

Output. A linear head maps $\mathbf{h}_D$ to $v_\theta \in \mathbb{R}^{C\times T'}$ (target channels only; $\mathbf{z}_s$ and $\mathbf{m}$ are not predicted).

3.5 Training and inference

Loss. With target velocity $\mathbf{u} = \boldsymbol{\epsilon} - \mathbf{z}$, we minimize

$$\mathcal{L} = \mathbb{E}_{t,\mathbf{z},\boldsymbol{\epsilon}}\left[\left\|v_\theta(\mathbf{X}, \{\tilde{\mathbf{E}}_k\}, t) - \mathbf{u}\right\|_2^2 \odot \mathbf{M}_{\text{loss}}\right], \tag{3}$$

where $\mathbf{M}_{\text{loss}}$ zeroes the reference prefix in zero-shot TTS so gradients apply only to frames to synthesize. Text conditions are dropped with probability 0.1 for classifier-free guidance (Ho and Salimans, 2021).

Inference. Starting from $\mathbf{z}_t \leftarrow \boldsymbol{\epsilon}$, we integrate the learned velocity with a 100-step Euler ODE solver, then VAE-decode the denoised $\mathbf{z}$ to waveform. At inference we use CFG scale ω=4.5.

3.6 Online multi-task data synthesis

Rather than constructing static datasets for each task, we implement a GPU-side online synthesis pipeline that constructs task-specific tuples on-the-fly from raw audio and speech clips. Table 10 in Appendix B summarizes, for every task, how $\mathbf{z}_s$, $\mathbf{z}$, and the instruction template are assembled. The pipeline handles RMS normalization for SNR-controlled mixing, boundary fade-in/out, and randomized temporal offsets; instructions are assembled from predefined template pools.

3.7 Curriculum Training with Homogeneous Batching

To prevent gradient conflicts between generation and editing objectives, we employ:

Two-stage curriculum. Stage 1 trains only on generation tasks (T2A, TTS, zero-shot TTS, T2AS) for the first 150K steps, establishing a stable generative prior. Stage 2 introduces all editing tasks with the full task probability distribution (approximately 70% generation, 30% editing).

Task-homogeneous batching. Each mini-batch contains samples from a single task type, preventing intra-batch gradient conflicts between opposing objectives (e.g., "add event" vs. "remove event").

4 Experiments

4.1 Implementation Details

Model configurations. We train two variants: (1) **D20** (44kHz): 20 double-stream blocks, 40 latent channels, 44.1 kHz MMAudio VAE, 621M parameters; (2) **D24** (16kHz): 24 double-stream blocks, 20 latent channels, 16 kHz MMAudio VAE, 732M parameters. Both use the same frozen Qwen2.5-Omni-7B text encoder.

Training data. We train on a combined corpus of approximately 36M clips (∼57K hours; 2.3M for audio and 33.7M for speech); per-dataset details are listed in Table 11 (Appendix C).

Training configuration. AdamW optimizer (β_1=0.9, β_2=0.95), learning rate 10^{-4} with cosine decay and 2000-step warmup, weight decay 0.01, gradient clipping 1.0. Batch size 56 per GPU on 8×H800. BF16 mixed precision. EMA with decay 0.999, updated every 10 steps. CFG dropout probability 0.1. Base models trained on 10 s max duration; fine-tuned to 22 s for long speech only. Inference details are described in §3.5.

4.2 Evaluation Setup

Metrics. We use FAD (VGGish) and FD (PANNs (Kong et al., 2020)) for distributional quality; KL divergence and IS for classifier-based evaluation; CLAP (LAION-CLAP (Wu et al., 2023)) for text–audio semantic alignment; WER/CER via Whisper-large-v3 (Radford et al., 2023) (EN) and Paraformer (Gao et al., 2023) (ZH) for intelligibility; LSD for spectral fidelity against reference targets; gender accuracy via wav2vec2-large-XLSR-53 (Conneau et al., 2021); and speech removal rate via Silero-VAD (Team, 2024).

Evaluation data. For T2A and TTS we use standard benchmarks: AudioCaps (Kim et al., 2019) test (881 clips) and Seed-TTS (Anastassiou et al., 2024) test (1088 EN + 2020 ZH). For tasks without public benchmarks, we construct evaluation sets with a fixed seed: T2AS (600 samples, Seed-TTS speech + non-speech AudioCaps SFX at 0 dB); audio editing (1200, 400/sub-task, non-speech AudioCaps pairs at random SNR $\in [-3, 3]$ dB); speech-in-scene editing (600, 200/sub-task, AudioCaps backgrounds + Seed-TTS speech at 10 dB); gender TTS (300, balanced gender assignment); timed composition (150, 2–3 segment timelines). "GT CLAP" values in tables denote the CLAP of the pseudo-GT against its caption, serving as an empirical ground truth reference since the pseudo-GT itself is artificially mixed.

Baselines. For T2A: AudioLDM 2 (Liu et al., 2024), Tango (Ghosal et al., 2023), Stable Audio Open (Evans et al., 2025), Make-An-Audio 2 (Huang et al., 2023), GenAU-L (Haji-Ali et al., 2026), Audio-Omni (Tian et al., 2026), MMAudio-L (Cheng et al., 2025), UniSonate (Qiang et al., 2026a). For TTS: MaskGCT (Wang et al., 2025), CosyVoice 2 (Du et al., 2024b), ZipVoice (Zhu et al., 2025), E2-TTS (Eskimez et al., 2024), F5-TTS (Chen et al., 2025), InstructAudio (Qiang et al., 2026b), UniSonate, Audio-Omni. For editing: SDEdit (Meng et al., 2022), ZETA (Manor and Michaeli, 2024), MMEDIT (Tao et al., 2025), Audio-Omni. TangoFlux (Hung et al., 2026) is ex-

Table 1: Text-to-Audio on AudioCaps (Kim et al., 2019) test set (881 clips). GT CLAP = 0.526, GT IS = 11.25. Models marked with † use substantially larger training data or preference optimization.

Model	Architecture	Params	FAD↓	FD↓	KL↓	IS↑	CLAP↑
AudioLDM 2-Large	UNet	712M	3.097	29.68	1.490	7.98	0.452
Tango	UNet	866M	1.846	24.52	1.305	7.45	0.498
Stable Audio Open	DiT	1.06B	10.83	52.03	3.049	6.13	0.203
Make-An-Audio 2	DiT	937M	2.142	20.14	1.597	10.02	0.441
GenAU-L†	DiT	1.25B	1.591	18.41	**1.290**	11.94	**0.561**
Audio-Omni	DiT	3.05B	2.535	31.42	1.337	9.55	0.486
MMAudio-L	MM-DiT	1.03B	5.893	16.53	1.421	11.98	0.441
UniSonate	MM-DiT	1.34B	4.210	30.21	2.440	8.22	–
UNISON (D24, 16kHz)	MM-DiT	732M	**1.558**	16.28	1.459	10.90	0.503
UNISON (D20, 44kHz)	MM-DiT	621M	1.756	**15.82**	1.455	**12.04**	0.467

cluded because its preference optimization (CRPO) is orthogonal to architecture; GenAU-L is included for reference despite its 20× larger training set.

4.3 Main Results

4.3.1 Text-to-Audio Generation

As in Table 1, UNISON (D24, 16 kHz) achieves the best FAD (1.558) and CLAP (0.503) among comparable models, while D20 (44 kHz) obtains the lowest FD (15.82) and highest IS (12.04). Both outperform Audio-Omni (3.05B) and MMAudio-L (1.03B) on FAD despite being smaller. The low FD and high CLAP scores validate the effectiveness of layer-wise deep LLM fusion for semantic alignment (further confirmed by Table 8, where the L-only variant shows lower CLAP). The two variants show complementary strengths: D24's larger capacity favors FAD/CLAP, while D20's 44.1 kHz VAE better preserves spectral detail for FD/IS. GenAU-L achieves higher CLAP (0.561) but uses a 20× larger audio dataset and is a single-task model without editing or TTS capability.

4.3.2 Text-to-Speech

Table 2: TTS results on Seed-TTS test set. *Pure TTS*: instruction-based generation without speaker reference. *ZS TTS*: speaker cloning from a reference utterance.

