A Unified Neural Codec Language Model for Selective Editable Text to Speech Generation

Anonymous ACL submission

Abstract

Neural codec language models achieve impressive zero-shot Text-to-Speech (TTS) by fully imitating the acoustic characteristics of a short speech prompt, including timbre, prosody, and paralinguistic information. However, such holistic imitation limits their ability to isolate and control individual attributes. In this paper, we present a unified codec language model SpeechEdit that extends zero-shot TTS with a selective control mechanism. By default, SpeechEdit reproduces the complete acoustic profile inferred from the speech prompt, but it selectively overrides only the attributes specified by explicit control instructions. To enable controllable modeling, SpeechEdit is trained on our newly constructed LibriEdit dataset, which provides delta (difference-aware) training pairs derived from LibriHeavy. Experimental results show that our approach maintains naturalness and robustness while offering flexible and localized control over desired attributes. Audio samples are available at https://speech-editing.github.io/speech-editing/.

1 Introduction

Recent zero-shot Text-to-Speech (TTS) generation has advanced rapidly with the rise of modern generative modeling, enabling high-fidelity voice cloning from short, unseen reference prompts. Existing systems leverage diverse acoustic representations, including discrete token-based approaches (Chen et al., 2025a; Łajszczak et al., 2024; Wang et al., 2025d), continuous representations (Meng et al., 2025; Eskimez et al., 2024; Chen et al., 2025b; Wang et al., 2025b), and hybrid token modeling (Du et al., 2024; Yang et al., 2025b; Anastassiou et al., 2024). Despite these advances, models treat the reference audio as a holistic, black-box condition, leaving key vocal attributes, such as timbre, emotion, prosody, and paralinguistic style, entangled and difficult to control independently.

This limitation has motivated growing interest in controllable speech synthesis, where fine-grained manipulation of attributes enables more flexible, expressive, and personalized voice generation (Xie et al., 2025). Existing control paradigms include text-driven, audio-driven, and hybrid approaches. Text-driven methods rely on textual directives, including style tags (Wang et al., 2025c), natural language descriptions (Guo et al., 2023), or instructions (Zhou et al., 2024), offering explicit high-level control but often failing to capture subtle acoustic details or reproduce a specific speaker's voice. Audio-driven approaches use dual speech prompts to separately specify timbre and style (Zhang et al.; Zhou et al., 2025), partially alleviating these limitations. Hybrid systems (Yang et al., 2025a; Du et al., 2024) combine textual instructions with audio prompts to balance explicit control and acoustic fidelity. However, when multiple prompts are used to control different aspects of speech, interactions between them can lead to attribute leakage and conflicts, where unintended prosodic or stylistic cues affect the output. These challenges necessitate a more precise approach that supports the fine-grained editing of individual speech attributes.

In this work, we formulate controllable speech generation as a *selective attribute editing problem*. Given a speech prompt p, a target text x, and an edit specification e, the goal is to generate speech that preserves the inherent attributes of p, such as speaker identity, while modifying only those explicitly indicated by e, such as the emotion. The editable space in this work spans three fundamental and interpretable dimensions of expressive speech: *(1) Emotion-related attributes* describe the affective state. *(2) Prosody-related attributes* characterize paralinguistic properties such as pitch, speaking speed, and energy, which jointly determine how the utterance is realized acoustically. *(3) Speaker-related attributes* primarily correspond to timbre.

Unlike conventional TTS or global style-transfer settings, this setup requires fine-grained, attribute-level editing while keeping all unspecified components faithful to the reference. For speaker editing, the system is given an extra speaker prompt for timbre mimic, much like voice conversion task, but within a TTS framework supporting random target text sequence.

Rather than explicitly disentangling speech attributes through specialized architectures or training schemes, we hypothesize that the in-context learning capability of neural codec language models (LMs)—trained on large, diverse datasets spanning multiple speakers, emotions, and vocal attributes—naturally provides implicit disentanglement. Building on this intuition, we design SpeechEdit, which treats the speech prompt as a base canvas and selectively modifies only the attributes specified by the user. This unified formulation enables a single model to seamlessly support zero-shot TTS, voice conversion, and fine-grained style editing. To train SpeechEdit, we construct a new dataset, LibriEdit, by labeling the speech attributes of utterances from LibriHeavy. We introduce a Delta-Pairs sampling method to generate training triplets (speech prompt, edit specification, speech target) by randomly sampling two utterances from LibriEdit and designating one as the prompt and the other as the target, with the differing attributes between them as the edit specification. Experimental results conducted on various speech editing tasks show that SpeechEdit achieves highly competitive performance on naturalness and robustness, while reaching state-of-the-art (SOTA) performance in selective speech editing. Our main contributions are as follows:

- We propose SpeechEdit, a unified selective editing framework that leverages the in-context learning capability of neural codec LMs to integrate zero-shot TTS, voice conversion, and style editing within a single model, enabling precise attribute-level control while faithfully preserving speaker identity.
- We introduce a data-driven implicit disentanglement strategy that combines assumption-free Delta-Pairs sampling with our newly annotated LibirEdit dataset, enabling promising separation of speaker identity and style attributes without complex auxiliary modules and providing a scalable paradigm for expressive speech synthesis.

2 Related Work

2.1 Neural Codec LM for Speech Synthesis

Neural codec language modeling treats speech synthesis as a sequence modeling problem over discrete acoustic tokens obtained from neural audio codecs. VALL-E (Chen et al., 2025a) pioneered this direction by proposing a hybrid Autoregressive (AR) and Non-Autoregressive (NAR) architecture. Subsequent studies have explored various aspects of neural codec LMs, including improving robustness (Chen et al., 2024; Han et al., 2024; Song et al., 2025), efficiency (Yang et al., 2025b; Chen et al., 2024; Kim et al., 2024). Across generation architectures, codec language models involve clear trade-offs. AR models achieve strong perceptual quality by modeling temporal dependencies, but suffer from slow inference and error accumulation, while NAR and partially NAR models improve efficiency via parallel generation and duration modeling, often at the cost of temporal coherence (Yang et al., 2025b; Wang et al., 2025d). Recent studies have explored enhancing expressive speech generation through richer conditioning signals, such as style tokens or textual instructions (Ji et al., 2024; Wang et al., 2025c; Zhou et al., 2025), while fine-grained, attribute-level control remains challenging.

