Audio Editing in the Era of Foundation Models: A Survey

Anonymous ACL submission

Abstract

Audio editing aims to modify a given synthetic or real-world audio signal to satisfy specific user needs. As a promising yet challenging direction in AIGC, it has attracted increasing attention. Recent advances in audio generation have made powerful generative models central to modern audio editing systems. This rapid progress has created a growing need to organize emerging tasks, methods, and resources into a coherent view. In this survey, we provide a comprehensive review of audio editing in the era of foundation models. We first present a unified taxonomy of existing editing tasks and then summarize the major foundation-model paradigms that support modern audio editing, covering representative approaches from both training-based and training-free perspectives. We further discuss related resources, including datasets, evaluation protocols, and data construction tools. Finally, we identify open challenges in this field and outline promising directions for future research. The accompanying repository is released at https://github.com/AudioEditSurvey/AudioEditSurvey.

1 Introduction

Nowadays, in the territory of AI-generated content (AIGC) (DeepSeek-AI, 2026; Seedance et al., 2026; Comanici et al., 2025), audio editing has emerged as a crucial technique for practical applications in commerce, entertainment, and scientific research (Du et al., 2025; Lin et al., 2025). Unlike audio generation, which creates new audio from scratch, audio editing aims to modify a given recording according to user intent, such as changing its content, background, style, speaker identity, or acoustic conditions. These edits may range from fine-grained local modifications to global switches over the entire timeline, while preserving task-irrelevant acoustic content and temporal coherence.

Early audio editing workflows were heavily dependent on digital audio workstations[1] and signal-processing tools[2], requiring substantial manual effort and domain expertise. With the rapid progress of deep generative models (Liu et al., 2023b; Ho et al., 2020; Goodfellow et al., 2014), this paradigm is shifting toward data-driven and foundation-model-based approaches. Recent advances in diffusion models (Rombach et al., 2022; Peebles and Xie, 2023; Esser et al., 2024) and audio-language models (Kong et al., 2024; Kreuk et al., 2022b) have significantly broadened the boundary of audio editing, enabling more flexible, semantic, and instruction-driven manipulation.

Despite these rapid advances, foundation-model-based audio editing has not been systematically reviewed. Existing surveys on the audio modality mainly focus on audio-language models (Latif et al., 2023; Su et al., 2025), spoken language models (Ji et al., 2024), and general audio generation (Xie et al., 2025; Božić and Horvat, 2024). While some of them briefly mention editing-related tasks (Ma et al., 2024b), audio editing is often discussed only as a minor component of broader audio generation or manipulation frameworks. Thus, to address this gap, we conduct this survey to offer a comprehensive analysis specifically focused on audio editing in the era of foundation models.

In particular, this survey offers a systematic analysis of audio editing from four perspectives: editing tasks, foundation-model architectures, learning paradigms, and resources. First, in Section 2, we review representative studies in recent audio editing research and propose a task taxonomy based on input conditions and expected editing outcomes. This taxonomy divides audio editing into three major categories, namely acoustic editing, semantic editing, and instance editing, covering twelve fine-grained editing tasks. It highlights how di-

[1] https://www.adobe.com/products/audition.html
[2] https://www.steinberg.net/cubase

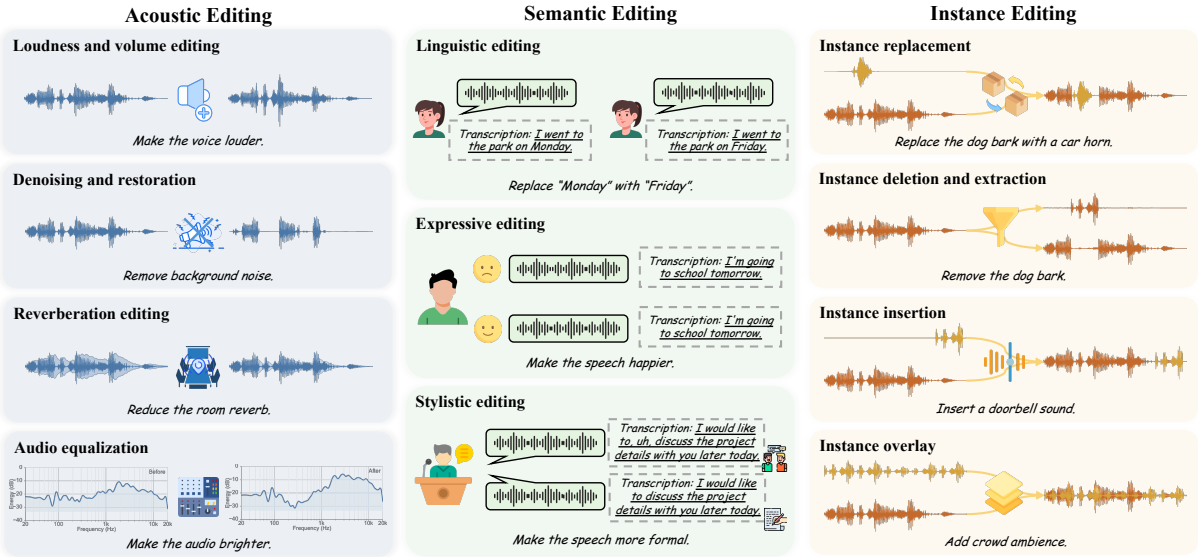

Figure 1: Taxonomy of audio editing tasks.

verse editing goals impose distinct requirements on controllability, temporal localization, and content preservation. Second, we examine the mainstream foundation-model architectures adopted for audio editing, and analyze their suitability for different editing scenarios in Section 3. Then, we organize existing methods according to their learning strategies, distinguishing between training-based and training-free approaches, and further categorize them by their core technical mechanisms. In addition, we provide a comprehensive summary of resources for audio editing, including public datasets, data construction tools, evaluation benchmarks, and metrics. We also discuss key challenges and future directions in this rapidly evolving field.

In summary, this survey aims to systematically classify and evaluate a large body of research in the field of audio editing under foundation models. We hope that our work can provide a comprehensive resource that not only synthesizes current findings but also guides future research directions in this rapidly evolving field.

2 Taxonomy

To systematically organize the audio editing in the foundational-model era, a taxonomy that is defined by the primary object itself is needed to answer one particular question: *which layer of the information of the audio is edited by the model?* As a result, we organize current audio editing tasks into the following three representative categories. Refer to Appendix C for task details.

Acoustic editing refers to transformations of signal-level attributes that affect auditory perception without changing the underlying semantic content, source identity, or temporal organization. It treats loudness, noise, reverberation, bandwidth, and spectral balance as controllable acoustic conditions. Representative tasks include *loudness and volume editing* for adjusting component salience (Jiang et al., 2024; Choi et al., 2021; Liang et al., 2024); *denoising and restoration* for removing recording or transmission degradations (Liu et al., 2022a; Fu et al., 2021); *reverberation editing* for modifying room response and decay patterns (Chen et al., 2022a; Bahrman et al., 2024); and *audio equalization* for reshaping spectral coloration (Venkatesh et al., 2022).

Semantic editing operates on high-level, interpretable information conveyed by audio, such as linguistic content, expressive intent, affective state, or stylistic semantics, while preserving non-target factors including speaker identity, source characteristics, acoustic context, and temporal coherence. Across speech, music, and general audio, it changes what the signal communicates or how it is perceived by listeners. *Linguistic editing* revises explicit language-bearing content within an existing recording (Tan et al., 2021; Zhang et al., 2022). *Expressive editing* modifies the manner in which audio content is delivered or emotionally perceived (Wang et al., 2024a; Kim et al., 2025). *Stylistic editing* changes the higher-level presentation style that shapes the overall character of speech, music, or sound events (Zhang et al., 2024b; Anastassiou et al., 2024; Jin et al., 2023).

Instance editing focuses on source-level entities in an audio scene, such as speakers, sound events, instruments, vocal tracks, or background components. Instead of primarily changing acoustic conditions or communicated meaning, it modifies the presence, identity, or source-to-mixture relation of selected entities while preserving non-target content, scene coherence, and temporal structure. *Instance replacement* substitutes a target entity with another source (Wang et al., 2021; Qian et al., 2019; Xu et al., 2024; Han et al., 2023; Fu et al., 2025). *Instance deletion and extraction* suppress or isolate selected sources from a mixture (Ellis et al., 2025; Liu et al., 2024c; Žmolíková et al., 2019; Défossez, 2021). *Instance insertion* introduces new entities into an existing recording (Wang et al., 2023b; Jin et al., 2017; Rouard et al., 2025; Peng et al., 2024). *Instance overlay* layers or emphasizes sources to enrich, rebalance, or remix an audio scene (Ellis et al., 2025; Strano et al., 2025).

3 Foundation Models for Audio Editing

3.1 Early Neural Editing Models

Before the foundation-model era, early neural audio editing methods explored task-specific generative models for local reconstruction and attribute control. For example, GAN-based speech inpainting reconstructs missing or corrupted time-frequency regions for localized repair (Zhao, 2023), while adversarial text-based speech editing modifies target regions in linguistic content and emotional attributes (Wang et al., 2024a). Although limited in scale and generalization, these methods introduced core ideas such as local generation, context-aware reconstruction, boundary consistency, and attribute-conditioned editing.

3.2 Token-based Codec Language Models

Token-based codec language models cast audio editing as conditional generation over discrete audio tokens. After continuous audio is converted into compact discrete token sequences, target regions are edited through autoregressive continuation, infilling, or selective regeneration conditioned on context, prompts, or task controls.

This paradigm is enabled by recent token-based audio generation models, which establish discrete tokens as a unified interface for modeling speech and music. For example, AudioLM (Borsos et al., 2023a) represents speech with semantic and acoustic tokens, while VALL-E (Wang et al., 2023a) applies neural codec language modeling to zero-shot speech synthesis with speaker prompts. MusicGen (Copet et al., 2023) extends this idea to text-conditioned music generation, and SoundStorm (Borsos et al., 2023b) improves decoding efficiency by parallel iterative token generation.

The explicit temporal structure of token sequences further makes this paradigm suitable for localized editing. VoiceCraft (Peng et al., 2024) treats a target region within an existing speech-token sequence as an infilling span, enabling zero-shot speech editing without regenerating the full utterance. SpeechX (Wang et al., 2024b) generalizes token-based editing to multiple speech transformation tasks, using discrete tokens as a shared representation for editing, extraction, and conversion. Later studies improve robustness and selectivity through multi-span editing, background-aware insertion and replacement, hallucination suppression, and stricter preservation of non-target areas (Wang et al., 2025d,b; Chen et al., 2025a).

3.3 Diffusion and Flow-Matching Models

Diffusion and flow-matching models formulate audio editing as conditional transformation in continuous acoustic spaces, such as mel-spectrograms or audio latents. Instead of infilling discrete tokens, they modify audio through conditional denoising, latent inversion, or continuous flow transformation, making them suitable for high-fidelity reconstruction, region-level refinement, and fine-grained acoustic control in complex scenarios.

Diffusion-based audio generation models provide the foundation for this paradigm by establishing iterative denoising as a strong framework for speech, general audio, and music synthesis. Representative examples include WaveGrad (Chen et al., 2020) for waveform generation, F5TTS (Chen et al., 2025c) and MegaTTS3 (Jiang et al., 2025b) for text-to-speech synthesis, AudioLDM (Liu et al., 2024a) for text-to-audio generation, and Moûsai (Schneider, 2023) and Stable Audio (Evans et al., 2024) for text-to-music generation. These models largely motivate later editing methods that adapt full-sample synthesis to source-conditioned and region-constrained modification.