Model	Params	English		Chinese	
		Pure WER↓	ZS WER↓	Pure CER↓	ZS CER↓
MaskGCT	1.05B	–	2.62	–	2.27
CosyVoice 2	618M	–	2.57	–	1.45
ZipVoice	123M	–	1.70	–	1.40
E2-TTS	333M	–	2.19	–	1.97
F5-TTS	336M	–	1.83	–	1.56
InstructAudio	1.30B	1.52	–	1.35	–
UniSonate	1.34B	1.47	–	1.25	–
Audio-Omni	3.05B	1.35	1.77	–	–
UNISON (D24)	732M	**1.27**	**1.50**	**0.92**	**0.89**
UNISON (D20)	621M	1.42	1.80	1.11	1.71

As shown in Table 2, UNISON (D24) achieves the lowest error rates across all settings: pure WER 1.27% (EN), CER 0.92% (ZH), zero-shot WER 1.50% and CER 0.89%. It outperforms Audio-Omni (3.05B, pure WER 1.35%) despite being ∼4× smaller, and surpasses dedicated TTS models such as ZipVoice (ZS WER 1.70%) and F5-TTS (ZS WER 1.83%). Notably, UNISON does not use a phoneme encoder (text conditioning is provided entirely via the frozen LLM), yet it matches models that rely on explicit G2P pipelines (UniSonate, MaskGCT). The D20 variant shows slightly higher WER (1.42% pure, 1.80% ZS), attributable to its smaller capacity and the increased modeling difficulty of 44.1 kHz audio. These results confirm that multi-task training does not degrade TTS quality.

4.3.3 Gender-Controlled TTS

As shown in Table 3, both variants achieve perfect gender accuracy (300/300) solely from text instructions (e.g., "A male voice saying...") without requiring explicit speaker embeddings or gender labels during training. WER/CER remain low (D24: 1.21% EN, 1.00% ZH; D20: 1.47% EN, 1.02% ZH), indicating that gender control introduces no intelligibility degradation compared to the standard TTS setting (Table 2).

Table 3: Gender-controlled TTS on a balanced bilingual test set (300 samples: 106 EN, 194 ZH, 150 male / 150 female). Test prompts from Seed-TTS (Anastassiou et al., 2024) with randomly assigned gender. Gender accuracy evaluated via wav2vec2-large-XLSR-53 fine-tuned on LibriSpeech (Conneau et al., 2021).

Metric	UNISON (D24, 16kHz)	UNISON (D20, 44kHz)
Gender Accuracy↑	100%	100%
WER-EN↓	1.21	1.47
CER-ZH↓	1.00	1.02
WER (male)↓	1.64	1.31
WER (female)↓	0.74	1.65
CER (male)↓	1.30	1.05
CER (female)↓	0.71	0.99

4.3.4 Mixed Speech + Sound Generation

As shown in Table 4, UNISON (D24) achieves CLAP 0.444 (93.3% of the pseudo-GT CLAP of

0.476), with WER 2.04% and CER 3.64% measured directly on the mixed output without source separation. The D20 variant shows slightly lower speech clarity (WER 3.44%, CER 5.80%) but achieves lower LSD (2.36 vs. 2.44), indicating better waveform-level fidelity to the pseudo-GT mixture. This task is unique in that no existing public benchmark or baseline exists for single-model joint speech+sound generation; UNISON handles it naturally by leveraging multi-task training on both T2A and TTS data without any dedicated mixing module or two-stage pipeline.

Table 4: T2AS: generating a unified output containing intelligible speech and a matching background soundscape from a joint instruction. The test set (600 samples) is constructed by pairing Seed-TTS speech entries with AudioCaps sound clips, mixed at 0 dB SNR as pseudo ground-truth. GT CLAP is computed on this pseudo-GT against the evaluation caption. WER/CER are measured directly on the mixed output (not separated). LSD is computed against the pseudo-GT.

Metric	UNISON (D24, 16kHz)	UNISON (D20, 44kHz)
CLAP↑ (GT: 0.476)	0.444	0.430
WER-EN↓	2.04	3.44
CER-ZH↓	3.64	5.80
LSD↓	2.44	2.36

4.3.5 Audio Editing

As shown in Table 5, UNISON (D24) achieves the best FD and CLAP across all sub-tasks, with overall FD 12.38 (vs. 20.60 for MMEDIT) and CLAP 0.364 (vs. 0.257), reaching 82% of the pseudo-GT CLAP. LSD remains $\leq$2.15 across all sub-tasks, confirming preservation of non-edited content. The "Remove" sub-task shows lower CLAP for all methods (D24: 0.308, MMEDIT: 0.221), reflecting the difficulty of spectral disentanglement. D20 achieves lower LSD due to its higher-bandwidth VAE but shows higher FD and lower CLAP.

By encoding source audio through the same frozen VAE used for the target, UNISON operates in a shared latent space, unlike SDEdit/ZETA (noise-injection/inversion) or Audio-Omni (separate mel encoder). Qualitative mel spectrograms are provided in Appendix F (Figs 3–4).

4.3.6 Speech-in-Scene Editing

Speech-in-scene editing manipulates spoken content within an existing audio scene (speech mixed with background sounds) by inserting, deleting, or rewriting speech while preserving the non-speech background intact. Results are shown in Table 6.

Table 5: Audio editing on 1200 constructed test samples (400 per sub-task). Source/target pairs are synthesized by mixing AudioCaps (Kim et al., 2019) test clips at random SNR (see §4.2). GT CLAP (in parentheses) is the CLAP score of the constructed target against the evaluation caption, serving as a pseudo-GT reference. LSD is computed between generated audio and the constructed target.

Task (GT CLAP)	Model	FD↓	LSD↓	CLAP↑
Add (0.429)	SDEdit	78.86	2.21	0.168
	ZETA	67.27	2.18	0.243
	MMEDIT	25.98	2.23	0.339
	Audio-Omni	34.92	1.99	0.332
	UNISON (D24)	**19.26**	1.49	**0.416**
	UNISON (D20)	20.18	**1.43**	0.391
Remove (0.485)	SDEdit	87.65	2.11	0.053
	ZETA	66.34	2.09	0.141
	MMEDIT	45.25	3.86	0.221
	Audio-Omni	64.00	2.51	0.112
	UNISON (D24)	**33.20**	2.15	**0.308**
	UNISON (D20)	37.93	2.18	0.169
Replace (0.417)	SDEdit	79.09	1.90	0.119
	ZETA	62.71	1.89	0.180
	MMEDIT	27.56	2.77	0.210
	Audio-Omni	55.39	1.82	0.202
	UNISON (D24)	**21.31**	1.68	**0.368**
	UNISON (D20)	23.09	**1.57**	0.307
Overall (0.444)	SDEdit	73.85	2.07	0.114
	ZETA	57.27	2.05	0.189
	MMEDIT	20.60	2.95	0.257
	Audio-Omni	36.29	2.11	0.217
	UNISON (D24)	**12.38**	1.77	**0.364**
	UNISON (D20)	13.44	**1.73**	0.289

D24 achieves 99.16% speech removal (Delete) with LSD 1.56, and maintains WER $\leq$1.35% for Insert/Rewrite, confirming effective voice suppression and high synthesized-speech intelligibility. D20 shows lower LSD across sub-tasks but lower CLAP and removal rate, consistent with the D24/D20 trade-off observed in audio editing. Qualitative examples are in Appendix F (Figs 5–6).

Table 6: Speech-in-scene editing (200 samples per sub-task). Test pairs constructed from AudioCaps backgrounds + Seed-TTS (Anastassiou et al., 2024) speech mixed at 10 dB SNR. CLAP computed on pseudo-GT. WER/CER measured directly on full output. LSD computed against constructed target. Removal rate via Silero-VAD (Team, 2024).