2.2 Controllable Speech Synthesis

Controllable speech synthesis generates natural, intelligible speech from text while enabling explicit control of specific speech attributes. Existing works explore different control dimensions, including prosody (Wang et al., 2025c), emotion (Gao et al., 2025), dialect (Du et al., 2024), and paralinguistic features (Liao et al., 2025). A critical challenge arises when additional controls are applied while preserving speaker identity: attribute conflict. The reference audio inherently carries its own timbre, prosody, and emotion, which can conflict with the target style specified by text or auxiliary prompts. To address this, systems typically employ either implicit or explicit disentanglement strategies. One approach, exemplified by EmoVoice (Yang et al., 2025a), uses neutral reference audio to mitigate conflicts. Explicit disentanglement methods resolve conflicts through mechanisms such as gradient reversal layers (Ju et al., 2024; Zhou et al., 2025) or information bottlenecks within codebooks (Zhang et al.), which may still suffer from incomplete attribute separation and require additional model components.

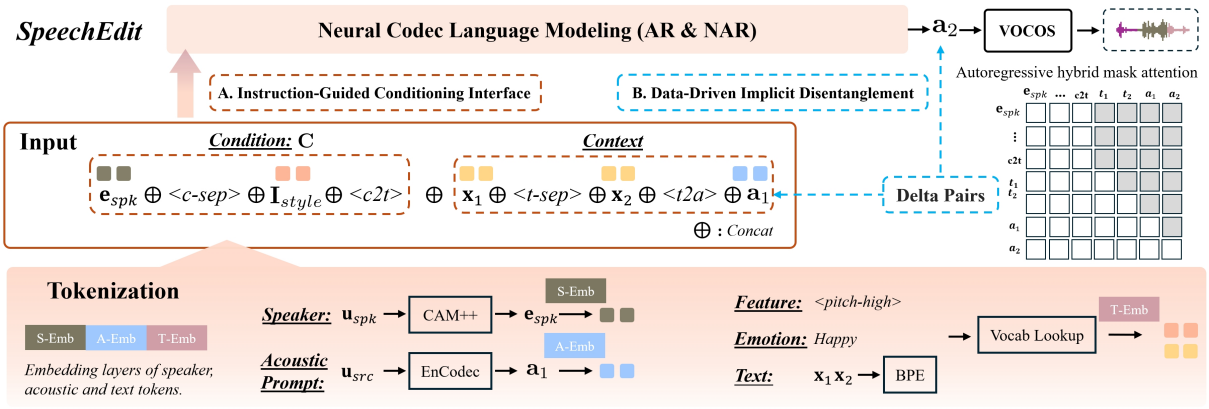

Figure 1: Overview of the SpeechEdit framework. Instruction tokens, textual content, and acoustic prompts are unified into a single token sequence through an instruction-guided conditioning interface. The codec language model performs selective attribute editing through data-driven implicit disentanglement with delta pairs.

3 Proposed Method

We formulate selective editable speech generation as a prompt-guided neural codec language modeling task, where editing is achieved by explicit instruction conditioning in the discrete codec token space. Following Encodec (Défossez et al., 2022), a speech waveform is represented as a sequence of discrete codec tokens $\mathbf{y} \in \mathbb{Z}^{T\times 8}$, where T is the number of time steps across 8 codebook layers. The token $\mathbf{y}_{t,j}$ denotes the discrete index at time step t from the j-th codebook layer. Building on the paradigm introduced in VALL-E (Chen et al., 2025a), the proposed SpeechEdit extends this framework to support attribute-level speech editing via unified instruction conditioning. As shown in Figure 1, both the AR and NAR stages share the same conditioning signals. Given a speech prompt $\mathbf{a}_1$, its transcription $\mathbf{x}_1$, a target text $\mathbf{x}_2$, and an editing specification condition $\mathbf{C}$, the AR model predicts the first codebook layer to capture the fundamental prosodic and phonetic structure:

$$\mathcal{L}_{\text{AR}} = -\sum_{t=1}^{T} \log p\left(\mathbf{y}_{t,1} \mid \mathbf{P}, \mathbf{y}_{<t,1};\, \boldsymbol{\theta}_{\text{AR}}\right), \quad (1)$$

where $\mathbf{y}_{<t,1}$ are previously generated tokens, $\mathbf{P} = [\mathbf{C}, \mathbf{x}_1, \mathbf{x}_2, \mathbf{a}_1]$ is the concatenated conditioning prompt, and $\boldsymbol{\theta}_{\text{AR}}$ denotes the AR model trainable parameters. Conditioned on the first-layer predictions, the NAR model refines acoustic details by generating the subsequent layers $\mathbf{y}_{:,j}, j \in [2, 8]$:

$$\mathcal{L}_{\text{NAR}} = -\sum_{t=1}^{T} \log p(\mathbf{y}_{t,j} \mid \mathbf{P}, \mathbf{y}_{:,<j};\, \boldsymbol{\theta}_{\text{NAR}}). \quad (2)$$

Unlike prior speech editing systems that rely on task-specific architectures or auxiliary disentanglement modules, SpeechEdit enables flexible and compositional attribute control through a unified instruction-driven framework.

3.1 Instruction Guided Interface

We adopt a discrete instruction guided interface to model multiple speech attributes during generation as show in the bottom of Figure 1.

Categorical Attributes Emotion and prosody attributes are represented as instruction tags. Emotion is modeled with five predefined classes: *Neutral*, *Happy*, *Sad*, *Angry*, and *Surprise*. Prosody attributes, including pitch, energy, and speaking speed, are discretized into five ordinal levels ranging from *Very Low* to *Very High* and expressed using a structured tag format such as <*pitch-high*> or <*speed-low*>. All instruction tags share the same vocabulary table and text embedding layer with Byte-Pair Encoding (BPE)-tokenized text.

Speaker Identity Speaker identity is inherently a continuous and high-dimensional factor. We extract a global speaker embedding $\mathbf{e}_{spk}$ from a speaker reference utterance using a pretrained voice print model[1]. This embedding is projected into the LM space via a dedicated speaker embedding layer.

