SE-Eval: A Multi-Dimensional Benchmark Dataset and Automatic Scorer for Speech Editing Evaluation

Anonymous ACL submission

Abstract

Speech editing has advanced rapidly, yet systematic evaluation of editing quality remains an open problem. Metrics designed for text-to-speech synthesis fail to capture edit-specific phenomena such as boundary artifacts and contextual inconsistency. We introduce **SE-Eval**, the first large-scale benchmark for speech editing evaluation, comprising 9,151 edited audio clips from 10 models across five sub-domains, together with 44,451 human ratings along five dimensions: MOS, Boundary MOS, Environment Consistency, Prosody Consistency, and Emotion Consistency. Building on this dataset, we develop an automatic scorer that feeds both original and edited audio, along with their transcriptions, into Qwen2.5-Omni and maps the resulting hidden states to continuous scores via a linear regression head. Under both random split and Leave-One-Model-Out protocols, our scorer reduces MSE by up to 41.8% and 40.0% over the UTMOS baseline, respectively. Ablation experiments confirm the contributions of the linear head and text input. The dataset is publicly available.[1]

1 Introduction

Speech Editing (SE) modifies a localized segment of an utterance to match a new target text while leaving the surrounding audio untouched. Recent work based on diffusion models (Jiang et al., 2023; Eskimez et al., 2024; Chen et al., 2025, 2024), language models (Peng et al., 2024; Wang et al., 2025, 2023a), and hybrid architectures has made this task increasingly practical for applications such as podcast post-production, stutter removal, and dialogue repair. Yet reliable evaluation of such systems remains an open problem.

Three classes of existing approaches each fall short in different ways.

Traditional objective metrics (WER, speaker similarity) operate only on the edited segment and ignore the core challenge of speech editing: seamless integration with surrounding context. A phonetically correct but acoustically jarring edit would score well on WER yet sound unnatural.

Automatic MOS predictors (Lo et al., 2019; Saeki et al., 2022), trained on fully synthesized TTS utterances, treat edited audio as if it were generated from scratch, making them blind to boundary quality and contextual coherence. Because the unedited portions of the signal are real recordings, these models encounter a distribution mismatch that undermines their predictions.

Human listening tests, while directly capturing perceptual quality, are expensive, difficult to reproduce, and lack standardized criteria for edit-specific artifacts. Different research groups use different scales, instructions, and annotator pools, making cross-paper comparison unreliable.

To address these shortcomings, we design a dedicated evaluation framework for speech editing built on five complementary dimensions—*MOS* for overall naturalness, *Boundary MOS* for transition quality at edit points, and three consistency measures (*Environment*, *Prosody*, *Emotion*) for alignment between edited and surrounding segments. Our contributions are:

1. We formalize multi-dimensional automatic evaluation for speech editing and propose a five-dimensional assessment framework tailored to this task.

2. We construct **SE-Eval**, a benchmark containing 9,151 edited clips with 44,451 human ratings independently scored by multiple speech quality assessment experts in a back-to-back setup, covering 10 models and five sub-domains.

3. We develop a Qwen2.5-Omni-based auto-

[1] https://huggingface.co/datasets/SE-Eval/SE-Eval

matic scorer and evaluate it under two protocols (random split and Leave-One-Model-Out), establishing baselines that outperform UTMOS across most dimensions.

2 Related Work

2.1 Speech Editing

Speech editing systems can be grouped into three paradigms.

Diffusion-based models refine mel-spectrograms from noise conditioned on context. FluentSpeech (Jiang et al., 2023) uses masked duration and pitch predictors for natural transitions. E2 TTS (Eskimez et al., 2024) simplifies the pipeline through non-autoregressive flow matching. F5-TTS (Chen et al., 2025) improves alignment with ConvNeXt blocks and Sway Sampling. DiffEditor (Chen et al., 2024) incorporates BERT embeddings and a first-order difference loss for boundary fluency.

Language-model-based approaches treat editing as token infilling in discrete codec spaces. VoiceCraft (Peng et al., 2024) uses a Transformer decoder with token rearrangement for bidirectional conditioning. SSR-Speech (Wang et al., 2025) adds classifier-free guidance and context-aware decoding to maintain background acoustic consistency.

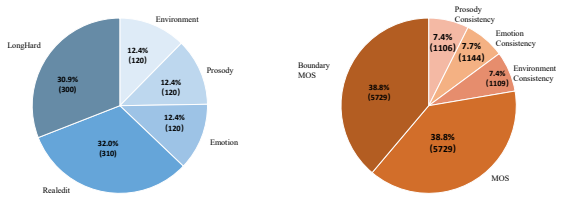
Hybrid and other approaches. PlayDiffusion[2] combines discrete tokens with non-autoregressive diffusion and non-causal attention. Concatenation-based pipelines splice TTS outputs (e.g., MaskGCT (Wang et al., 2024), VALL-E (Wang et al., 2023a)) into the original audio using forced alignment.