Sub-task	Model	CLAP↑	GT CLAP	LSD↓	WER↓	CER↓	Removal↑
Insert	D24	0.433	0.459	1.70	1.35	0.65	–
	D20	0.429	0.459	1.66	1.70	1.08	–
Delete	D24	0.412	0.468	1.56	–	–	99.16%
	D20	0.320	0.468	1.52	–	–	95.72%
Rewrite	D24	0.408	0.456	1.60	0.98	0.95	–
	D20	0.396	0.456	1.46	1.35	1.09	–

4.3.7 Timed Audio Generation

As in Table 7, both variants achieve per-segment CLAP $\geq$0.308, with overall CLAP exceeding the

per-segment value (D24: 0.345; D20: 0.405), indicating coherent holistic scenes despite some boundary softening. Temporal control relies purely on natural-language timestamp parsing by the frozen LLM without dedicated alignment modules. Mel spectrograms in Appendix G (Figs 7–8) confirm spectral alignment with specified time intervals.

Table 7: Timed composition (150 test samples with 2–3 segment temporal instructions). Per-segment CLAP measures semantic alignment within each time window; overall CLAP measures holistic scene quality.

Metric	UNISON (D24, 16kHz)	UNISON (D20, 44kHz)
Per-segment CLAP↑	0.308	0.311
Overall CLAP↑	0.345	0.405

4.4 Ablation Studies

We conduct ablations on three axes (LLM conditioning mode, stream architecture, and LLM scale) using the same training data and hyperparameters.

For LLM conditioning mode, we compare three strategies on the double-stream D24 architecture: (1) D24-O (deep fusion only): per-block projections from uniformly sampled LLM layers, no persistent text stream, text MLP disabled; (2) D24-L (penultimate layer only): a single projection from the penultimate LLM layer broadcast to all DiT blocks, text MLP enabled; (3) D24-OL (deep + penultimate): both mechanisms active simultaneously. **For stream architecture**, we compare D24-O (double-stream, 24 blocks, separate text/audio normalization and QKV) against S32-O (single-stream, 32 blocks, text and audio tokens share normalization, QKV projections, and MLP) with comparable FLOPs. **For LLM scale**, We test D24-O-3B, which replaces the default 7B Qwen2.5-Omni Thinker with a 3B variant to assess the effect of LLM capacity on conditioning quality.

Table 8 reveals several findings:

Deep fusion improves semantic following. D24-L (penultimate layer only) achieves the lowest CLAP (0.175) and highest FD (22.71) among D24 variants, confirming that broadcasting a single LLM layer provides weaker conditioning than depth-matched injection. Both D24-O and D24-OL achieve lower FD (20.46, 20.18) and higher CLAP (0.180, 0.187), demonstrating that per-block deep fusion better captures hierarchical text semantics for audio generation.

Redundant text tokens hurt TTS. D24-OL achieves the best FD (20.18) and CLAP (0.187), yet the highest WER (5.52%). In this variant, text tokens enter the DiT from *two* sources (the persistent last-layer stream and the per-block deep fusion projections), effectively duplicating the conditioning signal. This redundancy improves T2A semantic alignment but introduces noise that increases TTS difficulty. D24-O avoids this trade-off by using only per-block projections with *ephemeral* text tokens, achieving competitive FD/CLAP (20.46/0.180) while maintaining the lowest WER (4.33%).

Double-stream architecture is essential. S32-O (single-stream) shows the worst FD (23.19), lowest CLAP (0.169), and high WER (4.84%), despite using the same deep fusion as D24-O. Sharing normalization and QKV projections between text and audio tokens prevents modality-specific representations; the double-stream design avoids this by maintaining separate feature spaces with interaction only through joint attention.

LLM scale matters. D24-O-3B shows degraded performance on all metrics (FD: 20.46→21.53, CLAP: 0.180→0.174, WER: 4.33→5.61), confirming that richer LLM representations directly benefit both semantic following and speech intelligibility.

Table 8: Ablation on AudioCaps (T2A) and Seed-TTS EN (pure TTS). All variants use the same training data and hyperparameters (80K steps). "3B" denotes the Qwen2.5-Omni-3B Thinker in place of the default 7B.

Conditioning	Arch	Params	FD↓	CLAP↑	WER-EN↓
O (deep fusion, 7B)	D24	732M	20.46	0.180	**4.33**
L (penultimate-layer, 7B)	D24	975M	22.71	0.175	4.44
OL (deep + penultimate, 7B)	D24	1,063M	**20.18**	**0.187**	5.52
O (deep fusion, 7B)	S32	685M	23.19	0.169	4.84
O (deep fusion, 3B)	D24	694M	21.53	0.174	5.61

5 Conclusion

We presented UNISON, a framework that unifies audio generation and editing through layer-wise deep LLM fusion and a channel-concatenation architecture that routes all tasks through a single VAE, DiT backbone, and forward pass. A single 621M–732M parameter checkpoint achieves competitive or superior results across T2A, TTS, zero-shot cloning, audio editing, and temporal composition without task-specific modules, demonstrating that multi-task audio generation at scale does not necessarily require heterogeneous conditioning paths. These results suggest a practical path toward general-purpose audio systems that grow in capability through data and model scaling rather than architectural specialization.

Limitations

VAE reconstruction quality. UNISON relies on the pre-trained MMAudio VAE, which was originally designed for environmental sound synthesis. While it provides a compact and effective latent space for general audio, its reconstruction fidelity for speech—particularly high-frequency formant details, subtle prosodic variations, and breathy or whispered voice qualities—imposes an upper bound on overall output quality. This is especially noticeable for zero-shot TTS, where fine-grained speaker timbre nuances may be smoothed out during VAE encoding. A natural next step is to train a unified VAE with improved speech reconstruction, potentially adopting higher latent resolution or a multi-scale architecture that better preserves both spectral detail and temporal dynamics.

Synthetic training data for editing. Our editing and T2AS training data is constructed by algorithmically mixing open-source audio clips (RMS-based overlay with random temporal placement and fade-in/out). While this approach validates the architectural design and enables large-scale training without manual annotation, the resulting data distribution differs from naturalistic recordings in several ways: (i) real-world audio scenes exhibit complex acoustic interactions (e.g., reverberation, occlusion, Lombard effects) that simple mixing cannot capture; (ii) caption quality for AudioSet/WavCaps sources has not undergone rigorous human verification, introducing label noise; (iii) the SNR distribution and temporal alignment of synthetic mixtures may not reflect typical editing scenarios encountered in practice. Future work will explore more realistic synthesis pipelines (e.g., room impulse response convolution, physically-informed mixing) and incorporate human-verified editing pairs.

Scale and modality coverage. The current model (621M–732M DiT parameters) is trained on ∼36M clips (∼57K hours). Both model size and data quantity are moderate relative to recent scaling efforts such as GenAU (1.25B params, 47M clips with synthetic captions). The architecture is designed to scale: the channel-concatenation mechanism naturally extends to additional modalities (e.g., video features for V2A generation) and the deep fusion framework can accommodate larger LLM backbones. We have not yet explored these directions but anticipate substantial gains from increased scale.

Language and domain scope. The current system supports English and Chinese speech; extension to other languages requires additional multilingual speech data but no architectural changes. We do not target music generation in this work, primarily because large-scale, openly licensed music datasets with high-quality text annotations remain difficult to obtain due to copyright restrictions. Additionally, music generation involves distinct challenges—long-range harmonic structure, multi-instrument arrangement, and beat/tempo consistency (Copet et al., 2023)—that may benefit from domain-specific design choices (e.g., hierarchical latent representations or music-aware tokenization) beyond our current scope. Nevertheless, the architecture itself is domain-agnostic and could incorporate music data if suitable training corpora become available.

References

Philip Anastassiou, Jiawei Chen, Jitong Chen, Yuanzhe Chen, Zhuo Chen, Ziyi Chen, Jian Cong, Lelai Deng, Chuang Ding, Lu Gao, and 1 others. 2024. Seed-tts: A family of high-quality versatile speech generation models. *arXiv preprint arXiv:2406.02430*.

Qi Cai, Jingwen Chen, Yang Chen, Yehao Li, Fuchen Long, Yingwei Pan, Zhaofan Qiu, Yiheng Zhang, Fengbin Gao, Peihan Xu, and 1 others. 2025. Hidream-i1: A high-efficient image generative foundation model with sparse diffusion transformer. *arXiv preprint arXiv:2505.22705*.

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

Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, JianZhao JianZhao, Kai Yu, and Xie Chen. 2025. F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6255–6271.