3.2 Data-Driven Implicit Disentanglement

Instead of using auxiliary modules to decouple speech attributes, we adopt a *Delta Pair Sampling* strategy to achieve data-driven implicit disentanglement, as illustrated by the blue dashed box in Figure 1 where training pairs are deliberately constructed with amplified attribute discrepancies to

[1] https://github.com/alibaba-damo-academy/3D-Speaker/tree/main/egs/3dspeaker/sv-cam++

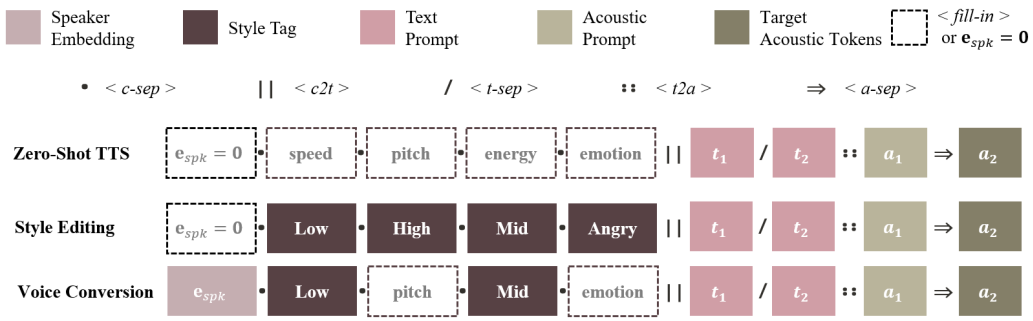

Figure 2: Token sequence composition for different tasks within SpeechEdit.

guide the model's attention to the explicit control signals.

Same-speaker Delta-Pair Sampling. Two utterances from the same speaker are sampled: a source speech $\mathbf{u}_{src}$ and a target speech $\mathbf{u}_{tgt}$ from different emotion categories (e.g., $\mathbf{u}_{src}$ is Happy while $\mathbf{u}_{tgt}$ is Angry). During training, LMs are conditioned on the style tags of $\mathbf{u}_{tgt}$ but the acoustic prompt of $\mathbf{u}_{src}$. This guides the attention mechanism to: (1) extract unspecified attribute from the prompt, and (2) derive the target attributes from the style instructions.

Cross-speaker Delta-Pair Sampling. The source speech $\mathbf{u}_{src,spk_1}$ and target speech $\mathbf{u}_{tgt,spk_2}$ are sampled from different speakers. The model is primarily conditioned on the source acoustic tokens of $\mathbf{u}_{src,spk_1}$ along with instruction prompts for target attributes. A separate speaker reference utterance $\mathbf{u}_{ref,spk_2}$ from the target speaker provides speaker embedding to define the target identity. This reference utterance is content-independent of $\mathbf{u}_{tgt,spk_2}$, which prevents content leakage.

By conditioning the model on mismatched acoustic prompts and target instructions, the explicit instruction tokens become the only consistent signal for the attention mechanism, enabling implicit disentanglement through Delta-Pair sampling.

3.3 Instruction Composition

SpeechEdit unifies multiple speech generation and editing tasks within a single model by reorganizing conditioning tokens, as shown in Figure 2. For *zero-shot TTS*, setting $\mathbf{e}_{spk} = \mathbf{0}$ and all style tokens to <*fill-in*> forces the model to rely entirely on the acoustic prompt for timbre and prosody. For *style editing*, specific style tags are explicitly overridden. By remaining assumption-free with respect to training-pair attributes, our Delta-Pair sampling strategy ensures that explicit style instructions consistently override the prompt when the two are in conflict. For *voice conversion*, a target speaker embedding $\mathbf{e}_{spk}$ specifies the new identity. Style tokens can be a hybrid of explicit tags and <*fill-in*>, allowing prosody transfer or partial editing. The final input sequence is structured as:

$$\mathbb{S}_{in} = \underbrace{[\mathbf{e}_{spk} \oplus \text{<c-sep>} \oplus \mathbf{I}_{style}]}_{\text{Conditioning}} \oplus \text{<c2t>} \oplus \quad (3)$$

$$\underbrace{[\mathbf{x}_1 \oplus \text{<t-sep>} \oplus \mathbf{x}_2 \oplus \text{<t2a>} \oplus \mathbf{a}_1]}_{\text{Context}},$$

where <*c-sep*>, <*t-sep*>, and <*a-sep*> separate elements within the same block, while <*c2t*> and <*t2a*> indicate transitions across modalities, marking boundaries between global conditioning, text, and acoustic prompts.

4 LibriEdit Dataset

4.1 Overview of LibriEdit

While emotionally or stylistically expressive speech can be collected at scale and efficiently annotated using LLMs, the effective data volume is often shrinks drastically once speaker annotations are required, as shown in Table 2, which severely limits the ability to learn fine-grained, speaker-preserving attribute control. To address this limitation, we build a style-labeled corpus based on the Libri-Heavy dataset (Kang et al., 2024) which is chosen for three reasons: (1) it provides a large-scale collection of read speech, with over 50k hours in its large split; (2) audiobook narration naturally contains expressive yet non-exaggerated emotional cues that well aligned with daily speaking styles, and (3) it offers reliable speaker identities, enabling speaker-consistent style mining. The resulting LibriEdit dataset comprises 2566 speakers with a total 708 hours of speech.

4.2 Dataset Construction Pipeline

Our LibriEdit is constructed following three steps: segmentation, emotion annotation and other at-

Emotion	Threshold
Happy	0.60
Sad	0.90
Angry	0.80
Neutral	0.60
Surprise	0.75

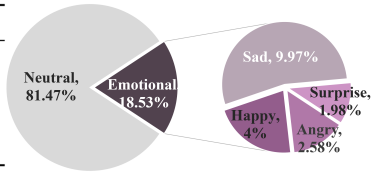

Figure 3: Confidence thresholds for emotion labeling and the resulting distribution of emotions in LibriEdit.

tribute annotation.

Step 1: Preprocessing and Fine-Grained Sentence Segmentation. Following the official LibriHeavy script, we begin by cutting long audiobook chapters into sentence-level segments. However, audiobook narration often exhibits style variation within a single sentence, such as neutral narration interleaved with emotionally expressive quoted speech, which remains too coarse for style labeling. So, we further refine the segmentation using Montreal Forced Aligner[2] by splitting at breath-group boundaries and punctuation-aligned pauses. This yields shorter prosodic segments with more consistent speaking styles. A minimum duration of 2 seconds is enforced to ensure sufficient acoustic context.