Despite the diversity of architectures, there is no consensus on how to evaluate them. Most papers report WER, speaker similarity, and small-scale MOS studies, making it difficult to draw reliable cross-system comparisons.

2.2 Automatic MOS Prediction

Automatic MOS prediction has become essential for scalable evaluation of generative speech. Representative systems exist for speech (Saeki et al., 2022; Lo et al., 2019), audio (Jia et al., 2025), music (Liu et al., 2025), and singing voice (Yao et al., 2025). In the speech domain, automatic MOS prediction has been applied to TTS (Cooper et al., 2022; Wang et al., 2023b), singing voice synthesis (Tang et al., 2024), and speech enhancement

[2] https://github.com/playht/PlayDiffusion



(a) Source Data Triplets (b) Annotated Ratings

Figure 1: Data distribution of SE-Eval: (a) five sub-datasets of source triplets, and (b) subjective ratings across five dimensions.

(Reddy et al., 2021; Mittag et al., 2021). These systems typically train on crowd-sourced or expert ratings of fully synthesized utterances.

The closest work to ours is UTMOS (Saeki et al., 2022), which uses self-supervised features and domain-specific fine-tuning for speech quality prediction. However, UTMOS and all existing MOS predictors share a common limitation: they are designed for signals that are *entirely* synthetic, whereas speech editing produces audio that is *partially* real and *partially* generated. This mismatch, combined with the lack of edit-specific dimensions such as boundary quality, leaves a gap that our work aims to fill.

3 The SE-Eval Dataset

3.1 Overview

SE-Eval contains 24.21 hours of audio across 9,151 edited samples generated by ten models from 970 source triplets (original text, target text, source audio). The triplets span five sub-domains: **RealEdit** (real-world daily usage), **LongHard** (complex long-form edits), **Environment** (ambient noise conditions), **Prosody** (expressive game speech), and **Emotion** (affective speech). Each of the resulting 14,817 evaluation tasks was scored independently by three annotators, yielding 44,451 ratings in total.

As shown in Figure 1, RealEdit and LongHard each account for roughly 30% of the triplets, while the remaining three sub-domains contribute about 12% each. The rating distribution differs because RealEdit and LongHard are assessed on both MOS and Boundary MOS, whereas the other three subsets are evaluated only on their respective consistency dimension.

Figure 2 shows the distribution of edit span lengths. The concentration within 0–10 words mirrors typical real-world editing needs, while the long-tail coverage ensures the benchmark exercises

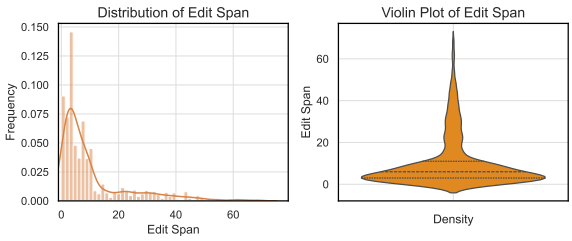


Figure 2: Distribution of edit span lengths (in words). Most samples fall within 0–10 words; long-tail cases exceeding 20 words are also represented.

models on challenging long edits as well.

3.2 Source Data Triplets

RealEdit reuses the open-source VoiceCraft benchmark (Peng et al., 2024), which contains naturally occurring corrections and rephrasing in conversational speech. **LongHard** draws 300 utterances from LibriTTS-clean (Zen et al., 2019), with target texts generated by a large language model to produce challenging edits involving long sentences, domain-specific terms, and atypical modifications. **Environment** covers ten acoustic conditions: 20 conference recordings from AMI (Kraaij et al., 2005), 20 telephonic samples from Earnings-21 (Del Rio et al., 2021), and 80 recordings from NOIZEUS (Hu and Loizou, 2006) spanning eight noise types (Airport, Babble, Car, Exhibition, Restaurant, Station, Street, Train). **Prosody** consists of 120 utterances from 24 Genshin Impact characters (5 each), leveraging expressive game voice-overs with distinctive pitch and rhythm patterns. **Emotion** includes 120 affective samples from IEMOCAP (Busso et al., 2008), covering emotions such as happiness, sadness, anger, and neutrality.

Although 970 triplets $\times$ 10 models would produce 9,700 samples, the actual count is 9,151 (94.4%) due to model-specific limitations in handling certain multi-edit operations and occasional synthesis failures.

3.3 Speech Editing Systems

We select ten open-source models spanning the major architectural families:

(1) Flow-matching / Diffusion NAR: FluentSpeech (Jiang et al., 2023), DiffEditor (Chen et al., 2024), E2-TTS (Eskimez et al., 2024), F5-TTS (Chen et al., 2025). These models employ mask-and-predict training for bidirectional context awareness but depend on external forced alignment at inference.

(2) LM-based AR: VoiceCraft (Peng et al., 2024), SSR-Speech (Wang et al., 2025). Bidirectional prompts enable context-aware infilling despite the causal nature of these models.

(3) End-to-end large speech model: Higgs (Boson AI, 2025), which generates target audio via task-specific prompting without external alignment.