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

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. 2023. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. 2019. What does bert look at? an analysis of bert's attention. In *Proceedings of the 2019 ACL workshop BlackboxNLP: analyzing and interpreting neural networks for NLP*, pages 276–286.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2021. Unsupervised cross-lingual representation learning for speech recognition. In *Interspeech 2021*, pages 2426–2430.

Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. 2023. Simple and controllable music generation. In *Advances in Neural Information Processing Systems*, volume 36.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, and 1 others. 2025. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*.

Zhihao Du, Qian Chen, Xian Shi, Xiang Lv, Zhifu Gao, Changfeng Gao, Hui Wang, Dong Yu, Jianzong Pan, and Fan Wang. 2024a. Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens. *arXiv preprint arXiv:2407.05407*.

Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, Zhifu Gao, Yexin Yang, Changfeng Gao, Hui Wang, and 1 others. 2024b. Cosyvoice 2: Scalable streaming speech synthesis with large language models. *arXiv preprint arXiv:2412.10117*.

Sefik Emre Eskimez, Xiaofei Wang, Manthan Thakker, Canrun Li, Chung-Hsien Tsai, Zhen Xiao, Hemin Yang, Zirun Zhu, Min Tang, Xu Tan, and 1 others. 2024. E2 tts: Embarrassingly easy fully non-autoregressive zero-shot tts. In *2024 IEEE spoken language technology workshop (SLT)*, pages 682–689. IEEE.

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

Zhifu Gao, Zerui Li, Jiaming Wang, Haoneng Luo, Xian Shi, Mengzhe Chen, Yabin Li, Lingyun Zuo, Zhihao Du, Zhangyu Xiao, and Shiliang Zhang. 2023. Funasr: A fundamental end-to-end speech recognition toolkit. In *Interspeech 2023*, pages 1593–1597.

Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780. IEEE.

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

Moayed Haji-Ali, Willi Menapace, Aliaksandr Siarohin, Guha Balakrishnan, and Vicente Ordonez. 2026. Taming data and transformers for audio generation. *International Journal of Computer Vision*, 134(3):87.

Haorui He, Zengqiang Shang, Chaoren Wang, Xuyuan Li, Yicheng Gu, Hua Hua, Liwei Liu, Chen Yang, Jiaqi Li, Peiyang Shi, and 1 others. 2024. Emilia: An extensive, multilingual, and diverse speech dataset for large-scale speech generation. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 885–890. IEEE.

Jonathan Ho and Tim Salimans. 2021. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*.

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

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

Ziyue Jiang, Qian Yang, Jialong Zuo, Zhenhui Ye, Rongjie Huang, Yi Ren, and Zhou Zhao. 2023. Fluentspeech: Stutter-oriented automatic speech editing with context-aware diffusion models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11655–11671.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. 2019. 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)*, pages 119–132.

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

Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi. 2023. Audiogen: Textually guided audio generation. In *International Conference on Learning Representations*.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. 2023. Flow matching for generative modeling. In *International Conference on Learning Representations*.

Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley. 2023a. Audioldm: Text-to-audio generation with latent diffusion models. In *International Conference on Machine Learning*, pages 21450–21474. PMLR.

Haohe Liu, Yi Yuan, Xubo Liu, Xinhao Mei, Qiuqiang Kong, Qiao Tian, Yuping Wang, Wenwu Wang, Yuxuan Wang, and Mark D Plumbley. 2024. Audioldm 2: Learning holistic audio generation with self-supervised pretraining. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:2871–2883.

Xingchao Liu, Chengyue Gong, and Qiang Liu. 2023b. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations*.

Hila Manor and Tomer Michaeli. 2024. Zero-shot unsupervised and text-based audio editing using DDPM inversion. In *International Conference on Machine Learning*, pages 34603–34629. PMLR.

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

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. 2022. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*.

William Peebles and Saining Xie. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205.

Chunyu Qiang, Xiaopeng Wang, Kang Yin, Yuzhe Liang, Yuxin Guo, Teng Ma, Ziyu Zhang, Tianrui Wang, Cheng Gong, Yushen Chen, and 1 others. 2026a. Unisonate: A unified model for speech, music, and sound effect generation with text instructions. *arXiv preprint arXiv:2604.22209*.

Chunyu Qiang, Kang Yin, Xiaopeng Wang, Yuzhe Liang, Jiahui Zhao, Ruibo Fu, Tianrui Wang, Cheng Gong, Chen Zhang, Longbiao Wang, and 1 others. 2026b. Instructaudio: Unified speech and music generation with natural language instruction. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 17722–17726. IEEE.

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

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. 2023. Roformer: Enhanced transformer with rotary position embedding. *Preprint*, arXiv:2104.09864.

Jaesung Tae, Hyeongju Kim, and Taesu Kim. 2022. Editts: Score-based editing for controllable text-to-speech. In *Interspeech 2022*, pages 421–425.

Bingda Tang, Boyang Zheng, Sayak Paul, and Saining Xie. 2025. Exploring the deep fusion of large language models and diffusion transformers for text-to-image synthesis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 28586–28595.

Ye Tao, Wen Wu, Chao Zhang, Mengyue Wu, Shuai Wang, and Xuenan Xu. 2025. Mmedit: A unified framework for multi-type audio editing via audio language model. *arXiv preprint arXiv:2512.20339*.

Silero Team. 2024. Silero vad: pre-trained enterprise-grade voice activity detector (vad), number detector and language classifier. https://github.com/snakers4/silero-vad.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. Bert rediscovers the classical nlp pipeline. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 4593–4601.

Zeyue Tian, Binxin Yang, Zhaoyang Liu, Jiexuan Zhang, Ruibin Yuan, Hubery Yin, Qifeng Chen, Chen Li, Jing Lv, Wei Xue, and 1 others. 2026. Audio-omni: Extending multi-modal understanding to versatile audio generation and editing. In *ACM SIGGRAPH*.

Apoorv Vyas, Bowen Shi, Matthew Le, Andros Tjandra, Yi-Chiao Wu, Baishan Guo, Jiemin Zhang, Xinyue Zhang, Robert Adkins, William Ngan, and 1 others. 2023. Audiobox: Unified audio generation with natural language prompts. *arXiv preprint arXiv:2312.15821*.

Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, and 1 others. 2023. Neural codec language models are zero-shot text to speech synthesizers. *arXiv preprint arXiv:2301.02111*.

Yuancheng Wang, Haoyue Zhan, Liwei Liu, Ruihong Zeng, Haotian Guo, Jiachen Zheng, Qiang Zhang, Xueyao Zhang, Shunsi Zhang, and Zhizheng Wu. 2025. Maskgct: Zero-shot text-to-speech with masked generative codec transformer. In *International Conference on Learning Representations*, volume 2025, pages 47127–47150.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. 2023. 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)*, pages 1–5. IEEE.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, and 1 others. 2025. Qwen2.5-omni technical report. *arXiv preprint arXiv:2503.20215*.

Dongchao Yang, Jinchuan Tian, Xu Tan, Rongjie Huang, Songxiang Liu, Haohan Guo, Xuankai Chang, Jiatong Shi, Sheng Zhao, Jiang Bian, Zhou Zhao, Xixin Wu, and Helen M. Meng. 2024. Uniaudio: Towards universal audio generation with large language models. In *International Conference on Machine Learning*, pages 56422–56447. PMLR.

Zhuoyuan Yao, Di Wu 0061, Xiong Wang, Binbin Zhang, Fan Yu, Chao Yang, Zhendong Peng, Xiaoyu Chen, Lei Xie, and Xin Lei. 2021. Wenet: Production oriented streaming and non-streaming end-to-end speech recognition toolkit. In *interspeech*, volume 2021, pages 4054–4058.

Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. 2019. Libritts: A corpus derived from librispeech for text-to-speech. In *Interspeech 2019*, pages 1526–1530.

Han Zhu, Wei Kang, Zengwei Yao, Liyong Guo, Fangjun Kuang, Zhaoqing Li, Weiji Zhuang, Long Lin, and Daniel Povey. 2025. Zipvoice: Fast and high-quality zero-shot text-to-speech with flow matching. In *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*.

A Architectural Comparison

Table 9 provides a detailed side-by-side comparison of UNISON with the two closest concurrent unified audio systems (Audio-Omni and UniSonate), highlighting differences in LLM conditioning strategy, transcript encoding, reference audio handling, task coverage, and model scale.