Step 2: Emotion Annotation. We begin by automatically labeling the emotion of each segment using a categorical speech emotion recognition model (SER) [3], which predicts an 8-way emotion distribution and outputs a confidence score for each category. Preliminary analysis shows that the categories *fear*, *disgust*, and *contempt* are highly ambiguous and low perceptual consistency. We therefore discard these classes and retain five reliably distinguishable emotions: *neutral*, *happy*, *sad*, *angry*, and *surprise*. To improve label reliability, we apply emotion-specific confidence thresholds and keep only segments whose predicted probabilities exceed the corresponding thresholds, as summarized in the left of Figure 3. We further refine the emotion labels via multi-model cross-validation with emotion2Vec-plus-large (Ma et al., 2024) and Audio Flamingo 3 (Ghosh et al., 2025). A majority voting scheme is adopted, preserving only segments agreed upon by at least two models and discarding those with disagreement. The final label is corrected to the majority decision and the prompting strategy used for Audio Flamingo 3 is provided in the Appendix A.1. The distribution of emotion-labeled data is shown in the right of Figure 3, totaling 129 hours of emotional speech.

Step 3: Prosody Attribute Annotation. In addition to emotion labels, we annotate speed, pitch, and energy using signal-processing-based estimators.

5 Experiment Setup

5.1 Implementation Detail

Training Dataset. We train the SpeechEdit model on the annotated LibriEdit corpus with same-speaker and cross-speaker delta pair sampling each accounting for 50% of the data, covering diverse variations in prosody and emotional expression.

Model Configuration. Both AR and NAR stages of SpeechEdit share a consistent backbone: a 12-layer decoder-only Transformer with 16 attention heads per layer, an embedding dimension of 1,024, and a feed-forward network with a dimensionality of 4,096 with ReLU activation. To enhance contextual modeling, the first-stage AR model employs a modified causal mask that allows bidirectional attention over prefix conditional tokens while maintaining causal attention on the following context tokens. Transcriptions are tokenized using BPE, and audio waveforms are discretized into speech tokens using the open-source EnCodec[4] operating at a 6 kbps bitrate for 24 kHz audio.

Training and inference. Both stages are trained on 16 NVIDIA Tesla V100 GPUs (32GB), with a maximum batch size of 10k tokens per GPU. The model is optimized using Adam with $\beta = (0.9, 0.98)$ and a weight decay of 0.01. We employ an inverse square-root learning rate schedule with linear warm-up, where the learning rate increases linearly from 0 to 5×10^{-4} over the first 32k update steps, followed by inverse square-root decay. SpeechEdit is first pretrained on LibriHeavy-large following the VALL-E setup for 800k updates, and then further trained on the target training dataset for an additional 800k updates. The same optimization strategy is applied in both stages, with all model parameters updated. During inference, we adopt the decoding strategy of Chen et al. (2024), using top-p sampling with a repetition penalty.

5.2 Baselines and Evaluation Metrics

We compare SpeechEdit with four SOTA systems: VALL-E (Chen et al., 2025a), which shares a similar backbone for fair zero-shot comparison; Step-

[2] https://github.com/MontrealCorpusTools/Montreal-Forced-Aligner

[3] https://huggingface.co/3loi/SER-Odyssey-Baseline-WavLM-Categorical

[4] https://github.com/facebookresearch/encodec

Audio-EditX (Yan et al., 2025) is included as the most relevant baseline, as it is the latest open-source LM-based framework specifically optimized for unified and iterative speech editing; CosyVoice 2 (Du et al., 2024) and IndexTTS 2 (Zhou et al., 2025) which are leading open-source models for instruction-based multi-style emotional synthesis.

We evaluate synthesized speech using four objective metrics:

Word Error Rate (WER): assesses the intelligibility by comparing the transcription of the generated audio from a Conformer-Transducer ASR model (Gulati et al., 2020) with ground-truth text.

Speaker Similarity (SIM): measures cosine similarity between speaker embeddings extracted from the reference speech and the synthesized speech using WavLM-TDNN[5] (Chen et al., 2022), indicating preservation of speaker identity.

DNSMOS: evaluates overall perceptual audio quality using a non-intrusive DNSMOS model (Reddy et al., 2021) trained on human ratings collected following the ITU-T P.808 protocol, with scores from 1 to 5.

Emotion Classification Accuracy (ECA): measures correctness of emotion expression using a WavLM-based classifier (Goncalves et al., 2024), with higher accuracy indicating stronger emotion controllability.controllability.

6 Evaluation Results

6.1 Objective Evaluation

Table 1 summarizes the objective results, showing zero-shot TTS performance in the upper section and emotion editing in the lower section, with the best-performing values highlighted in bold and the second-best underlined.

Zero-shot TTS. We first evaluate zero-shot TTS performance on the LibriSpeech test clean set, with comparisons to baselines reported in Table 1. Results marked with † are cited from (Chen et al., 2025a), focusing on ablation settings comparable training data scales. We follow the original evaluation protocol by performing five times samplings per utterance and reporting the final result by jointly ranking speaker similarity and WER. Under a restricted training budget of less than 1k hours, SpeechEdit achieves a WER of 1.9%, outperforming VALL-E-*A1* and VALL-E-*A2*. Compared to Step-Audio-EditX, our model uses much less training data and fewer parameters, yet achieves substantially higher perceptual quality. Speaker similarity is slightly lower than some baselines, which is expected given the expressive prosody and diversity of LibriEdit, but overall the model maintains a strong balance between intelligibility, speaker identity, and perceptual quality under limited data.

To comprehensively evaluate the model's capability in emotion editing, we designed two experimental setups based on the relationship between the speech prompt and the target emotion: *(1) Easy Task:* uses neutral prompts, presenting no emotional conflict with the target. It includes 80 test samples from 4 unseen speakers in the Step-audio-EditX benchmark, where the target emotions are balanced across the four non-neutral emotion categories. *(2) Hard Task:* includes prompts with conflicting emotions in 80% of the cases, using 100 samples from 4 unseen speakers in the LibriEdit dataset, with five target emotions roughly balanced.

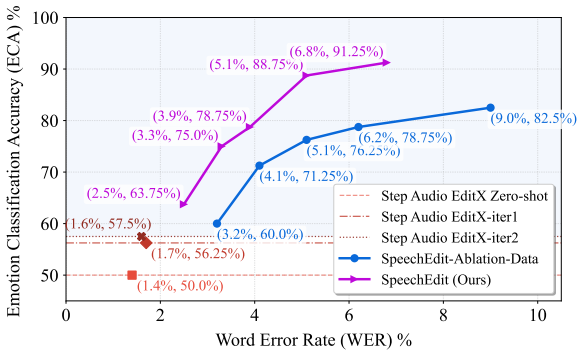

Figure 4: Emotion editing performance on the easy task.