The continuous generative process further makes this paradigm suitable for constrained audio modification. Editing-oriented methods adapt denoising or flow transformation by conditioning on source audio, masks, target regions, reference signals, or textual instructions, so that only the intended con-

Table 1: **Category definitions.** Row colors indicate domains: speech, general audio, and music. **Codec** and **Diff.** denote codec-based and diffusion-based models, respectively. **Text** denotes natural-language instructions, prompts, or transcript guidance; **Mask** denotes explicit mask or region guidance; **Ref.** denotes reference-audio conditioning. Editing categories follow our taxonomy: **Rest** denotes denoising and restoration; **Rev** denotes reverberation editing; **Loud** denotes loudness or volume editing; **EQ** denotes audio equalization; **Ling** denotes linguistic editing; **Expr** denotes expressive editing; **Sty** denotes stylistic editing; and **Rep**, **Del**, **Ins**, and **Ovl** denote instance replacement, deletion or extraction, insertion, and overlay, respectively.

Method	Arch	Cond	Acoustic Editing				Semantic Editing			Instance Editing			
			Rest	Rev	Loud	EQ	Ling	Expr	Sty	Rep	Del/Ext	Ins	Ovl
Training-based Methods													
FluentSpeech (Jiang et al., 2023)	Diff.	Text, Mask					✓	✓			✓	✓	
VoiceCraft (Peng et al., 2024)	Codec	Text					✓						
uSee (Yang et al., 2024)	Diff.	Text	✓	✓	✓							✓	✓
Step-Audio-EditX (Yan et al., 2025b)	Codec	Text	✓					✓	✓			✓	
InstructSpeech (Huang et al., 2024)	Codec	Text			✓		✓	✓	✓				
AUDIT (Wang et al., 2023b)	Diff.	Text	✓							✓	✓	✓	✓
SAO-Instruct (Ungersböck et al., 2026)	Diff.	Text	✓		✓	✓		✓	✓	✓	✓	✓	✓
Non-Rigid Prompt Edit (Paissan et al., 2023)	Diff.	Text	✓				✓		✓	✓		✓	✓
InstructME (Han et al., 2023)	Diff.	Text, Ref.			✓	✓			✓	✓	✓	✓	✓
Instruct-MusicGen (Zhang et al., 2024a)	Codec	Text			✓				✓		✓	✓	✓
Training-free Methods													
AST (Lv et al., 2026)	Diff.	Text					✓	✓	✓				
EdiTTS (Tae et al., 2021)	Diff.	Text, Mask					✓	✓					
DDPM Inversion (Manor and Michaeli, 2024)	Diff.	Text							✓	✓	✓		
AudioEditor (Jia et al., 2025a)	Diff.	Text							✓	✓	✓	✓	
PPAE (Xu et al., 2024)	Diff.	Text			✓				✓	✓	✓	✓	✓
AudioMorphix (Liang et al., 2025)	Diff.	Ref., Mask			✓	✓				✓	✓	✓	✓
MelodyFlow (Lan et al., 2024a)	Diff.	Text							✓				
MEDIC (Liu et al., 2024b)	Diff.	Text				✓		✓	✓	✓	✓	✓	
MusicMagus (Zhang et al., 2024b)	Diff.	Text				✓		✓	✓	✓			
MusRec (Boudaghi and Zare, 2025)	Diff.	Text				✓			✓	✓			

tent is regenerated. FluentSpeech (Jiang et al., 2023) and DiffEditor (Chen et al., 2024) apply diffusion models to localized speech refinement and text-based editing, improving contextual reconstruction and boundary consistency. uSee (Yang et al., 2024) extends diffusion editing to acoustic control with environmental conditions, while CosyEdit (Chen et al., 2026a) uses flow matching to improve region-specific consistency and alignment. Beyond speech, AUDIT (Wang et al., 2023b), PPAE (Xu et al., 2024), and AudioMorphix (Liang et al., 2025) develop instruction-guided, event-localized, and reference-guided audio editing methods, and MelodyFlow (Lan et al., 2024a) further extends continuous latent editing to music through flow matching and latent inversion.

3.4 Audio Editing Interfaces

Instructional and multimodal interfaces enable high-level control for foundation-model-based audio editing by converting textual instructions, reference audio, temporal regions, or visual cues into target spans, task embeddings, event locations, speaker references, or preservation constraints.

Representative works enable flexible, user-driven editing through such interfaces. Early efforts translate user requests into task-level or token-level control representations for unified editing, extraction, and conversion (Huang et al., 2024; Wang et al., 2024b). Recent studies extend these interfaces to finer-grained and more reliable control, including multi-span editing, background-aware modification, hallucination suppression, and non-target preservation (Wang et al., 2025b; Chen et al., 2025a; Huang et al., 2025; Wang et al., 2023b). These advances support broader scenarios such as sound-effect insertion, object replacement, soundscape remixing, and instructional music editing.

4 Training-based Approaches

Training-based approaches learn editing behaviors from paired, pseudo-paired, or instructional supervision. By optimizing editing objectives and preservation constraints, they provide stable control over predefined or user-specified edits. As shown in Figure 2, we categorize them according to their supervision and conditioning mechanisms.

4.1 Task-Specific Training

Task-specific training optimizes audio editing models for predefined functions, such as speech editing, prosody correction, source separation, and music stem separation. It typically relies on paired data with structured task-dependent conditions, including transcripts, masked regions, alignment informa-

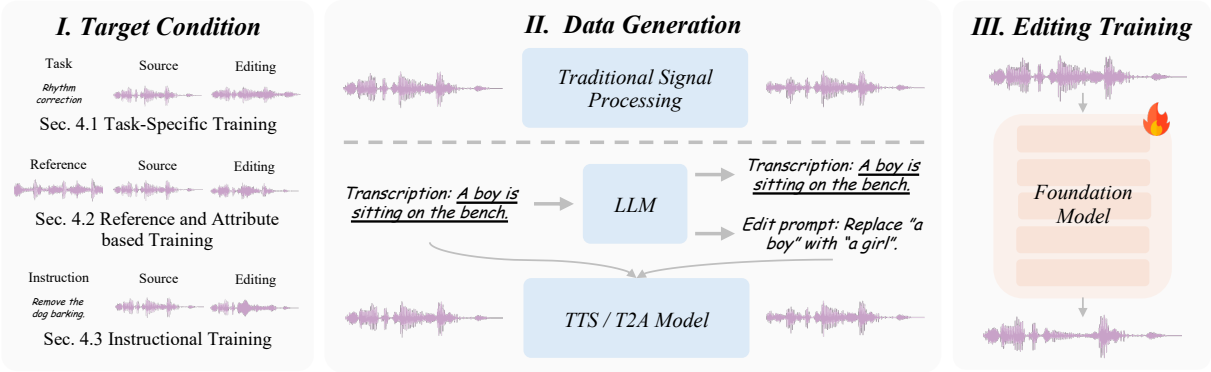

Figure 2: Overview of training-based audio editing methods.

tion, language queries, or source labels.

In speech editing, early methods focus on localized reconstruction and preservation, using partial inference, bidirectional fusion, or alignment-aware acoustic-text representations to reduce artifacts in edited and unmodified regions (Tan et al., 2021; Bai et al., 2022). Later studies further address context and boundary consistency through in-context prosody modeling and attention-based boundary stitching (Yu et al., 2023; Alexos and Baldi, 2024).

Beyond speech, task-specific training supports instance-level manipulation by isolating queried sounds or musical sources from mixtures through language-conditioned or dedicated separation models (Liu et al., 2022b, 2024c; Rouard et al., 2023).

This paradigm is reliable for well-defined edits, but its flexibility is often limited by the task and data distributions used in training.

4.2 Reference and Attribute based Training

Reference and attribute-based training specifies editing directions through external conditions, such as reference audio, style examples, or attribute labels. It typically relies on paired, weakly paired, or contrastive data linking input audio to target conditions. Compared with task-specific training, its conditions are more flexible and broadly applicable, covering references, categorical labels, descriptive attributes, or relative attribute differences.

Reference-based Training uses external audio examples to specify target styles, such as production characteristics, speaker timbre, accent, or emotion. In audio production, related methods, such as DeepAFx-ST, transfer reference production and mixing characteristics through effect-parameter prediction or style feature extraction (Steinmetz et al., 2022; Koo et al., 2023). In speech editing, reference speech guides timbre transfer and real-time voice style conversion (Li et al., 2024a; Liu et al., 2026). This paradigm lets users specify outputs through examples rather than manual parameters.

Attribute-based Training uses explicit labels, descriptors, or attribute degrees to specify target perceptual factors, such as speaker attributes, emotion, or speaking style. Early methods manipulate voice attributes by disentangling linguistic content from extra-linguistic factors (Benaroya et al., 2021). In recent works, taking emotion editing as an example, (Kreuk et al., 2022a) use target emotion as the condition, while (Du et al., 2021) disentangle emotional style from speaker identity to enable independent emotion transfer. EmoCampNet (Wang et al., 2024a) further incorporates emotion attributes into text-based speech editing to generate emotion-controlled edited regions. This paradigm supports controllable manipulation of interpretable perceptual factors.

4.3 Instructional Training

Instructional training supervises audio editing models with $(\texttt{instruction}, \texttt{input}, \texttt{output})$ triplets, where natural-language requests specify edit operations, target regions, styles, or preservation requirements. Compared with reference or attribute based training, it provides more flexible and compositional control through natural language.

Representative works show a progression from template-based instruction following to more open-ended and compositional audio editing. (Wang et al., 2023b) introduce instructional latent-diffusion editing for operations such as addition, removal, replacement, inpainting, and super-resolution. Later datasets and models, including SAO-Instruct (Ungersböck et al., 2026) and MMEdit (Tao et al., 2025), extend this paradigm to free-form instructions and broader operations such as reordering and attribute modification. This framework has also expanded to speech and

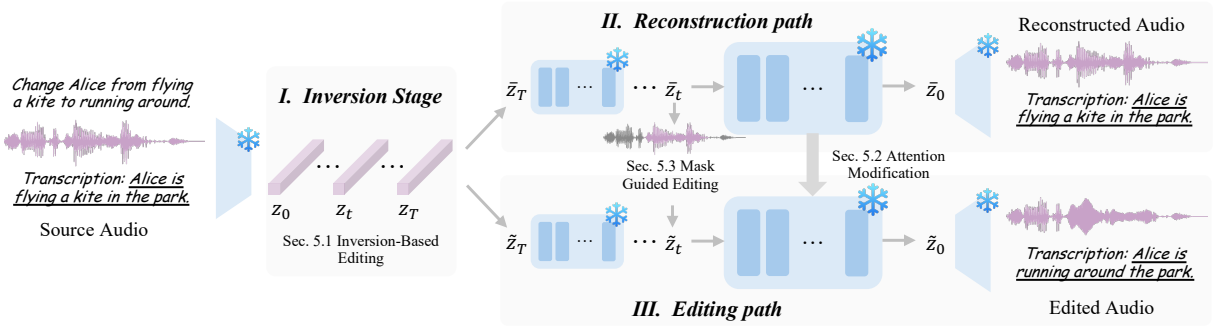

Figure 3: Overview of training-free audio editing methods.

music, where instruction-following models support natural-language control over speech content, acoustic attributes (Huang et al., 2024; Yan et al., 2025a), as well as music editing and remixing (Yan et al., 2025b; Han et al., 2023; Zhang et al., 2024a). Recent systems such as AudioChat (Chen et al., 2026b) further suggest a shift toward multi-task, multi-source, and interactive editing.