(4) Concatenation-based TTS pipelines: MaskGCT (Wang et al., 2024), VALL-E (Wang et al., 2023a). The original audio serves as an acoustic prompt; edited segments are spliced in using alignment tools.

(5) Hybrid Diffusion-LM: PlayDiffusion, which combines discrete tokens with non-causal attention for bidirectional context awareness.

This selection covers all major paradigms currently used for speech editing, ensuring that the benchmark captures a broad range of artifact types and quality levels.

3.4 Evaluation Dimensions

We define five perceptual dimensions on a 1–5 scale:

MOS rates overall naturalness—whether the output sounds indistinguishable from unedited human speech. **Boundary MOS** focuses on the edit boundaries: annotators receive the audio, its transcript, and marked boundary positions, and judge whether they can detect artifacts, spectral discontinuities, or unnatural silences at those points.

The three **Consistency** dimensions measure alignment between the edited region and its context:

- **Environment:** Are background noise characteristics and reverberation preserved across the edit?
- **Prosody:** Are rhythm, pitch contour, stress, and speaking rate consistent with the surrounding speech?
- **Emotion:** Does the emotion type and intensity remain coherent after editing?

All scores range from 1 (worst) to 5 (best). The first two dimensions (MOS, Boundary MOS) are evaluated on the RealEdit and LongHard subsets; the three consistency dimensions are each evaluated on their corresponding subset (Environment, Prosody, Emotion).

Table 1: Inter-annotator agreement across dimensions.

Dimension	Kripp.'s α	Fleiss' κ	Pearson	±1 Agr.
Emotion Con.	0.707	0.327	0.744	84.3%
Environ. Con.	0.529	0.222	0.671	83.2%
Prosody Con.	0.604	0.251	0.701	80.8%
LongHard (B-MOS)	0.230	0.097	0.413	64.7%
LongHard (MOS)	0.221	0.064	0.426	62.7%
RealEdit (B-MOS)	0.192	0.060	0.463	59.2%
RealEdit (MOS)	0.194	0.049	0.442	53.6%

3.5 Human Annotation Pipeline

Three annotators, each holding at least a bachelor's degree and experienced in speech quality assessment, performed all evaluations on a dedicated cloud platform. Before formal annotation, a calibration session with reference examples was conducted to align scoring standards. The annotators were trained on anchor samples representing each score level (1–5) for each dimension, with explicit descriptions of the perceptual criteria.

Each task presented the annotator with: (1) the edited audio, (2) original and target transcriptions, and (3) visual markers indicating the edit boundaries. Scoring was performed independently in a blind (back-to-back) setup with no access to model identities. The ground-truth score for each sample is the arithmetic mean of the three ratings. Periodic cross-checks and re-annotation of flagged items ensured consistency throughout the campaign. In total, the annotation process required approximately 450 person-hours.

3.6 Inter-Annotator Agreement

We report Krippendorff's α, Fleiss' κ, Pearson correlation, and ±1 agreement rate in Table 1.

The three consistency dimensions achieve $\alpha > 0.5$, with Emotion reaching 0.707 (Pearson 0.744), reflecting strong consensus on perceptually salient attributes. This pattern is intuitive: emotional shifts, background noise changes, and prosodic breaks are relatively easy to detect, even when annotators disagree on the exact severity.

MOS and Boundary MOS yield lower α values (0.19–0.23), consistent with the well-documented subjectivity of overall quality judgments (Cooper et al., 2022). The inherent ambiguity in mapping holistic perceptual impressions to a five-point scale makes exact agreement unlikely. Even so, the ±1 agreement rate remains above 53% in all cases and exceeds 84% for Emotion, confirming directional consistency. These figures are comparable to those reported in prior speech quality annotation campaigns and provide adequate supervision for training regression-based automatic scorers.

4 Automatic Evaluation Framework

4.1 Task Formulation

We cast automatic speech editing evaluation as a regression problem:

$$E = \mathcal{F}(A_{\text{orig}}, T_{\text{orig}}, T_{\text{target}}, A_{\text{edit}}) \qquad (1)$$

where $E = (e_{\text{MOS}}, e_{\text{B-MOS}}, e_{\text{Env}}, e_{\text{Pro}}, e_{\text{Emo}}) \in \mathbb{R}^5$ and each score lies in $[1, 5]$. We choose continuous regression over integer classification because the ground-truth labels are annotator averages and quantizing them would discard useful information.

The four-input design is deliberate. The original audio A_{orig} provides a reference for what the speech *should* sound like in the surrounding context. The text pair $(T_{\text{orig}}, T_{\text{target}})$ tells the model *what was changed*, helping it localize the edit region. The edited audio A_{edit} is the signal to be evaluated. Together, these inputs provide the scorer with sufficient information to assess both local edit quality and global contextual coherence.