Emotion Edit. In the *Easy Task*, Table 1 reports the results for Step-Audio-EditX iterative editing and SpeechEdit, with SpeechEdit achieving the best performance across all metrics except WER. Figure 4 further visualizes the relationship between WER and ECA. Step-Audio-EditX does not support direct emotion-conditioned generation and instead performs zero-shot TTS ($iter_0$) followed by iterative emotion editing with the speech content fixed. Zero-shot generation achieves 50% ECA, and iterative editing increases it only slightly to 56.25% and 57.5%, with negligible gain from the second iteration, indicating limited emotion controllability. In contrast, SpeechEdit performs direct emotion-controlled generation in a single stage. For each utterance, we generate five samples independently under a fixed inference configura-

[5] https://github.com/microsoft/UniSpeech/tree/main/downstreams/speaker_verification#pre-trained-models

Table 1: Overall objective performance comparison including zero-shot TTS results on the LibriSpeech test-clean set and emotion editing results under different task settings, with the best-performing values highlighted in bold and the second-best underlined.

Task	Model	Params	# / h	WER (%) ↓	SIM ↑	DNSMOS ↑	ECA (%) ↑
Zero-shot TTS	Step-Audio-EditX	3 B	-	1.6	**0.63**	3.32	-
	VALL-E-*A1*†	0.5 B	5 k	2.1	0.61	4.00	-
	VALL-E-*A2*†	0.5 B	1 k	2.7	0.48	**4.02**	-
	SpeechEdit	0.5 B	0.8 k*	**1.3**	0.48	4.00	-
	SpeechEdit-*Ablation-Data*	0.5 B	0.8 k*	1.9	0.45	4.01	-
	SpeechEdit-*Ablation-Task*	0.5 B	0.8 k*	1.5	0.53	**4.02**	-
Emotion Easy Task	Step-Audio-EditX-$iter_0$	3 B	-	1.4	0.49	3.39	50.00
	Step-Audio-EditX-$iter_1$	3 B	-	1.7	0.42	3.34	56.25
	Step-Audio-EditX-$iter_2$	3 B	-	1.6	0.36	3.29	57.50
	CosyVoice 2	0.5 B	<1.5 k*	4.1	**0.52**	**4.01**	43.75
	IndexTTS 2	1.5 B	135*	2.5	0.44	3.72	56.25
	SpeechEdit-*C1*	0.5 B	129*	2.5	0.45	**4.01**	63.75
	SpeechEdit-*C2*	0.5 B	129*	3.9	0.37	3.98	78.75
	SpeechEdit-*C3*	0.5 B	129*	6.8	0.25	4.00	**91.25**
	SpeechEdit-*Ablation-Data-C1*	0.5 B	129*	3.2	0.40	4.00	60.00
	SpeechEdit-*Ablation-Data-C2*	0.5 B	129*	5.1	0.30	3.93	76.25
	SpeechEdit-*Ablation-Data-C3*	0.5 B	129*	9.0	0.21	3.89	82.50
Emotion Hard Task	CosyVoice 2	0.5 B	<1.5 k*	5.8	**0.40**	3.70	79.00
	IndexTTS 2	1.5 B	135*	**2.0**	0.39	3.38	73.00
	SpeechEdit	0.5 B	129*	2.5	0.33	**4.03**	**92.00**
	SpeechEdit-*Ablation-Data*	0.5 B	129*	3.7	0.33	3.83	**92.00**

Params refers to the number of parameters in the AR model. **# / h** indicates the amount of training data in hours.
* Indicates the amount of task-specific training data used after model initialization.

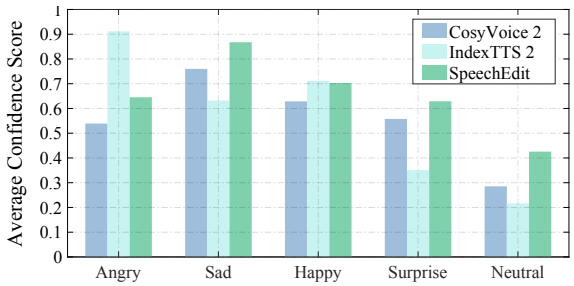

Figure 5: Average classification confidence scores for correctly predicted samples across five emotions.

tion. Selecting only the sample with the lowest WER, SpeechEdit achieves 63.75% ECA at an average WER of 2.5%, already surpassing Step-Audio-EditX in emotion expression. Including samples with higher WER, ECA rises monotonically to 75% at 3.3% WER and 91.25% at 6.8% WER, illustrating stronger emotion controllability and a clear trade-off between content fidelity and emotional expressiveness, which is consistent with the fact that automatic speech recognition models tend to be less accurate on emotional or expressive speech.

In the *Hard Task*, While IndexTTS 2 yields the lowest WER, SpeechEdit maintains a competitive WER of 2.5%. SpeechEdit achieves an ECA of 92%, substantially outperforming CosyVoice 2 (79%) and IndexTTS 2 (73%), indicating its ability to suppress the original emotional content from the prompt and accurately reconstruct the target emotion. It also achieves the highest DNSMOS, reflecting superior perceptual quality. Speaker similarity is slightly lower than the baselines, which is expected since SIM is computed with respect to the prompt speech. Stronger emotion modifications can alter emotion-related acoustic characteristics, naturally affecting similarity scores even when speaker identity is largely preserved. In addition, most baseline systems adopt flow-matching-based continuous-domain modeling in the second stage, which may contribute to better preservation of fine-grained acoustic details.

To further analyze emotion expression, we compute the average SER classification confidence for samples correctly generated with the target emotion. Higher confidence indicates stronger and more distinguishable emotion expression. As shown in Figure 5, SpeechEdit consistently achieves higher confidence than CosyVoice 2 across all five emotion categories. IndexTTS 2 attains particularly high confidence in the Angry category (0.91 versus SpeechEdit's 0.64), indicating especially strong angry expression for this baseline. Notably, SpeechEdit's confidence varies across emotions, with the highest for Sad, followed by Happy, Angry, and Surprise. This ordering aligns well with the distribution of emotions in the Lib-

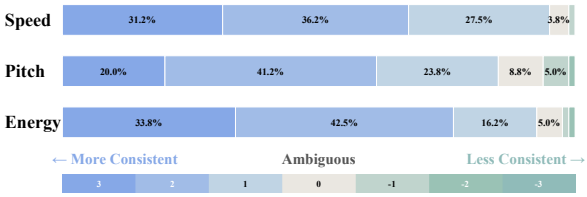

Figure 6: Result of a CMOS-style subjective test on three prosody attributes: speed, pitch, and energy.