B Online Multi-task Data Synthesis

Table 10 summarizes how each task's training tuple $(\mathbf{z}_s, \mathbf{z}, \text{instruction})$ is constructed on-the-fly from raw audio and speech clips during training. All synthesis is performed on GPU at data-loading time, requiring no pre-computed static datasets.

C Training Data Composition

Table 11 lists all datasets used for training, broken down by domain (audio vs. speech). The combined corpus contains approximately 36M clips totaling ∼57K hours. Audio clips with speech-heavy captions are filtered out at load time (∼390K removed); speech clips shorter than 3 s are excluded from zero-shot TTS sampling (∼29M eligible of 33.7M total).

D Task Probability Distribution

Table 12 shows the task sampling probabilities used during training. Stage 1 (first 150K steps) trains only on generation tasks; Stage 2 introduces editing tasks with the full probability distribution shown below.

E Architecture Details

Table 13 provides the complete set of architecture and training hyperparameters for both model variants (D20S0-44kHz and D24S0-16kHz), including VAE configuration, DiT dimensions, optimizer settings, and compute resources.

F Editing Qualitative Examples

Figures 3–4 and Figures 5–6 present mel spectrogram comparisons for audio editing and speech-in-scene editing tasks from both model variants, respectively. Each figure shows three columns—source audio (input to the model), UNISON's generated output, and the constructed ground truth—for one representative sample per subtask (add/remove/replace for audio editing; insert/delete/rewrite for speech editing). The instruction text is shown below each row. These visualizations complement the quantitative results in Tables 5 and 6, providing intuitive evidence that UNISON preserves non-edited content while accurately executing the specified modification.

G Timed Generation Qualitative Examples

Figures 7 and 8 visualize mel spectrograms of timed generation outputs from both model variants. Each panel corresponds to one generated sample; dashed vertical lines and colored shading mark the time boundaries specified in the natural-language prompt, with segment captions annotated above. The top two rows show sequential (non-overlapping) prompts, while the bottom two show overlapping segments. Across both models, the spectrograms confirm that distinct spectral patterns activate within the specified time intervals—validating that UNISON's temporal control operates purely through frozen-LLM instruction parsing without explicit alignment modules.

Table 9: Architectural comparison with Audio-Omni and UniSonate.

Aspect	Audio-Omni	UniSonate	UNISON (ours)
LLM conditioning	Penultimate MLLM layer → cross-attn	Qwen2.5-7B last layer → double-stream	Layer-wise: LLM layer i → DiT block i
Transcript encoding	ConvNeXt V2 character encoder	G2P phoneme + [SFX] tokens	Plain text via LLM; no extra module
Reference audio	Mel Encoder + channel-cat	Not supported	Same frozen VAE + channel-cat
Zero-shot TTS	English only (voice conversion)	Not supported	Bilingual (EN/ZH), VAE channel-cat
Audio editing	Hybrid stream (AudioEdit data)	Not supported	Same forward pass, mask=1
Music domain	Supported	Supported	Not targeted
Parameters	3.05B	1.30B	621M / 732M

Table 10: Online data synthesis: each task is constructed from base audio/speech clips at training time.

Task	Source latent $\mathbf{z}_s$	Target latent $\mathbf{z}$	Instruction format
T2A (generation)	Zeros	Audio clip	`[Audio] {caption}`
TTS (gender-controlled)	Zeros	Speech clip	`[Speech] A {gender} voice saying "{text}"`
Zero-shot TTS	VAE(ref + zeros pad)	Full utterance	`[Speech with voice] {text}`
T2AS (mixed speech+sound)	Zeros	RMS-mixed speech + SFX	`[Speech] ... [Audio] ...`
Audio edit — add	VAE(base)	Base + added event	`[Edit][Audio] Add {event}`
Audio edit — remove	VAE(base + event)	Base only	`[Edit][Audio] Remove {event}`
Audio edit — replace	VAE(base + old)	Base + new event	`[Edit][Audio] Remove {old} [Edit][Audio] add {new}`
Speech-in-scene — insert	VAE(background)	Background + speech	`[Edit][Speech] Add a voice saying "{text}"`
Speech-in-scene — delete	VAE(bg + speech)	Background only	`[Edit][Speech] Remove the speech`
Speech-in-scene — rewrite	VAE(bg + old speech)	Background + new speech	`[Edit][Speech] Change speech to "{new}"`
Timed composition	Zeros	Multi-event timeline	`[Audio] From {t1}s to {t2}s, {event}...`

Table 11: Training data composition. Sources include WavCaps (Mei et al., 2024), AudioSet (Gemmeke et al., 2017), VGGSound (Chen et al., 2020), LibriTTS (Zen et al., 2019), WenetSpeech (Yao et al., 2021), and Emilia (He et al., 2024).

Domain	Dataset	Clips	Hours
Audio	WavCaps	841K	2,189
	AudioSet (with labels)	1,718K	4,772
	VGGSound	174K	482
	Audio subtotal (after filter)	*2,342K*	*∼6,400*
Speech	LibriTTS	281K	332
	WenetSpeech	8,133K	7,693
	Emilia-EN (20M subset)	11,560K	19,802
	Emilia-ZH (20M subset)	13,759K	22,892
	Speech subtotal	*33,733K*	*∼50,700*
Total		**∼36M**	**∼57K**

Table 12: Task sampling probabilities in Stage 2 (joint training).

Task	Stage 1	Stage 2 Prob.
TTS (gender-controlled)	✓	0.15
Zero-shot TTS	✓	0.25
T2A (single event)	✓	0.10
T2A (mix)	✓	0.08
Speech + audio mix (T2AS)	✓	0.08
Timed composition	✓	0.04
Audio edit — add	×	0.04
Audio edit — remove	×	0.04
Audio edit — replace	×	0.03
Speech-in-scene — insert	×	0.04
Speech-in-scene — delete	×	0.03
Speech-in-scene — rewrite	×	0.04
Denoise	×	0.04
Other editing variants	×	0.04

Table 13: Full architecture and training hyperparameters.

Hyperparameter	**D20S0-44kHz (primary)**	**D24S0-16kHz**
Audio VAE	MMAudio 44.1kHz VAE	MMAudio 16kHz VAE
VAE latent channels (C)	40	20
DiT input channels (target + source + mask)	81 (40+40+1)	41 (20+20+1)
Sample rate	44,100 Hz	16,000 Hz
Max training duration	10 s (base) / 22 s (fine-tune)	10 s (base) / 22 s (fine-tune)
DiT hidden size	1,024	1,024
Attention heads	8	8
Head dimension	128	128
MLP width ratio	4.0	4.0
Double-stream blocks	20	24
Single-stream blocks	0	0
Patch embedder	`conv_mlp` (kernel=7)	`conv_mlp` (kernel=7)
RoPE dimensions	[96, 16, 16]	[96, 16, 16]
LLM encoder	Qwen2.5-Omni-7B (frozen)	Qwen2.5-Omni-7B (frozen)
LLM hidden dim	3,584	3,584
LLM total layers	28	28
Fusion mode	O (deep only)	O (deep only)
Layer selection	Uniform interval ($28 \rightarrow 20$)	Uniform interval ($28 \rightarrow 24$)
Per-block projector	Linear ($3584 \rightarrow 1024$)	Linear ($3584 \rightarrow 1024$)
Trainable parameters	621M	732M
Flow scheduler	FlowMatchEulerDiscrete	FlowMatchEulerDiscrete
Inference steps	100	100
CFG scale	4.5	4.5
Optimizer	AdamW (β_1=0.9, β_2=0.95)	AdamW (β_1=0.9, β_2=0.95)
Learning rate	1e-4 (cosine, 2000 warmup)	1e-4 (cosine, 2000 warmup)
Weight decay	0.01	0.01
Gradient clipping	1.0	1.0
Mixed precision	BF16	BF16
EMA decay	0.999 (update every 10 steps)	0.999 (update every 10 steps)
Training GPUs	8 $\times$ H800	8 $\times$ H800
Batch size per GPU	56	56