4.2 Model Architecture

The architecture (Figure 3) consists of three components:

Multi-modal encoder. Both audio streams (A_{orig}, A_{edit}) and their text transcriptions (T_{orig}, T_{target}) are fed into the Qwen2.5-Omni (Xu et al., 2025) Thinker module. We choose Qwen2.5-Omni because it natively supports multi-audio input and has shown strong paralinguistic perception in prior benchmarks. Its architecture processes audio and text in a shared representation space, enabling alignment of acoustic features with linguistic content.

Task prompt. A dimension-specific guideline is embedded in a unified template: "*You are a speech editing quality evaluation expert. Evaluate the edited speech from the perspective of {Score_Type}: {Guideline}. Score from 1 to 5.*" Varying the prompt allows a single model to produce all five scores. Each guideline describes the perceptual criteria for one dimension, mirroring the instructions given to human annotators.

Linear regression head. Rather than parsing a score from the model's token output, we project the last hidden state of the Thinker module through a linear layer to obtain a continuous prediction. This design choice is motivated by two observations: (1) discrete token output constrains predictions to a

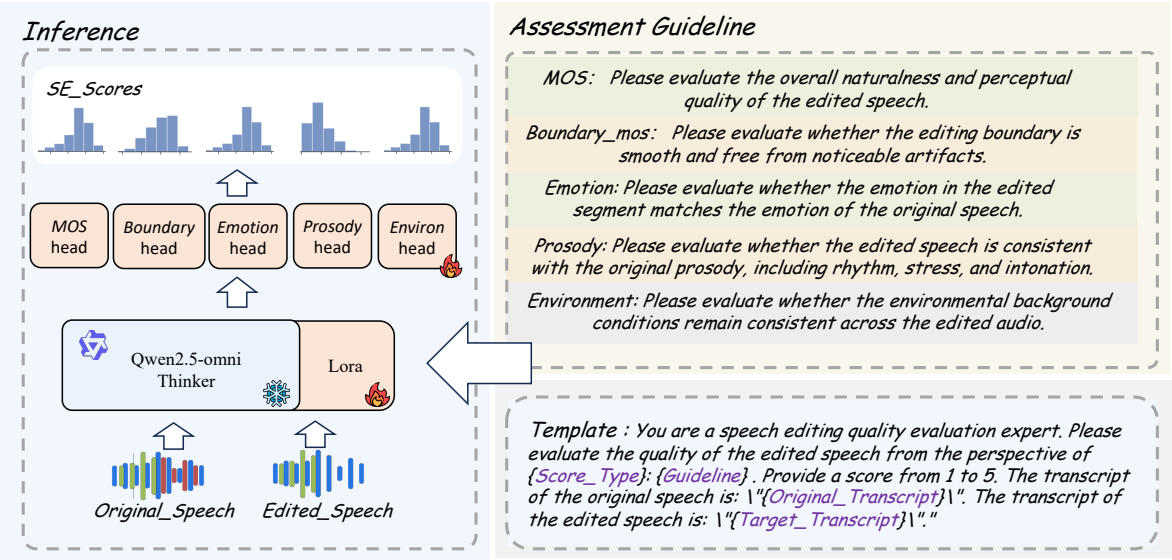


Figure 3: Architecture of the Qwen2.5-Omni-based automatic evaluation framework. Original and edited audio, together with their transcriptions, are encoded by the Thinker module; a linear regression head maps the last hidden state to a continuous score for each evaluation dimension.

small set of values (typically integers 1–5), introducing quantization error when the true label is a non-integer annotator average; (2) the hidden state already encodes rich perceptual information that a simple linear mapping can exploit. The same head is shared across all five dimensions, conditioned on different prompts, enabling multi-task learning with minimal parameter overhead.

4.3 Training

We use LoRA to fine-tune Qwen2.5-Omni-7B while keeping pretrained weights frozen. LoRA inserts low-rank decomposition matrices into the attention layers, reducing the number of trainable parameters while preserving the pretrained model's generalization capacity. We set LoRA rank to 8 and α to 32, applied to all linear layers with the vision encoder frozen. The training objective is MSE loss:

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} (\hat{e}_i - e_i)^2 \tag{2}$$

where $\hat{e}_i$ is the predicted score and e_i is the annotator mean. All five dimensions are trained jointly in a multi-task fashion: within each batch, samples from different dimensions are mixed, and the loss is averaged across all predictions. This joint training encourages the model to develop shared representations for overlapping perceptual phenomena (e.g., boundary artifacts affect both MOS and Boundary MOS) while still specializing through dimension-specific prompts. Training runs for 3 epochs with a learning rate of 10^{-4}, batch size 1 with gradient accumulation of 8, warmup ratio 0.05, and bfloat16 precision, using the ms-swift framework on a single NVIDIA A800 GPU. We select the checkpoint with the lowest validation loss.

5 Experiments

5.1 Human Evaluation Benchmark

Table 2 summarizes human ratings across all ten models and five dimensions.