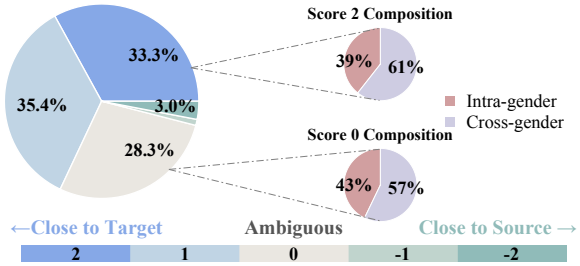

Figure 7: Result of a CMOS-style subjective test on voice conversion.

riEdit dataset, where Sad is most frequent. While this trend may partially reflect differences in perceptual salience across emotions, it also highlights the influence of data scale in emotion expressiveness, suggesting that increasing training data for underrepresented emotions, such as Angry, could further enhance emotion editing performance.

6.2 Subjective Evaluation

We assess the model's ability to follow style-control instructions through a subjective test on speed, pitch, and energy. For each test case, two speech samples are generated from the same source audio under opposite control specifications of a given prosodic attribute, such as low versus high pitch, while keeping all other factors unchanged. Ten listeners compare each pair using a comparative mean opinion score (CMOS), where +3 indicates strong consistency with the target specification, -3 indicates clear inconsistency, and 0 denotes ambiguous perception. In addition, subjective mean opinion score (SMOS) and subjective speaker similarity (SSIM) are evaluated. As shown in Figure 6, over 85% of the samples across all three attributes are rated consistent with the intended control direction, indicating reliable controllability. Energy control achieves the highest proportion of +3 scores (33.8%), followed by speed (31.2%) and pitch (20%). SMOS with details in the Figure 8 show that overall speech naturalness remains high, with average above 4.2. SSIM is best preserved under speed control, while pitch and energy manipulations result in slightly lower similarity and higher variance, reflecting the greater perceptual impact of these controls on speaker-related acoustic cues.

To evaluate the voice conversion capability of the proposed model, we conduct a subjective timbre similarity evaluation with four speakers, including two male and two female speakers. To ensure reliable comparison, intra-gender pairs with clearly distinct timbres are selected. A CMOS protocol is adopted, where listeners rate each sample on a five-point scale from -2 to +2. Negative scores indicate closer similarity to the source speaker, positive scores indicate closer similarity to the target speaker, and a score of 0 denotes an ambiguous identity. Figure 7 summarizes the evaluation results. Only 3.0% of samples receive negative scores, indicating that source speaker leakage is rare. Most samples, accounting for 68.7%, obtain positive scores, showing that the generated speech is generally perceived as closer to the target speaker. The remaining 28.3% of samples are rated as ambiguous. When further grouped by conversion type, 61% of the samples with the highest score and 57% of the ambiguous samples come from cross-gender conversion cases. This suggests that cross-gender conversion more readily departs from the source identity, whereas capturing fine-grained target characteristics across genders remains more difficult, sometimes leading to an intermediate or non-target timbre.

7 Conclusion

We presented SpeechEdit, a unified codec-LM framework for selective speech attribute editing that preserves the reference prompt's acoustic profile while modifying only user-specified attributes. Furthermore, we constructed the LibriEdit dataset and introduced a Delta-Pairs sampling strategy to generate difference-aware training triplets, facilitating implicit disentanglement of speaker identity, prosody, and emotion without requiring specialized architectural modules. Experiments across zero-shot TTS, voice conversion, and style editing show that SpeechEdit delivers strong naturalness, robustness, and state-of-the-art selective control, suggesting that in-context learning in neural codec LMs offers a promising direction for selective and partially disentangled speech generation.

Limitations

Despite the promising results, SpeechEdit has several limitations that warrant further investigation. First, the granularity of speaker modeling remains a challenge. We currently employ a global speaker embedding to represent identity. While effective, this static representation may fail to capture time-varying vocal nuances or idiosyncratic articulation patterns, occasionally leading to a loss of fine-grained timbre during voice conversion. Second, the model relies entirely on implicit disentanglement without explicit supervision. Unlike systems that employ auxiliary losses such as emotion classification or pitch regression, or use reinforcement learning to guide attribute control, SpeechEdit depends solely on in-context learning from contrastive pairs, which may limit robustness in extreme or rare attribute combinations. Third, the current controllable space is limited to emotion, prosody, and speaker identity, restricting more flexible or natural interactions, such as natural-language-based control or multi-attribute specifications. Expanding the controllable scope could enable richer and more expressive speech editing.

References

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

Sanyuan Chen, Shujie Liu, Long Zhou, Yanqing Liu, Xu Tan, Jinyu Li, Sheng Zhao, Yao Qian, and Furu Wei. 2024. Vall-e 2: Neural codec language models are human parity zero-shot text to speech synthesizers. *arXiv preprint arXiv:2406.05370*.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, and 1 others. 2022. WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Sanyuan Chen, Chengyi Wang, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, and 1 others. 2025a. Neural codec language models are zero-shot text to speech synthesizers. *IEEE Trans. Acoust., Speech, Signal Process.*

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

Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. 2022. High fidelity neural audio compression. *arXiv preprint arXiv:2210.13438*.

Zhihao Du, Qian Chen, Shiliang Zhang, Kai Hu, Heng Lu, Yexin Yang, Hangrui Hu, Siqi Zheng, Yue Gu, Ziyang Ma, and 1 others. 2024. Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens. *arXiv preprint arXiv:2407.05407*.

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

Xiaoxue Gao, Chen Zhang, Yiming Chen, Huayun Zhang, and Nancy F Chen. 2025. Emo-dpo: Controllable emotional speech synthesis through direct preference optimization. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Sreyan Ghosh, Arushi Goel, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang-gil Lee, Chao-Han Huck Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, and Bryan Catanzaro. 2025. Audio flamingo 3: Advancing audio intelligence with fully open large audio language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Lucas Goncalves, Ali N Salman, Abinay R Naini, Laureano Moro Velazquez, Thomas Thebaud, Leibny Paola Garcia, Najim Dehak, Berrak Sisman, and Carlos Busso. 2024. Odyssey 2024-speech emotion recognition challenge: Dataset, baseline framework, and results. *Development*, 10(9,290):4–54.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and 1 others. 2020. Conformer: Convolution-augmented transformer for speech recognition. In *Proceedings of Interspeech 2020*, pages 5036–5040.