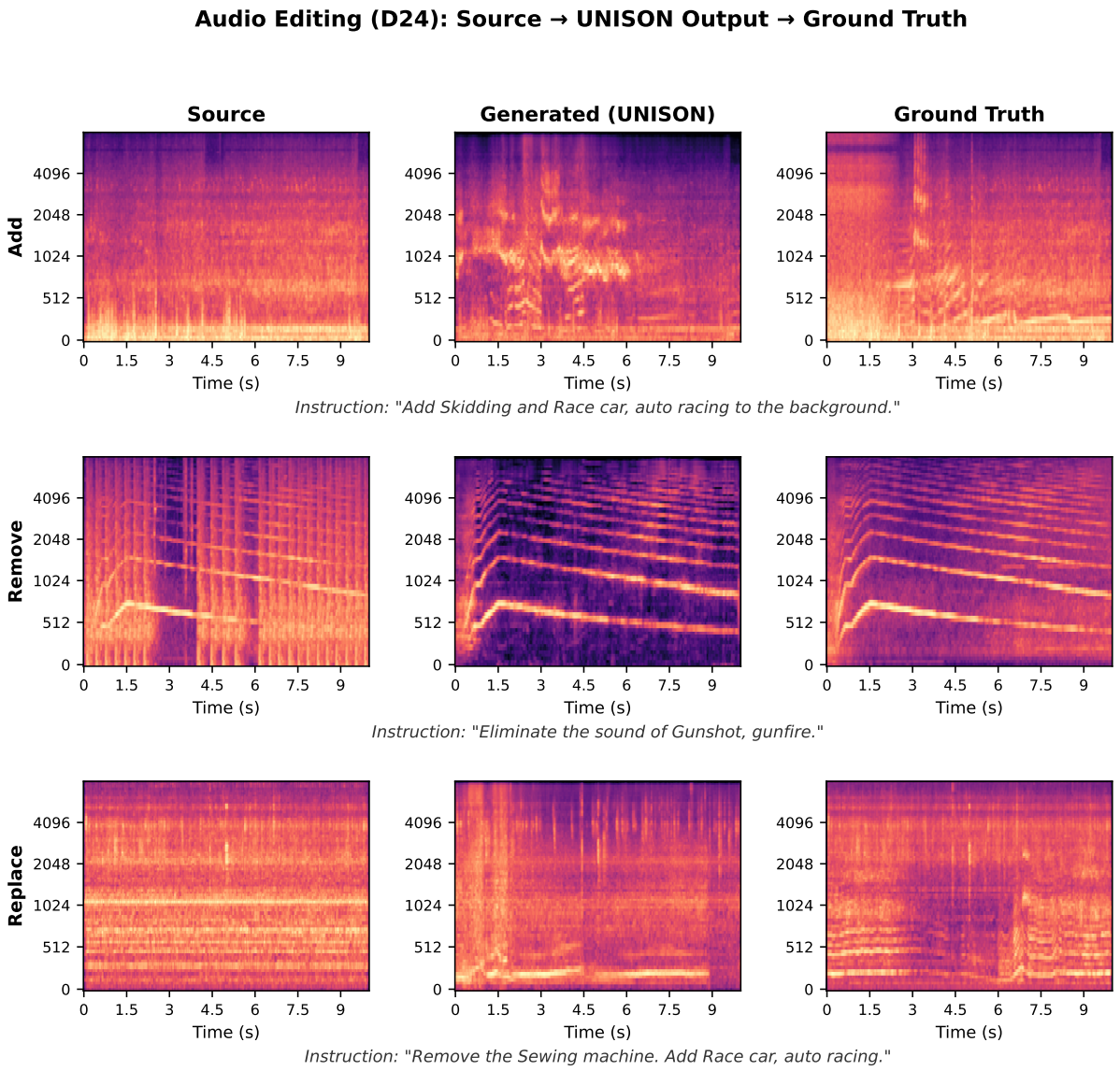

Figure 3: Audio editing qualitative examples from **UNISON (D24, 16 kHz)**. Each row shows one sub-task (Add / Remove / Replace). Left: source audio. Middle: UNISON output. Right: constructed ground truth.

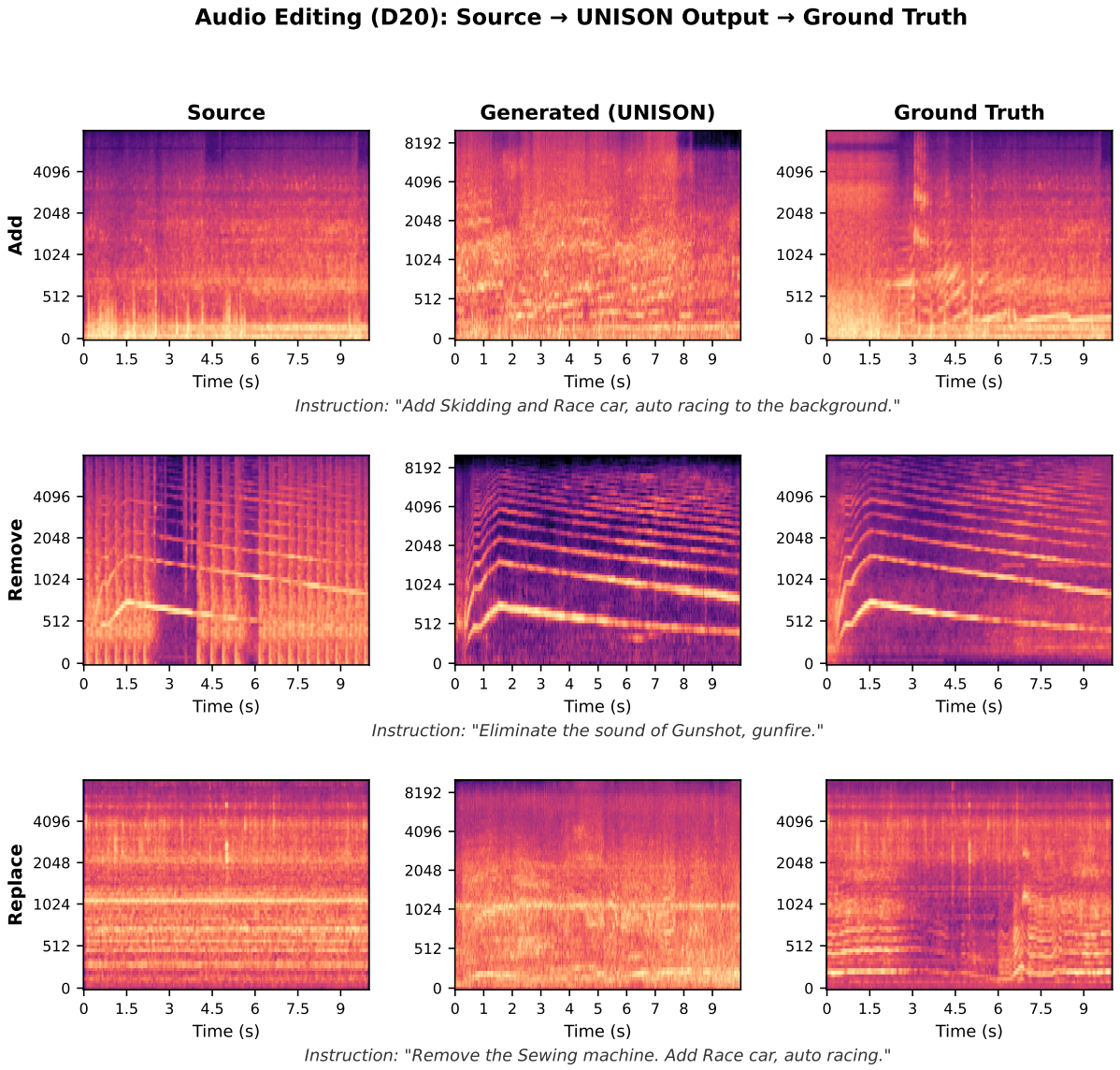

Figure 4: Audio editing qualitative examples from **UNISON (D20, 44.1 kHz)** on the same samples as Figure 3. The higher-bandwidth VAE preserves more spectral detail in both source and generated outputs.