PlayDiffusion attains the highest MOS on both RealEdit (4.53) and LongHard (4.21), as well as the highest Boundary MOS scores (4.79, 4.45). Its hybrid architecture with non-causal attention allows it to condition on both preceding and following context during generation, which benefits naturalness and boundary smoothness. Among the diffusion-based NAR models, E2-TTS and F5-TTS also perform well, with MOS above 3.79 on both sub-domains. VALL-E and DiffEditor lag behind, with MOS below 2.50 on LongHard—likely due to limited training data and model capacity.

In the consistency dimensions, no single model dominates. SSR-Speech leads in Environment Consistency (3.60) thanks to its context-aware decoding that explicitly preserves background acoustics. Higgs achieves the best Prosody Consistency (3.91), benefiting from end-to-end generation that

Table 2: Multi-dimensional human evaluation benchmark for ten speech editing models on SE-Eval. Results are Mean ± SD. Best result per column in **bold**.

Model	MOS		Boundary MOS		Env. Con.	Pro. Con.	Emo. Con.
	RealEdit	LongHard	RealEdit	LongHard			
FluentSpeech	2.96±0.95	2.53±0.59	3.01±0.80	2.85±0.53	2.24±0.89	3.07±0.87	2.41±0.91
DiffEditor	2.50±0.86	2.26±0.53	2.72±0.65	2.80±0.48	1.65±0.37	2.46±0.73	1.69±0.61
E2-TTS	3.88±0.97	4.01±0.86	4.27±0.73	4.35±0.65	2.39±1.29	3.66±1.02	**3.48**±1.26
F5-TTS	4.07±1.00	3.79±1.04	4.42±0.87	4.25±0.83	2.22±1.28	3.61±1.08	3.21±1.26
Higgs	4.10±0.80	4.03±0.94	4.21±0.98	4.17±0.95	3.11±0.93	**3.91**±1.04	3.34±1.29
MaskGCT	3.69±0.71	3.91±0.90	3.05±0.89	3.30±0.94	3.05±1.09	2.94±1.08	3.02±1.35
VALL-E	1.79±0.66	1.35±0.33	1.98±0.58	1.59±0.45	1.86±0.93	1.82±0.95	1.72±0.91
VoiceCraft	3.86±0.83	3.29±0.91	3.85±0.95	3.69±0.86	3.37±1.04	3.37±1.09	2.93±1.18
SSR-Speech	4.22±0.61	3.13±0.87	4.10±0.66	2.82±0.76	**3.60**±0.83	3.61±1.11	2.90±1.25
PlayDiffusion	**4.53**±0.54	**4.21**±0.76	**4.79**±0.44	**4.45**±0.69	2.90±0.86	3.46±0.98	3.06±1.30

avoids the splicing artifacts of pipeline approaches. E2-TTS scores highest on Emotion Consistency (3.48) through its continuous flow-matching process that maintains smooth affective transitions. Notably, PlayDiffusion's strong MOS and Boundary MOS performance does not carry over to consistency metrics, where it ranks mid-tier (Env: 2.90, Pro: 3.46, Emo: 3.06). This partial orthogonality between naturalness and consistency is one of the key takeaways from our benchmark: MOS alone would fail to distinguish models that sound natural but alter the acoustic environment.

Comparing sub-domains, most models degrade from RealEdit to LongHard. The drop is particularly pronounced for SSR-Speech (MOS: 4.22→3.13; B-MOS: 4.10→2.82), indicating that long, complex edits remain challenging for autoregressive architectures that must generate extended sequences without the ability to attend to future context. PlayDiffusion and Higgs show the smallest degradation, suggesting greater robustness to editing complexity. Concatenation-based models (MaskGCT) exhibit a reverse pattern on MOS (3.69→3.91), likely because their TTS component handles longer utterances well, though their Boundary MOS remains low due to unavoidable splicing discontinuities at edit points.

5.2 Setup for Automatic Scoring

We evaluate automatic scoring under two protocols:

Random Split: 90% of SE-Eval for training, 10% for testing, stratified by model and dimension. This measures standard in-distribution accuracy.

Leave-One-Model-Out: In each fold, all samples from one speech editing model serve as the test set; the remaining nine models form the training set. Results are averaged over all ten folds. This protocol tests whether the scorer can generalize to entirely unseen editing systems—a prerequisite for practical deployment where new models appear frequently.

The baseline is **UTMOS** (Saeki et al., 2022), a speech MOS predictor built on wav2vec 2.0 Base and pretrained on large-scale TTS MOS data. We fine-tune it separately for each dimension on the SE-Eval training set using Adam with a learning rate of 2×10^{-5}, batch size 4, and 3 epochs, selecting the checkpoint with the best validation SRCC.

We report Mean Squared Error (MSE), Linear Correlation Coefficient (LCC), Spearman Rank Correlation (SRCC), and Kendall's Tau (KTAU).

5.3 Random Split Results

Table 3 shows that our scorer outperforms UTMOS on 19 out of 20 metric–dimension combinations.