Zhifang Guo, Yichong Leng, Yihan Wu, Sheng Zhao, and Xu Tan. 2023. Prompttts: Controllable text-to-speech with text descriptions. In *Proc. IEEE ICASSP*, pages 1–5. IEEE.

Bing Han, Long Zhou, Shujie Liu, Sanyuan Chen, Lingwei Meng, Yanming Qian, Yanqing Liu, Sheng Zhao, Jinyu Li, and Furu Wei. 2024. Vall-e r: Robust and efficient zero-shot text-to-speech synthesis via monotonic alignment. In *Conference on Neural Information Processing Systems*.

Shengpeng Ji, Jialong Zuo, Minghui Fang, Ziyue Jiang, Feiyang Chen, Xinyu Duan, Baoxing Huai, and Zhou Zhao. 2024. Textrolspeech: A text style control speech corpus with codec language text-to-speech models. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10301–10305. IEEE.

Zeqian Ju, Yuancheng Wang, Kai Shen, Xu Tan, Detai Xin, Dongchao Yang, Yanqing Liu, Yichong Leng, Kaitao Song, Siliang Tang, and 1 others. 2024. Naturalspeech 3: Zero-shot speech synthesis with factorized codec and diffusion models. *arXiv preprint arXiv:2403.03100*.

Wei Kang, Xiaoyu Yang, Zengwei Yao, Fangjun Kuang, Yifan Yang, Liyong Guo, Long Lin, and Daniel Povey. 2024. Libriheavy: A 50,000 hours asr corpus with punctuation casing and context. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10991–10995. IEEE.

Jaehyeon Kim, Keon Lee, Seungjun Chung, and Jaewoong Cho. 2024. CLam-TTS: Improving neural codec language model for zero-shot text-to-speech. In *The Twelfth International Conference on Learning Representations*.

Mateusz Łajszczak, Guillermo Cámbara, Yang Li, Fatih Beyhan, Arent Van Korlaar, Fan Yang, Arnaud Joly, Álvaro Martín-Cortinas, Ammar Abbas, Adam Michalski, and 1 others. 2024. Base tts: Lessons from building a billion-parameter text-to-speech model on 100k hours of data. *arXiv preprint arXiv:2402.08093*.

Huan Liao, Qinke Ni, Yuancheng Wang, Yiheng Lu, Haoyue Zhan, Pengyuan Xie, Qiang Zhang, and Zhizheng Wu. 2025. Nvspeech: An integrated and scalable pipeline for human-like speech modeling with paralinguistic vocalizations. *arXiv preprint arXiv:2508.04195*.

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

Lingwei Meng, Long Zhou, Shujie Liu, Sanyuan Chen, Bing Han, Shujie Hu, Yanqing Liu, Jinyu Li, Sheng Zhao, Xixin Wu, and 1 others. 2025. Autoregressive speech synthesis without vector quantization. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1287–1300.

Tu Anh Nguyen, Wei-Ning Hsu, Antony d'Avirro, Bowen Shi, Itai Gat, Maryam Fazel-Zarani, Tal Remez, Jade Copet, Gabriel Synnaeve, Michael Hassid, and 1 others. 2023. Expresso: A benchmark and analysis of discrete expressive speech resynthesis. *arXiv preprint arXiv:2308.05725*.

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 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6493–6497. IEEE.

Julius Richter, Yi-Chiao Wu, Steven Krenn, Simon Welker, Bunlong Lay, Shinji Watanabe, Alexander Richard, and Timo Gerkmann. 2024. Ears: An anechoic fullband speech dataset benchmarked for speech enhancement and dereverberation. *arXiv preprint arXiv:2406.06185*.

Yakun Song, Zhuo Chen, Xiaofei Wang, Ziyang Ma, and Xie Chen. 2025. Ella-v: Stable neural codec language modeling with alignment-guided sequence reordering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 25174–25182.

Helin Wang, Jiarui Hai, Dading Chong, Karan Thakkar, Tiantian Feng, Dongchao Yang, Junhyeok Lee, Thomas Thebaud, Laureano Moro Velazquez, Jesus Villalba, and 1 others. 2025a. Capspeech: Enabling downstream applications in style-captioned text-to-speech. *arXiv preprint arXiv:2506.02863*.

Hui Wang, Shujie Liu, Lingwei Meng, Jinyu Li, Yifan Yang, Shiwan Zhao, Haiyang Sun, Yanqing Liu, Haoqin Sun, Jiaming Zhou, and 1 others. 2025b. Felle: Autoregressive speech synthesis with token-wise coarse-to-fine flow matching. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 10229–10238.

Xinsheng Wang, Mingqi Jiang, Ziyang Ma, Ziyu Zhang, Songxiang Liu, Linqin Li, Zheng Liang, Qixi Zheng, Rui Wang, Xiaoqin Feng, and 1 others. 2025c. Spark-tts: An efficient llm-based text-to-speech model with single-stream decoupled speech tokens. *arXiv preprint arXiv:2503.01710*.

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 *The Thirteenth International Conference on Learning Representations*.

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.

Chao Yan, Boyong Wu, Peng Yang, Pengfei Tan, Guoqiang Hu, Li Xie, Yuxin Zhang, Fei Tian, Xuerui Yang, Xiangyu Zhang, and 1 others. 2025. Step-audio-editx technical report. *arXiv preprint arXiv:2511.03601*.

Guanrou Yang, Chen Yang, Qian Chen, Ziyang Ma, Wenxi Chen, Wen Wang, Tianrui Wang, Yifan

Yang, Zhikang Niu, Wenrui Liu, and 1 others. 2025a. Emovoice: Llm-based emotional text-to-speech model with freestyle text prompting. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 10748–10757.

Yifan Yang, Shujie Liu, Jinyu Li, Yuxuan Hu, Haibin Wu, Hui Wang, Jianwei Yu, Lingwei Meng, Haiyang Sun, Yanqing Liu, and 1 others. 2025b. Pseudo-autoregressive neural codec language models for efficient zero-shot text-to-speech synthesis. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 9316–9325.

Xueyao Zhang, Xiaohui Zhang, Kainan Peng, Zhenyu Tang, Vimal Manohar, Yingru Liu, Jeff Hwang, Dangna Li, Yuhao Wang, Julian Chan, and 1 others. Vevo: Controllable zero-shot voice imitation with self-supervised disentanglement. In *The Thirteenth International Conference on Learning Representations*.

Siyi Zhou, Yiquan Zhou, Yi He, Xun Zhou, Jinchao Wang, Wei Deng, and Jingchen Shu. 2025. Indextts2: A breakthrough in emotionally expressive and duration-controlled auto-regressive zero-shot text-to-speech. *arXiv preprint arXiv:2506.21619*.