By using language as a unified control interface, this paradigm moves audio editing from predefined task execution toward open-ended intent following in foundation-model-based editing.

5 Training-Free Approaches

Training-free approaches adapt pretrained audio generative models to editing without parameter updates, mainly through inference-time inversion, attention, guidance, or mask constraints. As shown in Figure 3, we group them into three categories that are often combined to improve localization, preservation, and controllability. Since token-based autoregressive models are less suited to training-free editing, this section focuses on non-autoregressive paradigms, especially diffusion-based models.

5.1 Inversion-Based Editing

Inversion-based editing maps source audio into the latent, noise, or trajectory space of a pretrained generative model, enabling edits by modifying conditions or sampling paths. It is particularly suited to diffusion and flow-matching models, where audio can be reconstructed through diffusion, latent, or ODE-based inversion (Mokady et al., 2023).

Pretrained latent diffusion models first show the potential of zero-shot audio operations, such as text-guided style transfer, inpainting, and super-resolution (Liu et al., 2023a). Inversion-based methods in DDPM further formulate training-free audio editing by inverting source audio into diffusion trajectories for text-guided or unsupervised editing (Manor and Michaeli, 2024). Subsequent methods improve preservation and editability by refining inversion, suppressing EOT artifacts, relaxing prompt constraints, or balancing source preservation with prompt adherence (Jia et al., 2025a; Paissan et al., 2023). This paradigm also extends to music and speech, where regularized, disentangled, or ODE-based inversion supports text-guided music editing, source-preserving recomposition, and controllable speech editing (Lan et al., 2024a; Liu et al., 2024b; Lv et al., 2026; Tae et al., 2021).

Inversion-based methods reuse pretrained generators, but their performance depends on inversion accuracy, as reconstruction errors can weaken preservation or alter non-target regions.

5.2 Attention Modification

Attention modification guides pretrained generative models by modifying or reusing internal attention patterns, often together with inversion, to control how textual conditions, audio events, and structural patterns are edited or preserved.

Representative methods mainly rely on cross-attention control and self-attention preservation. Cross-attention methods utilize text-audio correspondences to localize or modify target events, using attention maps, word replacement, or prompt attention operations to reflect textual edits in latent or token representations (Xu et al., 2024; Zhang et al., 2024b; Sioros et al., 2025). Self-attention methods instead reuse source-side attention features to maintain musical or acoustic structure, with recent extensions to ReFlow (Liu et al., 2023b), Diffusion Transformer (Peebles and Xie, 2023), and localized temporal editing settings (Liu et al., 2024b; Boudaghi and Zare, 2025; Jiang et al., 2025a).

Attention modification improves localization, text-event alignment, and structural preservation, but depends on whether attention maps reflect perceptual events or musical structures.

Table 2: Overview of publicly available datasets related to audio editing. Row colors indicate speech, music, and general audio datasets. **Dur** denotes total hours, and **Diversity** denotes speakers, instruments, or event classes. **Paired**, **Timestamp**, and **Caption** indicate (source, target) pairs, temporal alignment, and annotations, respectively.

Name	Dur (hour)	Items	Diversity	Paired	Timestamp	Caption
LibriSpeech-Edit (Lv et al., 2026)	3.6	2000	40	✓	✓	✓
Emilia (He et al., 2025)	101654	–	–	×	✓	✓
WenetSpeech4TTS (Ma et al., 2024a)	12800	6283187	–	×	✓	✓
SeniorTalk (Chen et al., 2025b)	55.53	101	202	×	✓	✓
AliMeeting (Yu et al., 2022)	118.75	240	489	×	✓	✓
Slakh2100 (Manilow et al., 2019)	145	2100	34	✓	✓	✓
MUSDB18 (Rafii et al., 2017)	∼10	150	4	✓	×	×
MAESTRO (Hawthorne et al., 2018)	198.7	1276	1	✓	✓	✓
MusicCaps (Agostinelli et al., 2023)	15.34	5521	15	×	×	✓
MTG-Jamendo (Bogdanov et al., 2019)	3769.88	55609	20	×	×	×
NSynth (Engel et al., 2017)	339.98	305979	1006	×	✓	✓
GTSinger (Zhang et al., 2024c)	80.59	1366	20	✓	✓	×
AudioSet (Gemmeke et al., 2017)	5789.78	2084320	527	×	×	×
FSD50K (Fonseca et al., 2021)	108.3	51197	200	×	×	×
ESC-50 (Piczak, 2015)	2.8	2000	50	×	×	×
AudioCaps (Kim et al., 2019)	127.8	46000	527	×	×	✓
Clotho (Drossos et al., 2020)	∼43.59	6974	7853	×	×	✓
WavCaps (Mei et al., 2024)	7565.2	403050	–	×	×	✓
VoiceBank+DEMAND (Valentini-Botinhao, 2017)	∼7.1	12396	30	✓	×	✓
WHAM! (Wichern et al., 2019)	81.68	28000	117	✓	×	×
DNS Challenge (Dubey et al., 2024)	941.53	1560954	6360	✓	×	✓

5.3 Mask Guided Editing

Mask guided editing specifies where to edit or preserve source audio through masks defined in waveform, latent, or source component spaces. This makes it suitable for localized editing, inpainting, restoration, and source-level manipulation.

Representative methods differ mainly in mask design. Explicit time frequency or segment-level masks support localized modification, inpainting, and restoration (Liang et al., 2025; Moliner and Välimäki, 2023), while optimized masks enable source separation based manipulation (Lee et al., 2025). Latent channel and similarity guided masks further constrain which representations or regions are rewritten, supporting timbre transfer and music completion (Lee et al., 2026; Turland et al., 2025).

Mask guided methods offer explicit context or temporal control, but their performance depends on mask quality; inaccurate masks may cause boundary artifacts, incomplete edits, or unintended changes in preserved regions.

6 Resources

We review the resources supporting audio editing in the foundation model era, covering training data and evaluation protocols. And we further discuss scalable data construction tools in Appendix D.

6.1 Open-Source Datasets

Open-source audio datasets are essential for training and adapting audio foundation models. As summarized in Table 2, most are not direct editing datasets but provide transferable supervision from speech, singing, music, audio effects, and audio-text pairs for scalable task construction.

Speech Datasets Speech resources support content editing, speaker-preserving reconstruction, style control, and long-context speech modification. As shown in Table 2, they range from direct editing benchmarks such as LibriSpeech-Edit (Lv et al., 2026), to large-scale generation resources such as Emilia (He et al., 2025) and WenetSpeech (Ma et al., 2024a), and dialogue or meeting corpora for conversational, multi-speaker, and long-context scenarios (Lee et al., 2023; Chen et al., 2025b; Kraaij et al., 2005; Yu et al., 2022).

Musical Datasets Music editing relies on resources with stems, symbolic structure, lyrics, tags, or semantic descriptions. As shown in Table 2, these resources cover stem and score-based manipulation (Manilow et al., 2019; Rafii et al., 2017; Hawthorne et al., 2018), semantic control through captions (Agostinelli et al., 2023; Bogdanov et al., 2019), and timbre, note-level, lyrics, melody, or expressive singing-style editing (Zhang et al., 2024c).

General audio datasets General audio resources support sound-event editing, audio-language grounding, and acoustic restoration. As shown in Table 2, they cover event and scene annotations (Gemmeke et al., 2017; Fonseca et al., 2021; Piczak, 2015), caption-based audio-language supervision (Kim et al., 2019; Drossos et al., 2020; Mei

et al., 2024; Bai et al., 2025), and noisy, reverberant, or real-noise mixtures for denoising and restoration (Wichern et al., 2019; Dubey et al., 2024).

Despite these resources, large-scale instruction-based audio editing datasets remain scarce. Existing datasets provide only partial supervision and rarely pair input audio, edit instructions, target regions, and edited outputs in a unified format. Thus, current systems still depend on synthetic pairs or adapted resources, while standardized instruction-aligned datasets remain a key bottleneck.

6.2 Evaluation Protocols and Metrics

Compared with audio generation, audio editing evaluation must assess both edit success and preservation of non-target regions, in addition to output quality. We sum up existing protocols along four dimensions: edit success and instruction adherence, preservation and locality, temporal and structural consistency, and audio quality and naturalness.

Instruction Adherence Instruction adherence evaluates whether the edited audio follows the intended textual or task-specific requirement. In speech editing, it includes linguistic correctness and attribute-control accuracy: ASR-based WER/CER (Cao et al., 2023) quantifies mismatch with the target transcript, whereas emotion representations from emotion2vec (Ma et al., 2024c) support affective evaluation. For general audio, CLAP-based audio-text similarity (Elizalde et al., 2023) is widely used as a semantic proxy for instruction matching. Music editing typically decomposes adherence into task-specific criteria, including tag prediction, event classification, pitch accuracy, and rhythm errors (Copet et al., 2023).

Preservation and Locality Preservation and locality quantify whether content outside the intended edit scope remains unchanged. Speaker identity is commonly assessed with speaker-embedding cosine similarity (Snyder et al., 2018), while non-target acoustics can be compared at the waveform, spectrogram, or embedding level (Wang et al., 2020; Ragano et al., 2024). Restoration-oriented tasks often adopt reference-based metrics such as PESQ (Rix et al., 2001), STOI (Taal et al., 2010), and SI-SDR (Le Roux et al., 2019) to evaluate perceptual quality, intelligibility, and distortion relative to clean references. More generally, locality-aware evaluation should penalize both under- and over-editing within the specified time span, speaker, event, or source component.

Temporal and structural consistency Temporal and structural consistency examines whether the edited signal remains coherent over time and preserves relevant organization. Boundary accuracy is commonly quantified by the mean or median error between manual and automatic boundaries in speech editing (Williams et al., 2024); WDTW (Lv et al., 2026) further measures word-level temporal fidelity between edited and target utterances. For music editing, structure-aware metrics are particularly important: melody accuracy (Wu et al., 2024) evaluates whether the output follows the target contour, Rhythm F1 (Lan et al., 2024b) measures rhythmic alignment, and dynamics correlation (Sioros et al., 2025) captures consistency between loudness trajectories and generated dynamics. These metrics are useful for edits that must keep prosody, rhythm, melody, or long-range arrangement.

Audio quality and naturalness Audio quality and naturalness evaluate whether the edited audio sounds realistic, continuous, and free from obvious artifacts or synthetic discontinuities. MOS listening tests (Rec, 2006) remain the most direct protocol, while MOSNet (Lo et al., 2019), DNSMOS (Reddy et al., 2021), and NISQA (Mittag et al., 2021) provide scalable automatic estimates. Recent edit-aware evaluators, including AuditScore, AuditEval (Jia et al., 2025b), TTA-Bench (Wang et al., 2026), AudioEval (Wang et al., 2025c), T2A-Feedback (Wang et al., 2025e), and MLLM-based judges (Pu et al., 2025), better connect perceptual quality with instruction relevance and source fidelity. In music, FAD (Kilgour et al., 2018) estimates distributional distance to reference audio, while MuseCPBench (Vishe et al., 2025) targets structural preservation. Given the limits of automatic proxies, human evaluation remains essential for subjective style control, instruction following, and complex audio edits.