The gains are largest on dimensions where UTMOS, designed for general TTS quality, has no built-in inductive bias. For Boundary MOS, MSE drops from 0.537 to 0.428 (−20.3%), with LCC rising from 0.788 to 0.830. Emotion Consistency sees an MSE reduction of 41.8% (1.155→0.672), which we attribute to Qwen2.5-Omni's paralinguistic perception acquired during large-scale pretraining. Environment Consistency improves by 30.8% in MSE (0.559→0.387), and Prosody Consistency by 32.7% (0.997→0.671), though absolute Prosody correlations remain the lowest, consistent with the low annotator agreement on this metric.

For MOS, the two systems are competitive; UTMOS achieves a marginally higher LCC (0.811 vs. 0.806), likely because its pretraining data distribution aligns closely with overall speech quality. Our

Table 3: Random Split results. Best per dimension in **bold**.

Dimension	UTMOS (Baseline)				Ours			
	MSE↓	LCC↑	SRCC↑	KTAU↑	MSE↓	LCC↑	SRCC↑	KTAU↑
MOS	0.513	**0.811**	0.745	0.579	**0.499**	0.806	**0.760**	**0.596**
Boundary MOS	0.537	0.788	0.770	0.602	**0.428**	**0.830**	**0.800**	**0.627**
Prosody Con.	0.997	0.521	0.511	0.379	**0.671**	**0.664**	**0.614**	**0.466**
Emotion Con.	1.155	0.675	0.698	0.528	**0.672**	**0.774**	**0.787**	**0.612**
Environ. Con.	0.559	0.760	0.780	0.605	**0.387**	**0.830**	**0.829**	**0.660**

Table 4: Leave-One-Model-Out results. Best per dimension in **bold**.

Dimension	UTMOS (Baseline)				Ours			
	MSE↓	LCC↑	SRCC↑	KTAU↑	MSE↓	LCC↑	SRCC↑	KTAU↑
MOS	0.432	0.529	0.418	0.317	**0.279**	**0.657**	**0.476**	**0.367**
Boundary MOS	0.452	0.529	0.377	0.288	**0.271**	**0.684**	**0.498**	**0.398**
Prosody Con.	**0.724**	0.576	**0.516**	**0.382**	0.877	**0.625**	0.513	0.381
Emotion Con.	1.322	0.500	0.467	0.344	**0.839**	**0.723**	**0.663**	**0.488**
Environ. Con.	0.718	**0.722**	0.643	0.475	**0.453**	0.721	**0.671**	**0.498**

model nonetheless achieves lower MSE (0.499 vs. 0.513) and higher SRCC (0.760 vs. 0.745), indicating stronger rank-order agreement with human judgments overall.

5.4 Leave-One-Model-Out Results

Leave-One-Model-Out (Table 4) is the harder setting, as the test model is absent from training. Our scorer still leads on 16 out of 20 cells. MOS and Boundary MOS show MSE reductions of 35.4% and 40.0%, with LCC gains exceeding 0.12 for both, demonstrating that the model learns general perceptual criteria rather than overfitting to specific model artifacts.

Emotion Consistency transfers well (MSE: 1.322→0.839, SRCC: 0.467→0.663), suggesting that emotional cues are not highly model-specific. Prosody Consistency is the exception: UTMOS achieves a lower MSE (0.724 vs. 0.877) and a slightly higher SRCC, though our model retains an edge in LCC. This result is consistent with the annotation analysis, where Prosody already showed the lowest inter-annotator agreement among the consistency dimensions. We hypothesize that prosodic patterns vary more strongly across editing architectures, making cross-model generalization harder for any learned scorer.

Comparing the two splits, the gap is moderate for Emotion (MSE: 0.672→0.839) and Environment (0.387→0.453), indicating that these dimensions transfer reasonably well. The largest degradation occurs for Prosody (0.671→0.877). These observations offer practical guidance: when deploying automatic scorers on unseen models, Emotion and Environment scores are likely reliable out of the box, while Prosody evaluation may benefit from model-specific adaptation or few-shot calibration.

5.5 Ablation Studies

Linear head vs. token output. Table 5 compares the linear regression head with extracting a score from the model's discrete token output. The linear head achieves lower MSE on all five dimensions. The gains are largest for Emotion Consistency (1.238→0.672, −45.7%) and Environment Consistency (0.640→0.387, −39.6%), confirming that continuous projection better captures the granularity of human ratings than quantized token predictions. Interestingly, the token variant retains slightly higher SRCC and KTAU on MOS, Boundary MOS, and Prosody, indicating that discrete outputs can still rank samples reasonably even when their absolute predictions are less accurate. The MSE advantage of the linear head is nonetheless consistent and often large, making it the better choice when calibrated scores are needed.

Text input. Table 6 examines the role of text transcriptions. Providing text reduces MSE for four of five dimensions, with consistent gains across all four metrics. The benefit is most pronounced for Boundary MOS (MSE: 0.482→0.428, −11.2%) and Environment Consistency (MSE: 0.462→0.387, −16.2%), suggesting that knowing the edit content helps the model localize and evalu-

Table 5: Ablation: linear regression head vs. discrete token output (Random Split). Best per dimension in **bold**.