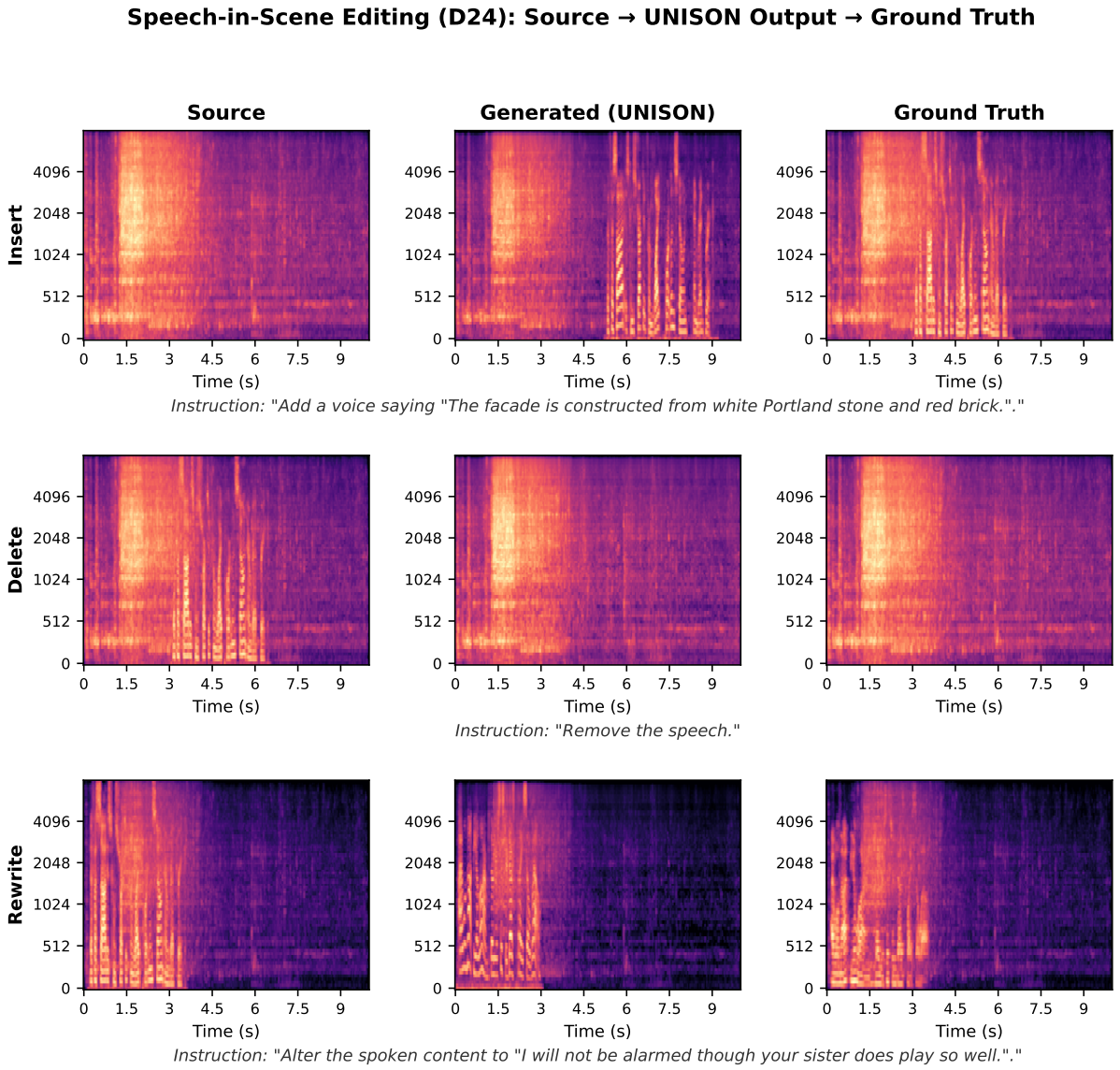

Figure 5: Speech-in-scene editing qualitative examples from **UNISON (D24, 16 kHz)**. Each row shows one sub-task (Insert / Delete / Rewrite). The model inserts speech, removes existing speech while preserving the soundscape, or rewrites spoken content.

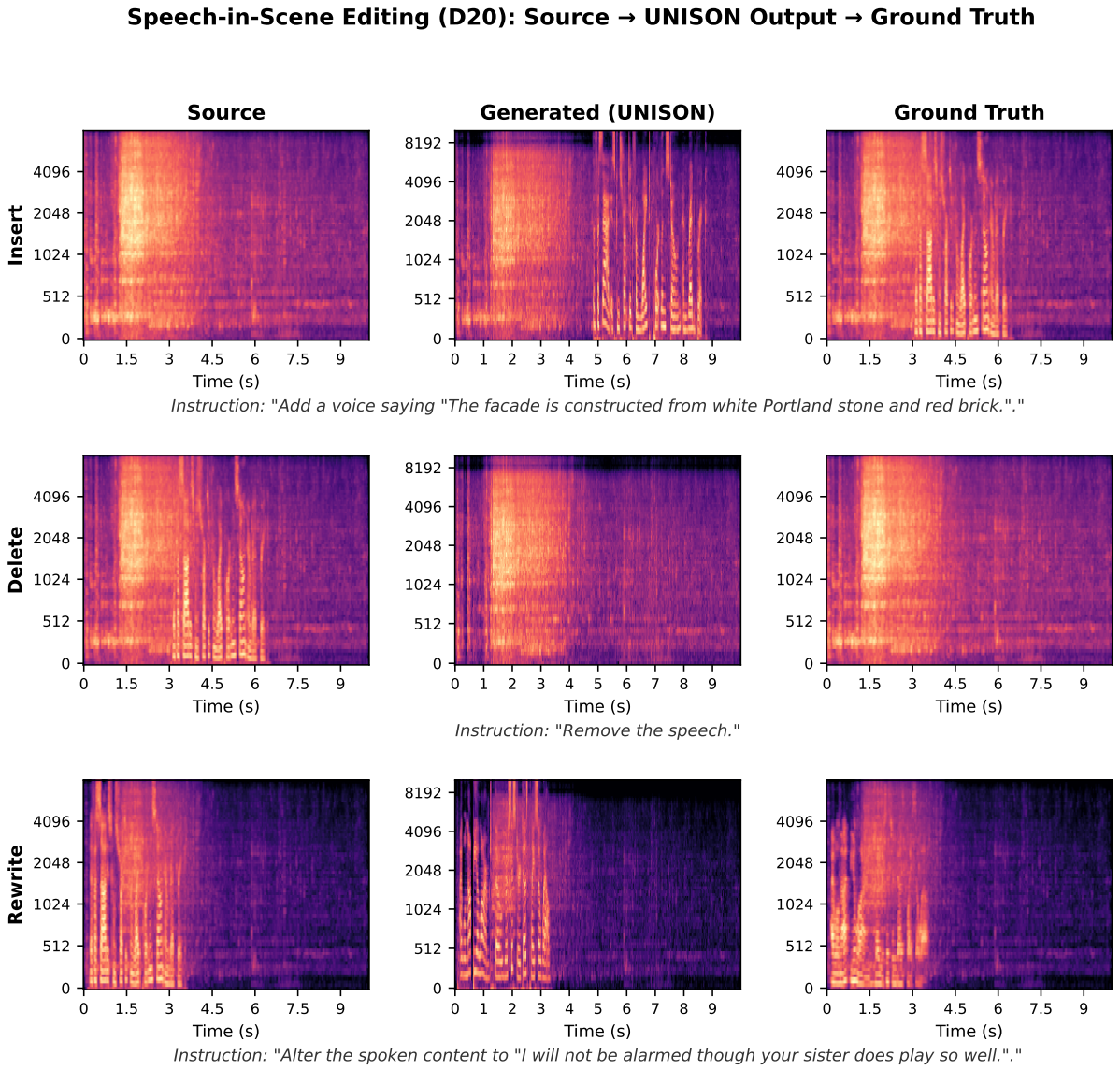

Figure 6: Speech-in-scene editing qualitative examples from **UNISON (D20, 44.1 kHz)** on the same samples as Figure 5.