7 Conclusion

This survey provides a structured review of audio editing in the foundation model era, connecting task formulation with foundational architectures, adaptation strategies, and supporting resources. Our analysis shows that foundation models have greatly improved editing controllability and flexibility, while future research can further advance precise local editing, source preservation, disentangled control, and reliable evaluation. We hope this survey can support future research toward more general and practical audio editing systems.

Limitations

This survey focuses on audio editing in the era of foundation models, where recent progress is largely driven by large-scale generative backbones, audio-language models, codec language models, diffusion models, and flow-matching frameworks. **Consequently, we mainly review foundation-model-based audio editing methods**. Earlier paradigms, including statistical parametric methods, signal-processing-based workflows, and early task-specific deep learning approaches, are discussed only as background rather than surveyed in detail. In addition, this survey primarily considers monaural audio editing. Spatial audio editing, such as stereo, binaural, or multi-channel Ambisonics-based editing, involves specialized issues of spatial rendering, room acoustics, and perceptual localization, and is therefore beyond the main scope of this review.

Ethical Considerations

Although this survey itself raises no immediate ethical concerns, the reviewed audio editing methods may introduce risks when applied in practice. We highlight two aspects. (1) **Data licensing and consent.** Users should respect dataset licenses and obtain proper permission before collecting or repurposing web-hosted speech, music, or sound recordings, especially when they involve personal voices, singing performances, copyrighted music, or private acoustic scenes. (2) **Misuse of editing models.** Foundation-model-based audio editing can realistically alter speech content, speaker identity, emotion, background sounds, and music. Without safeguards, it may enable voice impersonation, unauthorized modification, copyright infringement, or deceptive media generation. Practitioners should follow usage policies and regulations, while future work should explore provenance tracking, watermarking, and edited-audio detection.

References

Sharath Adavanne, Archontis Politis, Joonas Nikunen, and Tuomas Virtanen. 2018. Sound event localization and detection of overlapping sources using convolutional recurrent neural networks. *IEEE Journal of Selected Topics in Signal Processing*, 13(1):34–48.

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

Antonios Alexos and Pierre Baldi. 2024. Attentionstitch: How attention solves the speech editing problem. *arXiv preprint arXiv:2403.04804*.

Philip Anastassiou, Zhenyu Tang, Kainan Peng, Dongya Jia, Jiaxin Li, Ming Tu, Yuping Wang, Yuxuan Wang, and Mingbo Ma. 2024. Voiceshop: A unified speech-to-speech framework for identity-preserving zero-shot voice editing. *arXiv preprint arXiv:2404.06674*.

Louis Bahrman, Mathieu Fontaine, Jonathan Le Roux, and Gaël Richard. 2024. Speech dereverberation constrained on room impulse response characteristics. *arXiv preprint arXiv:2407.08657*.

He Bai, Renjie Zheng, Junkun Chen, Mingbo Ma, Xintong Li, and Liang Huang. 2022. A3 t: Alignment-aware acoustic and text pretraining for speech synthesis and editing. In *International Conference on Machine Learning*, pages 1399–1411. PMLR.

Jisheng Bai, Haohe Liu, Mou Wang, Dongyuan Shi, Wenwu Wang, Mark D Plumbley, Woon-Seng Gan, and Jianfeng Chen. 2025. Audiosetcaps: An enriched audio-caption dataset using automated generation pipeline with large audio and language models. *IEEE Transactions on Audio, Speech and Language Processing*.

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

Laurent Benaroya, Nicolas Obin, and Axel Roebel. 2021. Beyond voice identity conversion: manipulating voice attributes by adversarial learning of structured disentangled representations. *arXiv preprint arXiv:2107.12346*.

Paul Boersma and David Weenink. 2021. Praat: Doing phonetics by computer (6.1. 16)[computer software].

Dmitry Bogdanov, Minz Won, Philip Tovstogan, Alastair Porter, and Xavier Serra. 2019. The mtg-jamendo dataset for automatic music tagging. In *Machine learning for music discovery workshop, international conference on machine learning (ICML 2019)*, pages 1–3. Long Beach, CA, United States.

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, et al. 2023a. Audiolm: a language modeling approach to audio generation. *IEEE/ACM transactions on audio, speech, and language processing*, 31:2523–2533.

Zalán Borsos, Matt Sharifi, Damien Vincent, Eugene Kharitonov, Neil Zeghidour, and Marco Tagliasacchi. 2023b. Soundstorm: Efficient parallel audio generation. *arXiv preprint arXiv:2305.09636*.

Ali Boudaghi and Hadi Zare. 2025. Musrec: Zero-shot text-to-music editing via rectified flow and diffusion transformers. *arXiv preprint arXiv:2511.04376*.

Matej Božić and Marko Horvat. 2024. A survey of deep learning audio generation methods. *arXiv preprint arXiv:2406.00146*.

Hervé Bredin. 2023. pyannote. audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe. In *24th Interspeech Conference (INTERSPEECH 2023)*, pages 1983–1987. ISCA.

Jie Cao, Ananya Ganesh, Jon Cai, Rosy Southwell, E Margaret Perkoff, Michael Regan, Katharina Kann, James H Martin, Martha Palmer, and Sidney D'Mello. 2023. A comparative analysis of automatic speech recognition errors in small group classroom discourse. In *Proceedings of the 31st ACM conference on user modeling, adaptation and personalization*, pages 250–262.

Edresson Casanova, Julian Weber, Christopher D Shulby, Arnaldo Candido Junior, Eren Gölge, and Moacir A Ponti. 2022. Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone. In *International Conference on Machine Learning*, pages 2709–2720. PMLR.

Changan Chen, Ruohan Gao, Paul Calamia, and Kristen Grauman. 2022a. Visual acoustic matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18858–18868.

Junyang Chen, Yuhang Jia, Hui Wang, Jiaming Zhou, Yaxin Han, Mengying Feng, and Yong Qin. 2026a. Cosyedit: Unlocking end-to-end speech editing capability from zero-shot text-to-speech models. *arXiv preprint arXiv:2601.05329*.

Ke Chen, Xingjian Du, Bilei Zhu, Zejun Ma, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. 2022b. Htsat: A hierarchical token-semantic audio transformer for sound classification and detection. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 646–650. IEEE.

Kuan-Yu Chen, Jeng-Lin Li, and Jian-Jiun Ding. 2025a. Seamlessedit: Background noise aware zero-shot speech editing with in-context enhancement. *arXiv preprint arXiv:2505.14066*.

Nanxin Chen, Yu Zhang, Heiga Zen, Ron J Weiss, Mohammad Norouzi, and William Chan. 2020. Wavegrad: Estimating gradients for waveform generation. *arXiv preprint arXiv:2009.00713*.

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

Yang Chen, Yuhang Jia, Shiwan Zhao, Ziyue Jiang, Haoran Li, Jiarong Kang, and Yong Qin. 2024. Diffeditor: Enhancing speech editing with semantic enrichment and acoustic consistency. *arXiv preprint arXiv:2409.12992*.

Yang Chen, Hui Wang, Shiyao Wang, Junyang Chen, Jiabei He, Jiaming Zhou, Xi Yang, Yequan Wang, Yonghua Lin, and Yong Qin. 2025b. Seniortalk: A chinese conversation dataset with rich annotations for super-aged seniors. *arXiv preprint arXiv:2503.16578*.

Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, JianZhao JianZhao, Kai Yu, and Xie Chen. 2025c. 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.

Woosung Choi, Minseok Kim, Marco A Martínez Ramírez, Jaehwa Chung, and Soonyoung Jung. 2021. Amss-net: Audio manipulation on user-specified sources with textual queries. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1775–1783.

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

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

DeepSeek-AI. 2026. Deepseek-v4: Towards highly efficient million-token context intelligence. Technical report. Accessed: 2026-05-01.

Alexandre Défossez. 2021. Hybrid spectrogram and waveform source separation. *arXiv preprint arXiv:2111.03600*.

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

Zhihao Du, Changfeng Gao, Yuxuan Wang, Fan Yu, Tianyu Zhao, Hao Wang, Xiang Lv, Hui Wang, Chongjia Ni, Xian Shi, et al. 2025. Cosyvoice 3: Towards in-the-wild speech generation via scaling-up and post-training. *arXiv preprint arXiv:2505.17589*.

Zongyang Du, Berrak Sisman, Kun Zhou, and Haizhou Li. 2021. Disentanglement of emotional style and speaker identity for expressive voice conversion. *arXiv preprint arXiv:2110.10326*.

Harishchandra Dubey, Ashkan Aazami, Vishak Gopal, Babak Naderi, Sebastian Braun, Ross Cutler, Alex Ju, Mehdi Zohourian, Min Tang, Mehrsa Golestaneh, et al. 2024. Icassp 2023 deep noise suppression challenge. *IEEE Open Journal of Signal Processing*, 5:725–737.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. 2023. Clap learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Daniel PW Ellis, Eduardo Fonseca, Ron J Weiss, Kevin Wilson, Scott Wisdom, Hakan Erdogan, John R Hershey, Aren Jansen, R Channing Moore, and Manoj Plakal. 2025. Recomposer: Event-roll-guided generative audio editing. *arXiv preprint arXiv:2509.05256*.

Jesse Engel, Cinjon Resnick, Adam Roberts, Sander Dieleman, Mohammad Norouzi, Douglas Eck, and Karen Simonyan. 2017. Neural audio synthesis of musical notes with wavenet autoencoders. In *International conference on machine learning*, pages 1068–1077. PMLR.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*.

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

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

Szu-Wei Fu, Cheng Yu, Tsun-An Hsieh, Peter Plantinga, Mirco Ravanelli, Xugang Lu, and Yu Tsao. 2021. Metricgan+: An improved version of metricgan for speech enhancement. *arXiv preprint arXiv:2104.03538*.

Youquan Fu, Ruiyang Si, Hongfa Wang, Dongzhan Zhou, Jiacheng Sun, Ping Luo, Di Hu, Hongyuan Zhang, and Xuelong Li. 2025. Object-avedit: An object-level audio-visual editing model. *arXiv preprint arXiv:2510.00050*.

Liting Gao, Yi Yuan, Yaru Chen, Yuelan Cheng, Zhenbo Li, Juan Wen, Shubin Zhang, and Wenwu Wang. 2026. Rfm-editing: Rectified flow matching for text-guided audio editing. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 15467–15471. IEEE.

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

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.

Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27.

Karim Guirguis, Christoph Schorn, Andre Guntoro, Sherif Abdulatif, and Bin Yang. 2021. Seld-tcn: Sound event localization & detection via temporal convolutional networks. In *2020 28th European Signal Processing Conference (EUSIPCO)*, pages 16–20. IEEE.

Bing Han, Junyu Dai, Weituo Hao, Xinyan He, Dong Guo, Jitong Chen, Yuxuan Wang, Yanmin Qian, and Xuchen Song. 2023. Instructme: An instruction guided music edit and remix framework with latent diffusion models. *arXiv preprint arXiv:2308.14360*.

Curtis Hawthorne, Andriy Stasyuk, Adam Roberts, Ian Simon, Cheng-Zhi Anna Huang, Sander Dieleman, Erich Elsen, Jesse Engel, and Douglas Eck. 2018. Enabling factorized piano music modeling and generation with the maestro dataset. *arXiv preprint arXiv:1810.12247*.