Dimension	Discrete Token Output				Linear Regression Head			
	MSE↓	LCC↑	SRCC↑	KTAU↑	MSE↓	LCC↑	SRCC↑	KTAU↑
MOS	0.594	0.799	**0.761**	**0.629**	**0.499**	**0.806**	0.760	0.596
Boundary MOS	0.469	0.827	**0.807**	**0.671**	**0.428**	**0.830**	0.800	0.627
Emotion Con.	1.238	0.720	0.738	0.608	**0.672**	**0.774**	**0.787**	**0.612**
Prosody Con.	0.889	**0.674**	**0.635**	**0.500**	**0.671**	0.664	0.614	0.466
Environ. Con.	0.640	0.796	0.805	0.657	**0.387**	**0.830**	**0.829**	**0.660**

Table 6: Ablation: effect of text transcription input (Random Split). Best per dimension in **bold**.

Dimension	w/o Text				w/ Text			
	MSE↓	LCC↑	SRCC↑	KTAU↑	MSE↓	LCC↑	SRCC↑	KTAU↑
MOS	0.515	0.798	0.746	0.581	**0.499**	**0.806**	**0.760**	**0.596**
Boundary MOS	0.482	0.805	0.784	0.611	**0.428**	**0.830**	**0.800**	**0.627**
Emotion Con.	0.752	0.743	0.753	0.579	**0.672**	**0.774**	**0.787**	**0.612**
Prosody Con.	**0.642**	**0.681**	**0.626**	0.461	0.671	0.664	0.614	**0.466**
Environ. Con.	0.462	0.796	0.793	0.620	**0.387**	**0.830**	**0.829**	**0.660**

ate the edit region more precisely. The exception is Prosody Consistency, where the text-free variant performs slightly better across most metrics (MSE 0.642 vs. 0.671, LCC 0.681 vs. 0.664). We attribute this to the fact that prosodic assessment depends primarily on acoustic cues—pitch, rhythm, stress—rather than on what words were changed.

Generalization of ablation findings. We repeat both ablations under Leave-One-Model-Out (Tables 7 and 8 in Appendix A). The linear head advantage is even more pronounced when the test model is unseen: Environment Consistency MSE drops by 69.7% (1.493→0.453) and Emotion Consistency by 43.9% (1.495→0.839). The text input advantage on Boundary MOS is also amplified (-34.3% vs. -11.2% under random split). These results confirm that both design choices generalize robustly to new artifact distributions.

6 Discussion

Two key observations emerge from our experimental analysis.

Naturalness $\neq$ consistency. The benchmark (Table 2) shows that models ranking highest on MOS and Boundary MOS do not necessarily lead on consistency dimensions. PlayDiffusion, for instance, achieves MOS above 4.5 on RealEdit but falls to mid-tier on Environment and Emotion Consistency. This decoupling implies that single-score evaluation protocols—standard practice in most SE papers—can paint an incomplete picture. A model that sounds natural in isolation may still introduce subtle mismatches in background acoustics or emotional tone that only become apparent under targeted assessment.

Where does automatic scoring struggle? Prosody Consistency stands out as the hardest dimension for both humans and machines. Annotator agreement is lowest ($\alpha = 0.604$), and automatic scoring under Leave-One-Model-Out degrades more than for any other dimension. We speculate that prosodic patterns are highly speaker- and style-dependent, leaving less room for cross-model generalization, and that the boundary between "consistent" and "inconsistent" prosody is perceptually ambiguous.

7 Conclusion

We have presented SE-Eval, a dataset of 9,151 edited speech clips with 44,451 human ratings spanning five evaluation dimensions for speech editing. A ten-model benchmark reveals that naturalness and contextual consistency are partially independent, with no single model excelling across all dimensions—confirming the need for multi-dimensional evaluation. Our Qwen2.5-Omni-based scorer, using a linear regression head and multimodal text input, outperforms the UTMOS baseline under both random and Leave-One-Model-Out protocols, providing a practical tool for scalable speech editing assessment. The full dataset and evaluation code are publicly available.

Limitations

First, our prompt design uses fixed templates; exploring alternative wordings and structures may improve scoring accuracy. Second, Prosody Consistency remains the weakest dimension in both inter-annotator agreement and automatic scoring, especially under Leave-One-Model-Out, suggesting that prosodic evaluation may require specialized modeling or model-specific calibration. Third, all source data is in English; multilingual extension would broaden the benchmark's applicability. Fourth, the annotator pool is small; a larger pool would improve the robustness of ground-truth labels. Fifth, our scorer requires access to both original and edited audio together with their transcriptions, inputs that may not always be available in practice.

Ethics Statement

All audio in SE-Eval comes from publicly available research corpora (LibriTTS, AMI, Earnings-21, NOIZEUS, IEMOCAP, Genshin Impact game assets) or is generated by open-source models. No personally identifiable information is collected. Annotators were paid at fair market rates and gave informed consent. The dataset is released under CC BY-NC-SA 4.0.