Yixuan Zhou, Xiaoyu Qin, Zeyu Jin, Shuoyi Zhou, Shun Lei, Songtao Zhou, Zhiyong Wu, and Jia Jia. 2024. Voxinstruct: Expressive human instruction-to-speech generation with unified multilingual codec language modelling. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 554–563.

A Appendix

A.1 Prompt of Audio Flamingo 3

Task.
Identify the emotion in the utterance. Analyze ONLY the speaker's vocal emotion (prosody and tone), strictly ignoring the linguistic content.

Emotion Categories.
Classify the emotion into one of: {*Neutral*, *Happy*, *Sad*, *Angry*, *Surprise*}.

Confidence Estimation.
Provide a confidence score representing how certain you are about your prediction. The confidence should be a floating-point number between *0.0* and *1.0*, where *0.0* indicates complete uncertainty and *1.0* indicates complete certainty.

Output Format.
Return the result **strictly** as a JSON object with two keys:

- *"emotion"*: the predicted emotion category,
- *"confidence"*: the confidence score (*0.0–1.0*).

Do NOT include any explanations, commentary, or extra text outside this JSON object.

A.2 Subjective Evaluation

Beyond the average scores reported in the main text, Figure 8 illustrates the detailed score distributions for three key prosodic attributes: Speed, Pitch, and Energy. To ensure the reproducibility and consistency of our subjective testing, Table 3 explicitly outlines the 5-point scoring criteria used for both the Subjective Mean Opinion Score (SMOS) and Subjective Speaker Similarity (SSIM).

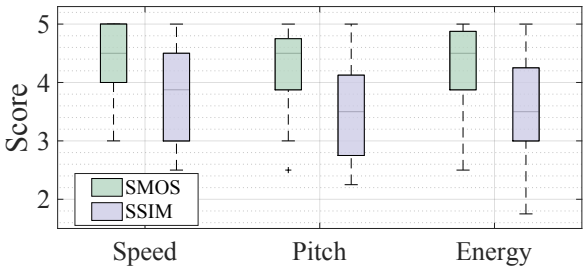

Figure 8: Score distributions of SMOS and SSIM for Speed, Pitch, and Energy.

A.3 Ablation Study

We conduct ablation studies from two perspectives: (i) the training data composition and (ii) the unified task formulation.

Data Ablation. Following prior works that adopt mixed training on collected emotional speech to enhance controllability, we investigate whether emotional data augmentation improves SpeechEdit,

Table 2: Comparison of open-sourced speech datasets in terms of fine-grained style control speech synthesis.

Dataset	Source	Speaker ID	Fine-Grained Feature Types				Duration (h)
			Emotion	Speed	Volume	Pitch	
EmoVoice-DB (Yang et al., 2025a)	Synthesis	✓	✓	✗	✗	✗	40
VoxBox (Wang et al., 2025c)	Collect	✗	✗	✓	✓	✗	102.5k
CapSpeech (Wang et al., 2025a)		✗	✓	✓	✓	✓	33.6k
TextrolSpeech (Ji et al., 2024)		✗	✓	✓	✓	✓	300
Expresso (Nguyen et al., 2023)	Record	✓	✓	✗	✗	✗	47
EARS (Richter et al., 2024)		✓	✓	✓	✓	✓	60
LibriEdit (Ours)	Audiobooks	✓	✓	✓	✓	✓	700

Table 3: Evaluation criteria for SMOS and SSIM.

Metric	Score	Description
SMOS	5	Excellent; natural and clear quality.
	4	Good; minor flaws or barely perceptible noise.
	3	Fair; perceptible degradation but intelligible.
	2	Poor; very annoying or unpleasant to listen to.
	1	Bad; unintelligible or totally corrupted.
SSIM	5	Identical; sounds exactly like the target speaker.
	4	Very Similar; confident it is the same speaker.
	3	Similar; sounds like the target but with noticeable differences.
	2	Different; sounds like a different person.
	1	Totally Different; no resemblance to the target speaker.

denoting this mixed-training variant as SpeechEdit-*Ablation-data*. Specifically, we train SpeechEdit on a mixture of the annotated LibriEdit corpus, an internal emotional speech dataset, and the Expresso dataset (Nguyen et al., 2023). The internal dataset contains approximately 30 h of acted emotional speech, while Expresso contributes an additional 5 h of professionally recorded expressive speech. In total, the training set comprises 743 h of speech. Same-speaker and cross-speaker delta pair sampling are equally balanced.

Table 1 (below the dashed line) reports the ablation results across three tasks. Contrary to expectations, mixed training with additional emotional speech leads to consistent degradation across most objective metrics. Despite the stronger emotional expressions in the internal and Expresso datasets, Emotion-Easy task shows no improvement, while Emotion-Hard task remains comparable to the default SpeechEdit. We attribute this to data distribution mismatch and imbalance: LibriEdit contains spontaneous emotional expressions in read speech, whereas the internal and Expresso datasets comprise elicited, exaggerated emotions. This mismatch introduces a distribution shift that adversely affects training stability and weakens generalization to subtle emotional variations emphasized in Emotion-Easy. Moreover, the relatively limited scale of elicited emotional data leads to an imbalanced optimization signal, causing the model to bias toward salient emotional cues without improving fine-grained emotional controllability. These results suggest that naive emotional data augmentation via mixed training is insufficient, and that better-aligned emotional distributions and sampling strategies are essential for controllable speech editing.

Taks Ablation We observe that SpeechEdit exhibits slightly inferior speaker similarity, raising the concern that the inclusion of the voice conversion task may affect similarity preservation. To examine this effect, we train a task-ablated variant, denoted as SpeechEdit-*Ablation-Task*, on the combined dataset using same-speaker delta pair sampling only, thereby removing cross-speaker supervision. As shown in Table 1 under the zero-shot TTS setting, this ablated model yields a non-negligible improvement in speaker similarity from 0.45 to 0.53, indicating that the voice conversion objective introduces an inherent trade-off between identity preservation and cross-speaker controllability. While the proposed speaker-embedding-based control mechanism effectively supports voice conversion, qualitative results in Figures 7 and 8 show that the generated speech may reflect blended characteristics of the source and target speakers. This observation suggests that how to represent speaker identity in a controllable and robust manner remains an open research question. More expressive and structured speaker representations may further improve conversion fidelity while preserving high speaker similarity.