Haorui He, Zengqiang Shang, Chaoren Wang, Xuyuan Li, Yicheng Gu, Hua Hua, Liwei Liu, Chen Yang, Jiaqi Li, Peiyang Shi, et al. 2025. Emilia: A large-scale, extensive, multilingual, and diverse dataset for speech generation. *IEEE Transactions on Audio, Speech and Language Processing*.

Romain Hennequin, Anis Khlif, Felix Voituret, and Manuel Moussallam. 2020. Spleeter: a fast and efficient music source separation tool with pre-trained models. *Journal of Open Source Software*, 5(50):2154.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851.

Rongjie Huang, Ruofan Hu, Yongqi Wang, Zehan Wang, Xize Cheng, Ziyue Jiang, Zhenhui Ye, Dongchao Yang, Luping Liu, Peng Gao, et al. 2024. Instructspeech: Following speech editing instructions via large language models. In *Forty-first International Conference on Machine Learning*.

Sung-Feng Huang, Heng-Cheng Kuo, Zhehuai Chen, Xuesong Yang, Pin-Jui Ku, Ante Jukic, Huck Yang, Yu Tsao, Yu-Chiang Frank Wang, Hung-yi Lee, et al. 2025. Voicenong: Robust high-quality speech editing model without hallucinations. In *Proc. Interspeech 2025*, pages 3469–3473.

Yannick Jadoul, Bill Thompson, and Bart De Boer. 2018. Introducing parselmouth: A python interface to praat. *Journal of Phonetics*, 71:1–15.

Shengpeng Ji, Yifu Chen, Minghui Fang, Jialong Zuo, Jingyu Lu, Hanting Wang, Ziyue Jiang, Long Zhou, Shujie Liu, Xize Cheng, Xiaoda Yang, Zehan Wang, Qian Yang, Jian Li, Jingzhen Jiang, Yidi ang He, Yunfei Chu, Jin Xu, and Zhou Zhao. 2024. Wavchat: A survey of spoken dialogue models. *arXiv preprint arXiv:2411.13577*.

Yuhang Jia, Yang Chen, Jinghua Zhao, Shiwan Zhao, Wenjia Zeng, Yong Chen, and Yong Qin. 2025a. Audioeditor: A training-free diffusion-based audio editing framework. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Yuhang Jia, Hui Wang, Xin Nie, Yujie Guo, Lianru Gao, and Yong Qin. 2025b. Towards automatic evaluation and high-quality pseudo-parallel dataset construction for audio editing: A human-in-the-loop method. *arXiv preprint arXiv:2508.11966*.

Xilin Jiang, Cong Han, Yinghao Aaron Li, and Nima Mesgarani. 2024. Listen, chat, and remix: Text-guided soundscape remixing for enhanced auditory experience. *arXiv preprint arXiv:2402.03710*.

Yuxuan Jiang, Zehua Chen, Zeqian Ju, Chang Li, Weibei Dou, and Jun Zhu. 2025a. Freeaudio: Training-free timing planning for controllable long-form text-to-audio generation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 9871–9880.

Ziyue Jiang, Yi Ren, Ruiqi Li, Shengpeng Ji, Boyang Zhang, Zhenhui Ye, Chen Zhang, Bai Jionghao, Xiaoda Yang, Jialong Zuo, et al. 2025b. Megatts 3: Sparse alignment enhanced latent diffusion transformer for zero-shot speech synthesis. *arXiv preprint arXiv:2502.18924*.

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.

Mumin Jin, Prashant Serai, Jilong Wu, Andros Tjandra, Vimal Manohar, and Qing He. 2023. Voice-preserving zero-shot multiple accent conversion. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Zeyu Jin, Gautham J Mysore, Stephen Diverdi, Jingwan Lu, and Adam Finkelstein. 2017. Voco: Text-based insertion and replacement in audio narration. *ACM Transactions on Graphics (TOG)*, 36(4):1–13.

Biswajit Karan, Joshua Jansen van Vüren, Febe de Wet, and Thomas Niesler. 2024. A transformer-based voice activity detector. In *Interspeech*.

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

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.

Jong Wook Kim, Justin Salamon, Peter Li, and Juan Pablo Bello. 2018. Crepe: A convolutional representation for pitch estimation. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 161–165. IEEE.

Seung-Bin Kim, Jun-Hyeok Cha, Hyung-Seok Oh, Heejin Choi, and Seong-Whan Lee. 2025. Fillerspeech: Towards human-like text-to-speech synthesis with filler insertion and filler style control. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 34096–34113.

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.

Zhifeng Kong, Arushi Goel, Rohan Badlani, Wei Ping, Rafael Valle, and Bryan Catanzaro. 2024. Audio flamingo: A novel audio language model with few-shot learning and dialogue abilities. In *International Conference on Machine Learning*, pages 25125–25148. PMLR.

Junghyun Koo, Marco A Martínez-Ramírez, Wei-Hsiang Liao, Stefan Uhlich, Kyogu Lee, and Yuki Mitsufuji. 2023. Music mixing style transfer: A contrastive learning approach to disentangle audio effects. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Wessel Kraaij, Thomas Hain, Mike Lincoln, and Wilfried Post. 2005. The ami meeting corpus. In *Proc. International Conference on Methods and Techniques in Behavioral Research*, pages 1–4.

Felix Kreuk, Adam Polyak, Jade Copet, Eugene Kharitonov, Tu Anh Nguyen, Morgan Rivière, Wei-Ning Hsu, Abdelrahman Mohamed, Emmanuel Dupoux, and Yossi Adi. 2022a. Textless speech

emotion conversion using discrete & decomposed representations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11200–11214.

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

Gael Le Lan, Bowen Shi, Zhaoheng Ni, Sidd Srinivasan, Anurag Kumar, Brian Ellis, David Kant, Varun Nagaraja, Ernie Chang, Wei-Ning Hsu, et al. 2024a. High fidelity text-guided music editing via single-stage flow matching. *arXiv preprint arXiv:2407.03648*.

Yun-Han Lan, Wen-Yi Hsiao, Hao-Chung Cheng, and Yi-Hsuan Yang. 2024b. Musicongen: Rhythm and chord control for transformer-based text-to-music generation. *arXiv preprint arXiv:2407.15060*.

Siddique Latif, Moazzam Shoukat, Fahad Shamshad, Muhammad Usama, Yi Ren, Heriberto Cuayáhuitl, Wenwu Wang, Xulong Zhang, Roberto Togneri, Erik Cambria, et al. 2023. Sparks of large audio models: A survey and outlook. *arXiv preprint arXiv:2308.12792*.

Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R Hershey. 2019. Sdr–half-baked or well done? In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 626–630. IEEE.

Ching Ho Lee, Javier Nistal, Stefan Lattner, Marco Pasini, and George Fazekas. 2026. Diffusion timbre transfer via mutual information guided inpainting. *arXiv preprint arXiv:2601.01294*.

Geonyoung Lee, Geonhee Han, and Paul Hongsuck Seo. 2025. Dgmo: Training-free audio source separation through diffusion-guided mask optimization. *arXiv preprint arXiv:2506.02858*.

Keon Lee, Kyumin Park, and Daeyoung Kim. 2023. Dailytalk: Spoken dialogue dataset for conversational text-to-speech. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Junjie Li, Yiwei Guo, Xie Chen, and Kai Yu. 2024a. Sef-vc: Speaker embedding free zero-shot voice conversion with cross attention. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12296–12300. IEEE.

Kang Li, Yan Song, Li-Rong Dai, Ian McLoughlin, Xin Fang, and Lin Liu. 2023. Ast-sed: An effective sound event detection method based on audio spectrogram transformer. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Ruiqi Li, Yu Zhang, Yongqi Wang, Zhiqing Hong, Rongjie Huang, and Zhou Zhao. 2024b. Robust singing voice transcription serves synthesis. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9751–9766.

Yinghao Aaron Li, Cong Han, and Nima Mesgarani. 2025. Styletts: A style-based generative model for natural and diverse text-to-speech synthesis. *IEEE Journal of Selected Topics in Signal Processing*, 19(1):283–296.

Jinhua Liang, Yuanzhe Chen, Yi Yuan, Dongya Jia, Xiaobin Zhuang, Zhuo Chen, Yuping Wang, and Yuxuan Wang. 2025. Audiomorphix: Training-free audio editing with diffusion probabilistic models. *arXiv preprint arXiv:2505.16076*.

Jinhua Liang, Huan Zhang, Haohe Liu, Y Cao, Q Kong, X Liu, W Wang, MD Plumbley, H Phan, and E Benetos. 2024. Wavcraft: Audio editing and generation with natural language models. ICLR 2024 Workshop on LLM Agents.

Gaojie Lin, Jianwen Jiang, Jiaqi Yang, Zerong Zheng, Chao Liang, Yuan Zhang, and Jingtuo Liu. 2025. Omnihuman-1: Rethinking the scaling-up of one-stage conditioned human animation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13847–13858.

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. *arXiv preprint arXiv:2301.12503*.

Haohe Liu, Xubo Liu, Qiuqiang Kong, Qiao Tian, Yan Zhao, DeLiang Wang, Chuanzeng Huang, and Yuxuan Wang. 2022a. Voicefixer: A unified framework for high-fidelity speech restoration. *arXiv preprint arXiv:2204.05841*.

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

Huadai Liu, Jialei Wang, Xiangtai Li, Wen Wang, Qian Chen, Rongjie Huang, Yang Liu, Jiayang Xu, and Zhou Zhao. 2024b. Medic: Zero-shot music editing with disentangled inversion control. *arXiv preprint arXiv:2407.13220*.

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

Xubo Liu, Qiuqiang Kong, Yan Zhao, Haohe Liu, Yi Yuan, Yuzhuo Liu, Rui Xia, Yuxuan Wang, Mark D Plumbley, and Wenwu Wang. 2024c. Separate anything you describe. *IEEE Transactions on*

Audio, Speech and Language Processing, 33:458–471.

Xubo Liu, Haohe Liu, Qiuqiang Kong, Xinhao Mei, Jinzheng Zhao, Qiushi Huang, Mark D Plumbley, and Wenwu Wang. 2022b. Separate what you describe: Language-queried audio source separation. *arXiv preprint arXiv:2203.15147*.

Yisi Liu, Nicholas Lee, and Gopala Anumanchipalli. 2026. Stylestream: Real-time zero-shot voice style conversion. *arXiv preprint arXiv:2602.20113*.

Chen-Chou Lo, Szu-Wei Fu, Wen-Chin Huang, Xin Wang, Junichi Yamagishi, Yu Tsao, and Hsin-Min Wang. 2019. Mosnet: Deep learning based objective assessment for voice conversion. *arXiv preprint arXiv:1904.08352*.

Sihan Lv, Yechen Jin, Zhen Li, Jintao Chen, Jinshan Zhang, Ying Li, Jianwei Yin, and Meng Xi. 2026. Ast: Adaptive, seamless, and training-free precise speech editing. *arXiv preprint arXiv:2604.16056*.

Linhan Ma, Dake Guo, Kun Song, Yuepeng Jiang, Shuai Wang, Liumeng Xue, Weiming Xu, Huan Zhao, Binbin Zhang, and Lei Xie. 2024a. Wenetspeech4tts: A 12,800-hour mandarin tts corpus for large speech generation model benchmark. *arXiv preprint arXiv:2406.05763*.

Yinghao Ma, Anders Øland, Anton Ragni, Bleiz MacSen Del Sette, Charalampos Saitis, Chris Donahue, Chenghua Lin, Christos Plachouras, Emmanouil Benetos, Elona Shatri, et al. 2024b. Foundation models for music: A survey. *arXiv preprint arXiv:2408.14340*.