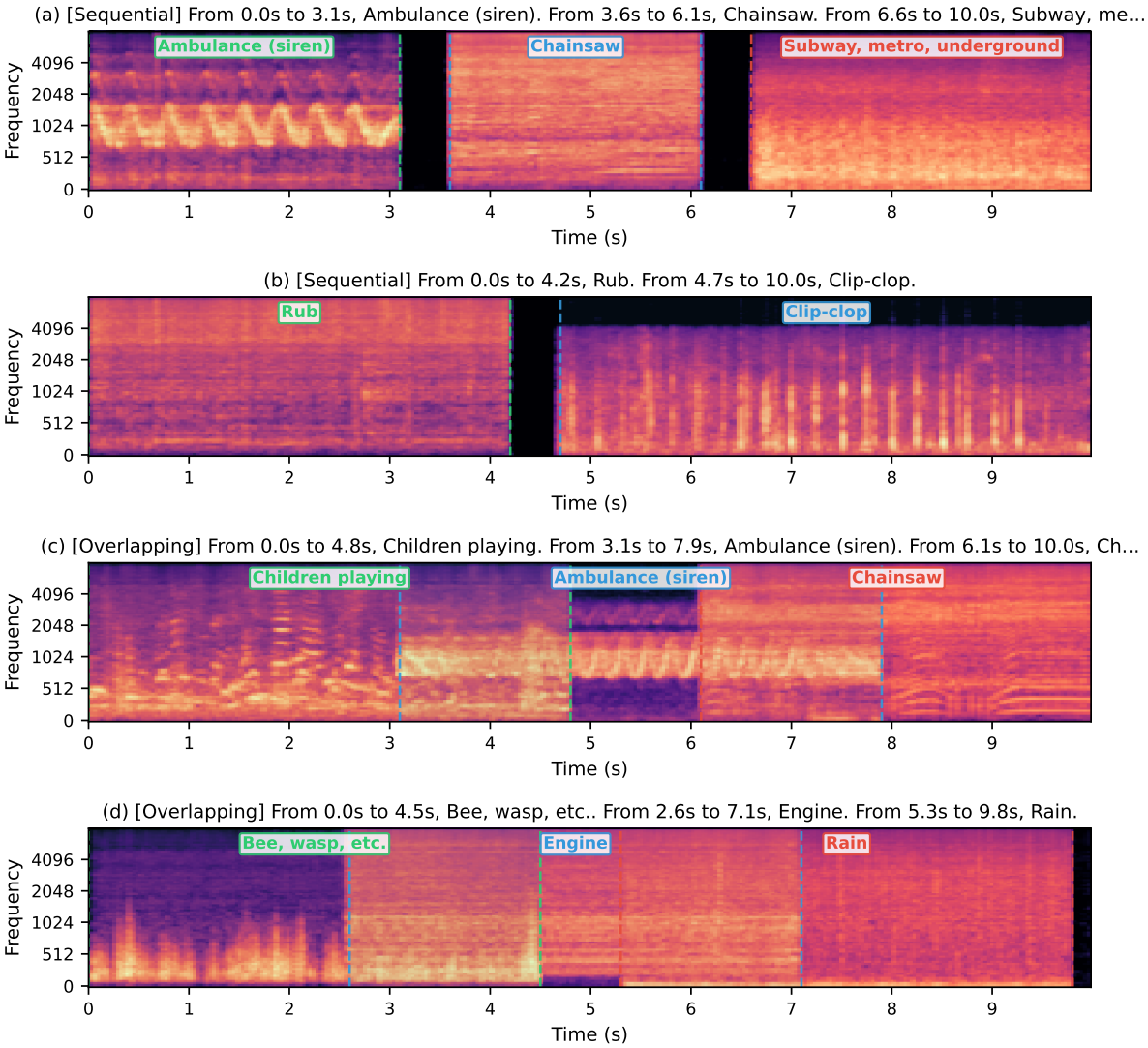

Figure 7: Timed generation mel spectrograms from **UNISON (D24, 16 kHz)**. Colored dashed lines and shading denote the time boundaries from the input prompt; segment captions are annotated above each region. (a)–(b): sequential segments. (c)–(d): overlapping segments. The model produces distinct spectral patterns that align with the specified time intervals.

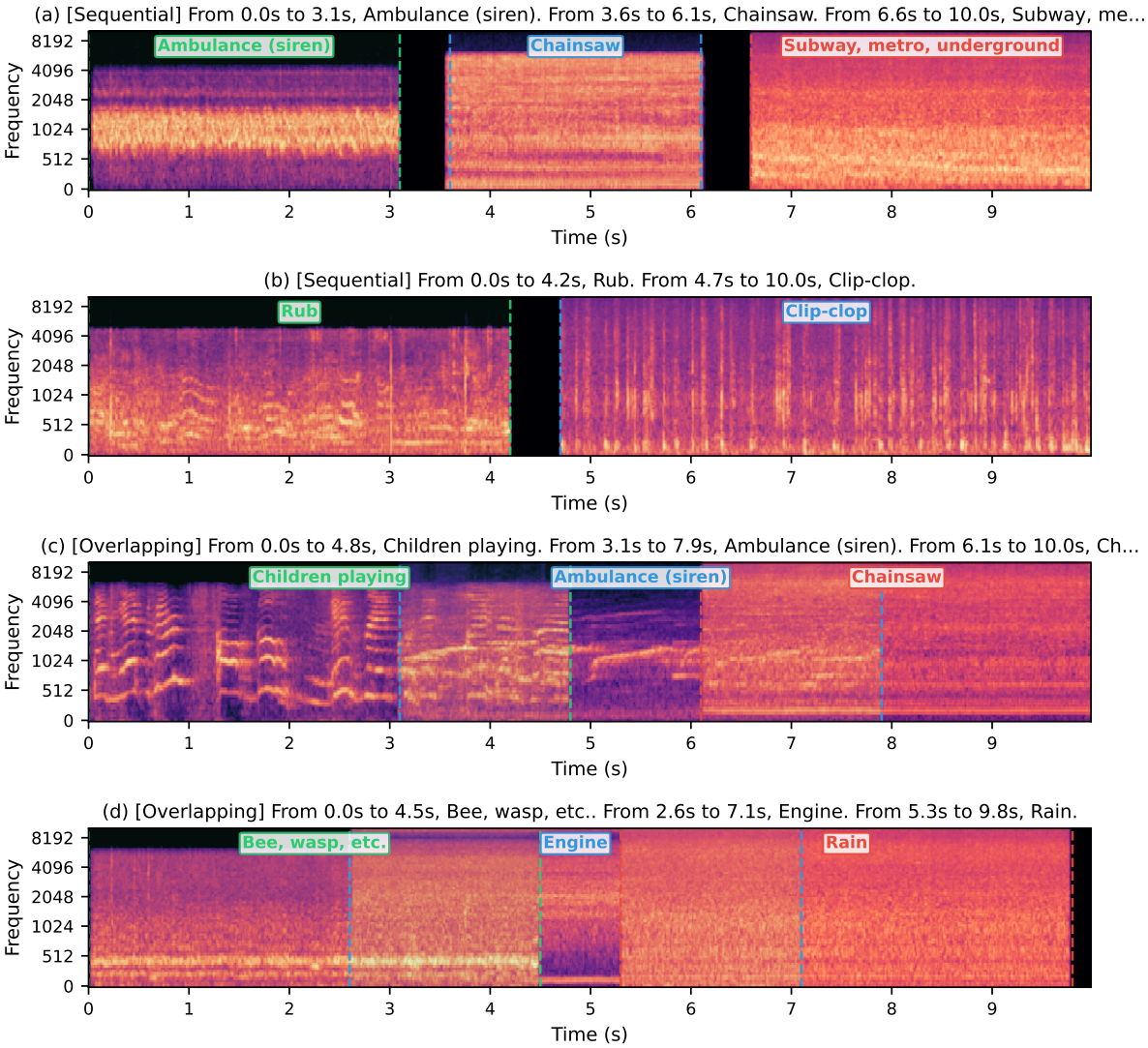

Figure 8: Timed generation mel spectrograms from **UNISON (D20, 44.1 kHz)** on the same prompts as Figure 7. The higher sample rate and 40-channel VAE produce richer spectral detail, particularly in the upper frequency bands. Temporal alignment with prompt boundaries remains consistent across both model variants.