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

Ethan Manilow, Gordon Wichern, Prem Seetharaman, and Jonathan Le Roux. 2019. Cutting music source separation some slakh: A dataset to study the impact of training data quality and quantity. In *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 45–49. IEEE.

Hila Manor and Tomer Michaeli. 2024. Zero-shot unsupervised and text-based audio editing using ddpm inversion. *arXiv preprint arXiv:2402.10009*.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal forced aligner: Trainable text-speech alignment using kaldi. In *Interspeech*, volume 2017, pages 498–502.

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.

Gabriel Mittag, Babak Naderi, Assmaa Chehadi, and Sebastian Möller. 2021. Nisqa: A deep cnn-self-attention model for multidimensional speech quality prediction with crowdsourced datasets. *arXiv preprint arXiv:2104.09494*.

Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2023. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6038–6047.

Eloi Moliner and Vesa Välimäki. 2023. Diffusion-based audio inpainting. *arXiv preprint arXiv:2305.15266*.

Andrew Cameron Morris, Viktoria Maier, and Phil D Green. 2004. From wer and ril to mer and wil: improved evaluation measures for connected speech recognition. In *Interspeech*, 4-8, page 2004.

Francesco Paissan, Luca Della Libera, Zhepei Wang, Mirco Ravanelli, Paris Smaragdis, and Cem Subakan. 2023. Audio editing with non-rigid text prompts. *arXiv preprint arXiv:2310.12858*.

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.

Puyuan Peng, Po-Yao Huang, Shang-Wen Li, Abdelrahman Mohamed, and David Harwath. 2024. Voicecraft: Zero-shot speech editing and text-to-speech in the wild. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12442–12462.

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

Shu Pu, Yaochen Wang, Dongping Chen, Yuhang Chen, Guohao Wang, Qi Qin, Zhongyi Zhang, Zhiyuan Zhang, Zetong Zhou, Shuang Gong, et al. 2025. Judge anything: Mllm as a judge across any modality. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 5742–5753.

Kaizhi Qian, Yang Zhang, Shiyu Chang, Xuesong Yang, and Mark Hasegawa-Johnson. 2019. Autovc: Zero-shot voice style transfer with only autoencoder loss. In *International Conference on Machine Learning*, pages 5210–5219. PMLR.

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.

Zafar Rafii, Antoine Liutkus, Fabian-Robert Stöter, Stylianos Ioannis Mimilakis, and Rachel Bittner. 2017. The musdb18 corpus for music separation.

Alessandro Ragano, Jan Skoglund, and Andrew Hines. 2024. Nomad: Unsupervised learning of perceptual embeddings for speech enhancement and non-matching reference audio quality assessment. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1011–1015. IEEE.

ITUT Rec. 2006. P. 800.1, mean opinion score (mos) terminology. *International Telecommunication Union, Geneva*, 2.

Chandan KA Reddy, Vishak Gopal, and Ross Cutler. 2021. Dnsmos: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6493–6497. IEEE.

Antony W Rix, John G Beerends, Michael P Hollier, and Andries P Hekstra. 2001. Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)*, volume 2, pages 749–752. IEEE.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695.

Simon Rouard, Francisco Massa, and Alexandre Défossez. 2023. Hybrid transformers for music source separation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Simon Rouard, Robin San Roman, Yossi Adi, and Axel Roebel. 2025. Musicgen-stem: Multi-stem music generation and edition through autoregressive modeling. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

F Schneider. 2023. Moûsai: Text-to-music generation with long-context latent diffusion [internet]. *arXiv [cs. CL]*.

Team Seedance, De Chen, Liyang Chen, Xin Chen, Ying Chen, Zhuo Chen, Zhuowei Chen, Feng Cheng, Tianheng Cheng, Yufeng Cheng, et al. 2026. Seedance 2.0: Advancing video generation for world complexity. *arXiv preprint arXiv:2604.14148*.

Bowen Shi, Andros Tjandra, John Hoffman, Helin Wang, Yi-Chiao Wu, Luya Gao, Julius Richter, Matt Le, Apoorv Vyas, Sanyuan Chen, et al. 2025. Sam audio: Segment anything in audio. *arXiv preprint arXiv:2512.18099*.

Vassilis Sioros, Alexandros Potamianos, and Giorgos Paraskevopoulos. 2025. Editgen: Harnessing cross-attention control for instruction-based auto-regressive audio editing. *arXiv preprint arXiv:2507.11096*.

David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. 2018. X-vectors: Robust dnn embeddings for speaker recognition. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5329–5333. IEEE.

Christian J Steinmetz, Nicholas J Bryan, and Joshua D Reiss. 2022. Style transfer of audio effects with differentiable signal processing. *arXiv preprint arXiv:2207.08759*.

Fabian-Robert Stöter, Stefan Uhlich, Antoine Liutkus, and Yuki Mitsufuji. 2019. Open-unmix-a reference implementation for music source separation. *Journal of Open Source Software*, 4(41):1667.

Giorgio Strano, Chiara Ballanti, Donato Crisostomi, Michele Mancusi, Luca Cosmo, and Emanuele Rodolà. 2025. Stage: Stemmed accompaniment generation through prefix-based conditioning. *arXiv preprint arXiv:2504.05690*.

Yi Su, Jisheng Bai, Qisheng Xu, Kele Xu, and Yong Dou. 2025. Audio-language models for audio-centric tasks: A survey. *arXiv preprint arXiv:2501.15177*.

Cees H Taal, Richard C Hendriks, Richard Heusdens, and Jesper Jensen. 2010. A short-time objective intelligibility measure for time-frequency weighted noisy speech. In *2010 IEEE international conference on acoustics, speech and signal processing*, pages 4214–4217. IEEE.

Jaesung Tae, Hyeongju Kim, and Taesu Kim. 2021. Editts: Score-based editing for controllable text-to-speech. *arXiv preprint arXiv:2110.02584*.

Daxin Tan, Liqun Deng, Yu Ting Yeung, Xin Jiang, Xiao Chen, and Tan Lee. 2021. Editspeech: A text based speech editing system using partial inference and bidirectional fusion. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 626–633. IEEE.

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*.

Fei Tian, Xiangyu Tony Zhang, Yuxin Zhang, Haoyang Zhang, Yuxin Li, Daijiao Liu, Yayue Deng, Donghang Wu, Jun Chen, Liang Zhao, et al. 2025. Step-audio-r1 technical report. *arXiv preprint arXiv:2511.15848*.

Sean Turland, Eloi Moliner, and Vesa Välimäki. 2025. Similarity-guided diffusion for long-gap music inpainting. *arXiv preprint arXiv:2509.16342*.

Michael Ungersböck, Florian Grötschla, Luca Lanzendörfer, June Young Yi, Changho Choi, and Roger Wattenhofer. 2026. Sao-instruct: Free-form audio editing using natural language instructions. *Advances in Neural Information Processing Systems*, 38:83411–83437.

Cassia Valentini-Botinhao. 2017. Noisy speech database for training speech enhancement algorithms and TTS models. Clean and noisy parallel speech database.

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

Satvik Venkatesh, David Moffat, and Eduardo Reck Miranda. 2022. Word embeddings for automatic equalization in audio mixing. *arXiv preprint arXiv:2202.08898*.

Yash Vishe, Eric Xue, Xunyi Jiang, Zachary Novack, Junda Wu, Julian McAuley, and Xin Xu. 2025. Musecpbench: an empirical study of music editing methods through music context preservation. *arXiv preprint arXiv:2512.14629*.

Bin Wang, Xunlong Zou, Geyu Lin, Shuo Sun, Zhuohan Liu, Wenyu Zhang, Zhengyuan Liu, AiTi Aw, and Nancy Chen. 2025a. Audiobench: A universal benchmark for audio large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4297–4316.

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

Disong Wang, Liqun Deng, Yu Ting Yeung, Xiao Chen, Xunying Liu, and Helen Meng. 2021. Vqmivc: Vector quantization and mutual information-based unsupervised speech representation disentanglement for one-shot voice conversion. *arXiv preprint arXiv:2106.10132*.

Helin Wang, Meng Yu, Jiarui Hai, Chen Chen, Yuchen Hu, Rilin Chen, Najim Dehak, and Dong Yu. 2025b. Ssr-speech: Towards stable, safe and robust zero-shot text-based speech editing and synthesis. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

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

Hui Wang, Jinghua Zhao, Junyang Cheng, Cheng Liu, Yuhang Jia, Haoqin Sun, Jiaming Zhou, and Yong Qin. 2025c. Audioeval: Automatic dual-perspective and multi-dimensional evaluation of text-to-audio-generation. *arXiv preprint arXiv:2510.14570*.

Tao Wang, Jiangyan Yi, Ruibo Fu, Jianhua Tao, Zhengqi Wen, and Chu Yuan Zhang. 2024a. Emotion selectable end-to-end text-based speech editing. *Artificial Intelligence*, 329:104076.

Xianke Wang, Wei Xu, Weiming Yang, and Wenqing Cheng. 2022. Musicyolo: A sight-singing onset/offset detection framework based on object detection instead of spectrum frames. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 396–400. IEEE.

Xiaofei Wang, Manthan Thakker, Zhuo Chen, Naoyuki Kanda, Sefik Emre Eskimez, Sanyuan Chen, Min Tang, Shujie Liu, Jinyu Li, and Takuya Yoshioka. 2024b. Speechx: Neural codec language model as a versatile speech transformer. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:3355–3364.

Yuancheng Wang, Zeqian Ju, Xu Tan, Lei He, Zhizheng Wu, Jiang Bian, et al. 2023b. Audit: Audio editing by following instructions with latent diffusion models. *Advances in Neural Information Processing Systems*, 36:71340–71357.

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

Zehan Wang, Ke Lei, Chen Zhu, Jiawei Huang, Sashuai Zhou, Luping Liu, Xize Cheng, Shengpeng Ji, Zhenhui Ye, Tao Jin, et al. 2025e. T2a-feedback: Improving basic capabilities of text-to-audio generation via fine-grained ai feedback. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 23535–23547.

Zhong-Qiu Wang, Peidong Wang, and DeLiang Wang. 2020. Complex spectral mapping for single-and multi-channel speech enhancement and robust asr. *IEEE/ACM transactions on audio, speech, and language processing*, 28:1778–1787.

Haojie Wei, Xueke Cao, Tangpeng Dan, and Yueguo Chen. 2023. Rmvpe: A robust model for vocal pitch estimation in polyphonic music. *arXiv preprint arXiv:2306.15412*.

Gordon Wichern, Joe Antognini, Michael Flynn, Licheng Richard Zhu, Emmett McQuinn, Dwight Crow, Ethan Manilow, and Jonathan Le Roux. 2019. Wham!: Extending speech separation to noisy environments. *arXiv preprint arXiv:1907.01160*.

Samantha Williams, Paul Foulkes, and Vincent Hughes. 2024. Analysis of forced aligner performance on l2 english speech. *Speech Communication*, 158:103042.

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

Tianxin Xie, Yan Rong, Pengfei Zhang, Wenwu Wang, and Li Liu. 2025. Towards controllable speech synthesis in the era of large language models: A systematic survey. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 764–791.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. 2025. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*.

Manjie Xu, Chenxing Li, Dan Su, Wei Liang, Dong Yu, et al. 2024. Prompt-guided precise audio editing with diffusion models. *arXiv preprint arXiv:2406.04350*.

Canxiang Yan, Chunxiang Jin, Dawei Huang, Haibing Yu, Han Peng, Hui Zhan, Jie Gao, Jing Peng, Jingdong Chen, Jun Zhou, et al. 2025a. Ming-uniaudio: Speech llm for joint understanding, generation and editing with unified representation. *arXiv preprint arXiv:2511.05516*.

Chao Yan, Boyong Wu, Peng Yang, Pengfei Tan, Guoqiang Hu, Li Xie, Yuxin Zhang, Fei Tian, Xuerui Yang, Xiangyu Zhang, et al. 2025b. Step-audio-editx technical report. *arXiv preprint arXiv:2511.03601*.

Dongchao Yang, Jinchuan Tian, Xu Tan, Rongjie Huang, Songxiang Liu, Xuankai Chang, Jiatong Shi, Sheng Zhao, Jiang Bian, Xixin Wu, et al. 2023. Uniaudio: An audio foundation model toward universal audio generation. *arXiv preprint arXiv:2310.00704*.

Muqiao Yang, Chunlei Zhang, Yong Xu, Zhongweiyang Xu, Heming Wang, Bhiksha Raj, and Dong Yu. 2024. usee: Unified speech enhancement and editing with conditional diffusion models. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7125–7129. IEEE.

Cheng Yu, Yang Li, Weiqin Zu, Fanglei Sun, Zheng Tian, and Jun Wang. 2023. Cross-utterance conditioned coherent speech editing. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, volume 2023, pages 2108–2112. ISCA.

Fan Yu, Shiliang Zhang, Yihui Fu, Lei Xie, Siqi Zheng, Zhihao Du, Weilong Huang, Pengcheng Guo, Zhijie Yan, Bin Ma, et al. 2022. M2met: The icassp 2022 multi-channel multi-party meeting transcription challenge. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6167–6171. IEEE.

Ruibin Yuan, Yinghao Ma, Yizhi Li, Ge Zhang, Xingran Chen, Hanzhi Yin, Yiqi Liu, Jiawen Huang, Zeyue Tian, Binyue Deng, et al. 2023. Marble: Music audio representation benchmark for universal evaluation. *Advances in Neural Information Processing Systems*, 36:39626–39647.

Lichao Zhang, Zhou Zhao, Yi Ren, and Liqun Deng. 2022. Editsinger: Zero-shot text-based singing voice editing system with diverse prosody modeling. In *IJCAI*, pages 4503–4509.

Yixiao Zhang, Yukara Ikemiya, Woosung Choi, Naoki Murata, Marco A Martínez-Ramírez, Liwei Lin, Gus Xia, Wei-Hsiang Liao, Yuki Mitsufuji, and Simon Dixon. 2024a. Instruct-musicgen: Unlocking text-to-music editing for music language models via instruction tuning. *arXiv preprint arXiv:2405.18386*.

Yixiao Zhang, Yukara Ikemiya, Gus Xia, Naoki Murata, Marco A Martínez-Ramírez, Wei-Hsiang Liao, Yuki Mitsufuji, and Simon Dixon. 2024b. Musicmagus: Zero-shot text-to-music editing via diffusion models. *arXiv preprint arXiv:2402.06178*.

Yu Zhang, Changhao Pan, Wenxiang Guo, Ruiqi Li, Zhiyuan Zhu, Jialei Wang, Wenhao Xu, Jingyu Lu, Zhiqing Hong, Chuxin Wang, et al. 2024c. Gtsinger: A global multi-technique singing corpus with realistic music scores for all singing tasks. *Advances in Neural Information Processing Systems*, 37:1117–1140.

Haixin Zhao. 2023. A gan speech inpainting model for audio editing software. In *Proc. Interspeech 2023*.

Kateřina Žmolíková, Marc Delcroix, Keisuke Kinoshita, Tsubasa Ochiai, Tomohiro Nakatani, Lukáš Burget, and Jan Černockỳ. 2019. Speakerbeam: Speaker aware neural network for target speaker extraction in speech mixtures. *IEEE Journal of Selected Topics in Signal Processing*, 13(4):800–814.

A Scope

In this survey, we focus on works that make direct contributions to audio editing in the era of foundation models. To ensure a precise and focused discussion, we adopt two main inclusion criteria: (1) the task should center on audio editing, which we define as modifying the acoustic attributes, instances, or content of an existing audio recording, without transformations so substantial that they amount to generating an entirely new audio sample; (2) the method should rely on mainstream audio foundation model paradigms. Accordingly, we do not cover works primarily focused on audio generation, nor do we provide an extensive discussion of signal-processing-based audio editing methods. In addition, to maintain a focused scope, spatial audio[3] and related editing techniques are beyond the main scope of this survey.

B Discussion

B.1 Foundation Models with Discrete and Continuous Representations

The surveyed methods suggest that audio editing with foundation models is primarily shaped by the underlying audio representation and generation process. Codec language models discretize audio into token sequences, which provides a compact interface for infilling, continuation, and generation conditioned on editing tasks. This makes them particularly effective for speech editing, where linguistic content, speaker identity, and local temporal structure can be explicitly modeled. By contrast, diffusion and flow matching models operate on continuous acoustic representations, such as spectrograms or latent features, and are therefore more suitable for high fidelity reconstruction, fine grained acoustic manipulation, and generation constrained to target regions.

These architectural choices lead to different failure modes. Discrete token models may lose spectral details due to codec compression and can suffer from autoregressive latency or error propagation, especially in music and general audio with complex timbral dynamics. Continuous generative models better preserve local acoustic structure, but their editing behavior is highly sensitive to target localization, source conditioning, mask design, and preservation objectives. In both cases, the key technical difficulty is not only to generate plausible target audio, but also to enforce edit locality, so that the modified region satisfies the editing condition while the surrounding content remains perceptually unchanged. Overall, progress in audio editing will likely depend on architectures that jointly support semantic controllability, continuous high fidelity synthesis, and explicit preservation of regions outside the edit target, rather than optimizing generation quality alone

B.2 Training-based and Training-free Editing Paradigms

Training-based and training-free methods reflect different assumptions about how editing ability is acquired. Training-based methods learn editing behaviors from paired data, pseudo pairs, or instruction style supervision. Their advantage lies in stability: when the task and data distribution are clearly defined, they can directly optimize edit success, preservation of non-edited regions, and output naturalness. This makes them suitable for reliable instruction following, attribute control, and repeated use of fixed editing functions. However, their performance depends heavily on the coverage and quality of training data. Given the scarcity of large-scale instruction aligned audio editing datasets, many methods still rely on synthetic pairs, task specific pipelines, or adapted resources, which may introduce distribution bias and limit generalization to open domain requests.

Training-free methods instead reuse pretrained generative models without parameter updates. Through inversion, attention modification, or mask constraints, they exploit the editing capacity already encoded in foundation models. This paradigm avoids costly editing specific training and can be applied to new prompts, domains, or edit types, especially when paired editing data are unavailable. Its main limitation is weaker controllability and less predictable preservation. Editing quality often depends on accurate inversion, meaningful attention localization, and reliable separation between edited and preserved regions.

The two paradigms are complementary. Training-based methods are better suited to robust systems under well specified task settings, while training-free methods offer flexibility for open domain and low resource editing. Future audio editing systems may combine training-based adaptation for instruction following and preservation with training-free mechanisms for localization,

[3] Spatial audio refers to multi-channel audio formats, such as binaural stereo and first-order Ambisonics (FOA).

source control, and test time refinement.

C Detailed Description of the Task

We note that the proposed taxonomy does not define mutually exclusive task boundaries. In practice, audio editing operations are often compositional, and a single instruction may belong to multiple task classes. We therefore view the taxonomy as a functional organization based on the primary edited information layer, rather than a strict one-to-one mapping between an instruction and a category.

A typical example is the instruction "remove the background crowd noise from the speech". This operation can be viewed as denoising or restoration when the crowd sound is treated as an unwanted degradation. It can also be regarded as instance-level deletion when the crowd is modeled as a separable sound source in the audio scene. This example shows that real-world audio editing tasks are often multi-label by nature. In the following descriptions, we assign each task to the category that best reflects its dominant editing goal, while allowing overlaps across acoustic, semantic, and instance-level editing.

C.1 Acoustic Editing

Loudness and volume editing constitute one of the most common forms of acoustic editing. This capability reorganizes auditory attention by making certain components more prominent or by adjusting the balance between foreground and background sounds (Jiang et al., 2024; Choi et al., 2021; Liang et al., 2024). For example, in advertising scenarios, adjusting key segments and the foreground-background balance can largely enhance expressiveness and listener engagement.

Denoising and restoration are another important class of acoustic editing, focusing on the suppression of undesired degradations, such as background noise, codec artifacts, and local corruption, while preserving useful acoustic cues and source characteristics (Liu et al., 2022a). Rather than erasing all environmental traces, these tasks aim to remove unwanted degradations while preserving the naturalness of the original audio (Fu et al., 2021).

Reverberation editing modifies room-acoustic attributes, such as room response, early reflections, and reverberation tails (Chen et al., 2022a). It adjusts spatial and temporal decay patterns to change the perceived acoustic environment while preserving the underlying source. Related tasks such as dereverberation and room-acoustic control suggest that room impulse responses can serve as editable acoustic conditions rather than merely distortions to be removed (Bahrman et al., 2024).

Audio equalization is an essential manifestation of acoustic editing that traditionally requires substantial domain expertise. It modifies the spectral energy distribution of the raw audio to control perceptual qualities such as brightness, warmth, sharpness, or muffledness. Also, recent studies represent that equalization can be learned and parameterized by neural networks, which enables descriptive language control, such as “brighter”, “warmer”, or “more distant” (Venkatesh et al., 2022).

C.2 Semantic Editing

Linguistic editing revises explicit language-bearing content. In speech, it inserts, deletes, or replaces words and phrases while preserving speaker timbre, surrounding acoustics, and smooth transitions across edited regions (Tan et al., 2021). In music, related operations include modifying lyrics or short vocal phrases while maintaining melody, singer identity, and accompaniment consistency (Zhang et al., 2022). Such capabilities support correction, updating, and localized revision without regenerating the entire recording.

Expressive editing changes how audio is delivered or emotionally perceived. For speech, it covers emotion (Wang et al., 2024a), prosody, speaking rate, pitch contours, accent, and paralinguistic cues such as whispers, laughter, sighs, and hesitation (Kim et al., 2025). For general audio, it may adjust the affective character of a sound event, making it more tense, gentle, dramatic, or playful. These tasks require controlled transformation of expressive attributes while preserving content, source identity, and naturalness in edited audio.

Stylistic editing targets high-level musical and stylistic properties, such as genre, mood, melodic expression, rhythm, or arrangement-level characteristics (Zhang et al., 2024b). Beyond music, stylistic editing may also change the presentation style of speech or general audio, such as making speech more formal, colloquial, or advertisement-like, or making a sound more realistic (Anastassiou et al., 2024; Jin et al., 2023). Since these attributes are often subjective and language-described, this type of editing requires models to align style descriptions with coherent audio transformations.

C.3 Instance Editing

Instance replacement substitutes a target entity with another instance while preserving the surrounding context. In speech, this often corresponds to voice conversion, where the speaker identity is changed while the linguistic content and local acoustics are retained (Wang et al., 2021; Qian et al., 2019). In general audio, it may involve replacing one sound event with another (Xu et al., 2024); in music, it may replace a vocal, instrumental, or local musical component (Han et al., 2023). The main challenge is to ensure that the substituted instance remains coherent with the acoustic context, temporal alignment, and source relationships of the original recording (Fu et al., 2025).

Instance deletion and extraction remove or isolate a target entity from a mixture. Deletion suppresses the target while preserving the remaining scene (Ellis et al., 2025), whereas extraction isolates it for separate use or further editing (Liu et al., 2024c). Typical examples include target-speaker extraction (Žmolíková et al., 2019), sound-event removal (Fu et al., 2025), and music stem separation (Défossez, 2021), all requiring accurate instance discrimination and preservation of non-target component within complex mixtures.

Instance insertion introduces a new entity into an existing recording, such as a speaker turn, sound event, instrument stem, or vocal phrase. Common examples include inserting sound effects into an acoustic scene (Wang et al., 2023b) or extending a dialogue with an additional utterance (Jin et al., 2017). Unlike generation from scratch, insertion requires the added instance to be coherent with the original acoustic environment, temporal layout, and, when applicable, musical arrangement (Rouard et al., 2025; Peng et al., 2024).

Instance overlay layers or emphasizes an instance without fully replacing existing components. It covers foreground enhancement, soundscape remixing, background ambience enrichment, and music-layer addition (Ellis et al., 2025). It may enhance a target speaker in a noisy conversation, add crowd ambience to an outdoor scene, or layer a rhythmic component over existing music. Unlike insertion, overlay editing preserves the original mixture while adjusting the salience, density, or texture of selected instances, making it useful for enriching or rebalancing an audio scene under global coherence constraints (Strano et al., 2025).

D Data Construction Tools

Given the scarcity of complete instruction-based editing datasets, data construction tools are essential for converting raw audio resources into reliable, structured, and reusable editable supervision. We group these tools by the artifacts they produce: temporal boundaries for locating edit regions, semantic annotations for describing edit targets, and paired examples for defining desired edited outputs.

Temporal localization tools Temporal localization tools determine where an edit should be applied. In speech editing, this is often achieved through text-audio alignment: Praat (Boersma and Weenink, 2021) supports manual boundary inspection and acoustic analysis, MFA (McAuliffe et al., 2017) provides automatic phone- or word-level forced alignment, and WhisperX (Bain et al., 2023) generates word-level timestamps for long-form speech. When the editable unit is defined by speaker activity rather than text, pyannote (Bredin, 2023) and VAD tools (Karan et al., 2024) can provide speaker-active segments. For general audio, sound event detection models (Kong et al., 2020; Li et al., 2023) produce event-level activity boundaries. In music editing, temporal localization often relies on pitch or note-level cues: Parselmouth (Jadoul et al., 2018), RMVPE (Wei et al., 2023), and CREPE (Kim et al., 2018) extract time-varying F0 contours, while ROSYOT (Li et al., 2024b) and MusicYOLO (Wang et al., 2022) provide precise note-level onset and offset annotations for editing in structured workflows.

Semantic annotation tools After editable areas are localized, semantic annotation tools assign interpretable descriptions or labels to them. For speech, ASR and alignment systems (Gao et al., 2023; Radford et al., 2023) provide transcripts that can serve as content annotations. For non-speech audio, tagging, detection, and captioning tools can produce sound-event labels, scene labels, or natural-language descriptions (Chen et al., 2022b); SELD models (Adavanne et al., 2018; Guirguis et al., 2021), for example, provide event-oriented annotations with localization information. Attribute annotations are also important for expressive editing: emotion2vec (Ma et al., 2024c) offers speech emotion representations that can support large-scale affective labeling. Recent audio-language models, such as Qwen3-Omni (Xu et al., 2025), can further assist caption generation and

tag-to-description conversion, making raw labels more suitable for instruction-style supervision.

Pair construction tools Pair construction tools define the desired edited output by creating input-target examples. For speech editing, controllable TTS systems such as MaskGCT (Wang et al., 2025d) and StyleTTS (Li et al., 2025) can synthesize paired examples with controlled content, speaking style, or prosody. Voice conversion systems such as AutoVC (Qian et al., 2019) and YourTTS (Casanova et al., 2022) can transform speaker identity while preserving linguistic content, providing paired examples for speaker replacement or timbre-level editing. For acoustic editing, degradation simulation can construct restoration pairs by adding noise, reverberation, bandwidth limitation, or codec artifacts to clean recordings. For instance-level editing, source separation and remixing tools provide another route: Open-Unmix (Stöter et al., 2019), Spleeter (Hennequin et al., 2020), and Demucs (Rouard et al., 2023) decompose music into stems, while AudioSep (Liu et al., 2024c) and SAM-Audio (Shi et al., 2025) extend separation to open-domain or prompt-based scenarios with text, visual, or temporal-span prompts. These tools enable the construction of insertion, deletion, extraction, replacement, and overlay pairs from existing audio mixtures for supervised editing.

E Future Directions and Challenges

Despite the success achieved in audio editing in the foundation model era, there are still limits to address in future work across practical applications.

E.1 Complex Editing

Although recent audio editing methods have already achieved promising progress in specific domains, their editing capability remains limited when facing complex and entangled audio scenes. Existing audio editing methods are still largely shaped by domain-specific task definitions. Speech editing mainly focuses on text replacement, speaker identity preservation, and speech naturalness (Peng et al., 2024; Huang et al., 2024). General audio editing is often formulated as event-level addition, deletion, or replacement (Wang et al., 2023b). Music editing further requires the preservation of harmony, rhythm, instrumentation, and long-range musical structure (Han et al., 2023). Although these task-specific designs have enabled effective editing in constrained settings, they remain far from a unified editing framework that can precisely modify arbitrary audio components while preserving task-irrelevant content in open scenarios.

The central challenge stems from the highly coupled nature of audio. Real-world recordings often entangle semantic events, speaker identity, acoustic texture, background ambience, rhythm, spatial cues, and reverberation within a single waveform. As a result, audio editing must not only generate plausible target content, but also localize the editable region, modify the intended attributes, and preserve non-target context in timbre, timing, loudness, and acoustic coherence. Such requirements become especially difficult for overlapping sources, long-form speech, dense soundscapes, and music, where object boundaries are often weak or implicit.

Future research should move toward disentangled and object-aware audio editing. One possible direction is to develop source-aware representations that separate editable objects from non-target contexts, enabling more reliable local generation and preservation. Another direction is to introduce structured control signals, such as temporal regions, event boundaries, speaker labels, source masks, acoustic attributes, or spatial trajectories, to reduce the ambiguity of free-form instructions. Recent unified audio generation models (Yang et al., 2023; Valle et al., 2025) suggest the potential of handling speech, sound, singing, and transformation within a general framework. However, future audio editing systems should go beyond unified generation and clearly support object localization, attribute-level modification, and non-target preservation across speech, music, and general audio.

E.2 Robustness Under Open-Domain

Robustness under open-domain and real-world conditions remains a central challenge for audio editing. Existing methods often perform well on predefined tasks, clean inputs, or carefully constructed instructions, but their behavior becomes unstable when the audio contains noise, reverberation, overlapping sources, long-range dependencies, or ambiguous user intents. These failures arise from two coupled difficulties: the model must ground the editing instruction to the correct acoustic target, and it must preserve non-target content under complex acoustic conditions in real-world mixtures.

Different modeling paradigms expose this challenge in different forms. Codec language model-based methods formulate editing as token infilling, contextual continuation, or text-conditioned gener-

ation, and have shown strong potential in speech editing, as demonstrated by VoiceCraft (Peng et al., 2024) and SpeechX (Wang et al., 2024b). However, their effectiveness depends heavily on stable text-acoustic alignment and high-quality tokenization. When applied to music or general audio, weak semantic boundaries, dense source mixtures, and long-term structural constraints may lead to token-level errors, context drift, and fidelity degradation. Diffusion and flow-matching models, by contrast, are strong at generating high-fidelity acoustic details in continuous latent spaces, making them suitable for editing timbre, background, sound texture, and musical details. Yet their local controllability remains limited: instruction-guided latent diffusion editing has shown promising results (Wang et al., 2023b; Gao et al., 2026), but local modifications may affect non-target regions, especially in music and complex soundscapes where acoustic components are strongly correlated (Han et al., 2023).

Future work should therefore move beyond improving generation quality alone and focus on robust editing behavior. One direction is to improve instruction grounding, enabling models to infer the intended target, editable region, desired attribute, and preservation scope from natural language. Another direction is to strengthen long-context modeling and preservation mechanisms, so that edits remain stable in long-form or multi-source audio with extended dependencies. Iterative refinement, self-verification, retrieval-augmented editing, and multi-stage correction may further help models recover from uncertain or failed edits, especially in multi-turn editing scenarios.

E.3 Faithful and Specific Evaluation

Faithful evaluation remains a critical bottleneck for audio editing. The first challenge lies in the mismatch between existing benchmarks and editing tasks. Current audio resources mainly focus on generation, tagging, captioning, understanding, or audio-language reasoning, rather than editing behavior itself. Datasets such as AudioSet (Gemmeke et al., 2017) and AudioCaps (Kim et al., 2019) support audio tagging, captioning, and audio-language alignment, but lack the (`instruction`, `input`, `output`) triplets needed to verify whether a specific edit is correctly performed. Similarly, benchmarks such as MARBLE (Yuan et al., 2023) and AudioBench (Wang et al., 2025a) assess music understanding or general audio-language understanding and reasoning, but are not designed to measure edit success, non-target preservation, or instruction-following quality in edited outputs. As a result, current protocols often conflate generation quality with editing quality.

The second challenge comes from the multi-dimensional nature of audio editing. Unlike audio generation, editing requires modifying only the intended content while preserving the remaining audio and maintaining perceptual coherence. A faithful protocol should therefore jointly assess target modification, non-target preservation, acoustic naturalness, and instruction following (Wang et al., 2023b). These dimensions are related but not equivalent: an edited sample may sound natural yet fail the requested edit, or follow the instruction while unintentionally changing speaker identity, background ambience, rhythm, or acoustic texture.

The third challenge is the limitation of existing evaluation tools. Most automatic metrics capture only partial aspects of editing quality. Text-audio similarity reflects semantic alignment but not preservation (Elizalde et al., 2023); ASR-based accuracy and speaker similarity are useful for speech editing but less applicable to music or general audio (Morris et al., 2004; Snyder et al., 2018); acoustic quality predictors can detect artifacts but cannot judge whether the edit is correct (Reddy et al., 2021). Recent audio-language and multimodal models (Tian et al., 2025; Xu et al., 2025; Comanici et al., 2025) offer a promising direction for automatic judges, but they must be calibrated against human listening results to avoid overestimating instruction alignment while missing subtle artifacts, temporal discontinuities, or unintended non-target changes in realistic evaluations.

Future evaluation should shift from generation-centered to editing-centered assessment. Reliable benchmarks should provide explicit annotations of target regions, edit operations, preservation regions, and control signals, enabling separate measurement of edit success and non-target preservation. Metrics should also be multi-dimensional, covering instruction following, target modification, content preservation, acoustic naturalness, and perceptual coherence. Human listening tests remain essential for calibrating automatic evaluators. Although human-aligned audio editing judges that combine semantic understanding with perceptual sensitivity offer a scalable direction, their reliability must be validated across diverse tasks, domains, and acoustic conditions. Such benchmarks and evaluators are crucial for advancing audio editing toward robust

real-world applications at practical scale.