Speech editing technology carries a risk of misuse for creating deceptive audio. We believe that rigorous evaluation, as facilitated by SE-Eval, is a prerequisite for the responsible development and deployment of such technology.

References

Boson AI. 2025. Higgs Audio V2: Redefining Expressiveness in Audio Generation. https://github.com/boson-ai/higgs-audio.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359.

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*.

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

Erica Cooper, Wen-Chin Huang, Tomoki Toda, and Junichi Yamagishi. 2022. Generalization ability of mos prediction networks. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8442–8446. IEEE.

Miguel Del Rio, Natalie Delworth, Ryan Westerman, Michelle Huang, Nishchal Bhandari, Joseph Palakapilly, Quinten McNamara, Joshua Dong, Piotr Żelasko, and Miguel Jetté. 2021. Earnings-21: A practical benchmark for ASR in the wild. In *Interspeech*, Interspeech 2021, pages 3465–3469. ISCA.

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.

Yi Hu and P. C. Loizou. 2006. Subjective comparison of speech enhancement algorithms. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pages 153–156.

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

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.

Wessel Kraaij, Thomas Hain, Mike Lincoln, and Wilfried Post. 2005. The AMI meeting corpus. In *Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research*.

Cheng Liu, Hui Wang, Jinghua Zhao, Shiwan Zhao, Hui Bu, Xin Xu, Jiaming Zhou, Haoqin Sun, and Yong Qin. 2025. Musiceval: A generative music dataset with expert ratings for automatic text-to-music evaluation. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

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*.

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*.

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.

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.

Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari. 2022. Utmos: Utokyo-sarulab system for voicemos challenge 2022. *Preprint*, arXiv:2204.02152.

Yuxun Tang, Jiatong Shi, Yuning Wu, and Qin Jin. 2024. Singmos: An extensive open-source singing voice dataset for mos prediction. *arXiv preprint arXiv:2406.10911*.

Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, Lei He, Sheng Zhao, and Furu Wei. 2023a. Neural codec language models are zero-shot text to speech synthesizers. *Preprint*, arXiv:2301.02111.

Helin Wang, Meng Yu, Jiarui Hai, Chen Chen, Yuchen Hu, Rilin Chen, Najim Dehak, and Dong Yu. 2025. 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, Shiwan Zhao, Xiguang Zheng, and Yong Qin. 2023b. Ramp: Retrieval-augmented mos prediction via confidence-based dynamic weighting. *arXiv preprint arXiv:2308.16488*.

Yuancheng Wang, Haoyue Zhan, Liwei Liu, Ruihong Zeng, Haotian Guo, Jiachen Zheng, Qiang Zhang, Xueyao Zhang, Shunsi Zhang, and Zhizheng Wu. 2024. Maskgct: Zero-shot text-to-speech with masked generative codec transformer. *Preprint*, arXiv:2409.00750.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. 2025. Qwen2.5-omni technical report. *Preprint*, arXiv:2503.20215.

Jixun Yao, Guobin Ma, Huixin Xue, Huakang Chen, Chunbo Hao, Yuepeng Jiang, Haohe Liu, Ruibin Yuan, Jin Xu, Wei Xue, and 1 others. 2025. Songeval: A benchmark dataset for song aesthetics evaluation. *arXiv preprint arXiv:2505.10793*.

Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J. Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. 2019. Libritts: A corpus derived from librispeech for text-to-speech. *Preprint*, arXiv:1904.02882.

A Leave-One-Model-Out Ablation Results

Linear head vs. token output. Table 7 shows that the linear head advantage persists under Leave-One-Model-Out and is in fact more pronounced. Environment Consistency MSE drops by 69.7% (1.493→0.453) and Emotion Consistency by 43.9% (1.495→0.839), indicating that continuous regression provides substantially better generalization to unseen editing systems than discrete token prediction.

Table 7: Ablation: linear regression head vs. discrete token output under Leave-One-Model-Out. Best per dimension in **bold**.

Dimension	Discrete Token Output				Linear Regression Head			
	MSE↓	LCC↑	SRCC↑	KTAU↑	MSE↓	LCC↑	SRCC↑	KTAU↑
MOS	0.409	0.580	0.412	0.359	**0.279**	**0.657**	**0.476**	**0.367**
Boundary MOS	0.305	0.622	**0.511**	**0.462**	**0.271**	**0.684**	0.498	0.398
Emotion Con.	1.495	0.644	0.633	**0.501**	**0.839**	**0.723**	**0.663**	0.488
Prosody Con.	0.982	0.604	0.478	**0.393**	**0.877**	**0.625**	**0.513**	0.381
Environ. Con.	1.493	0.686	0.634	0.505	**0.453**	**0.721**	**0.671**	**0.498**

Text input. Table 8 presents the text input ablation under Leave-One-Model-Out. Providing text yields lower MSE for Boundary MOS (0.271 vs. 0.412, −34.3%), Prosody Consistency, and Environment Consistency. MOS is the exception, where the text-free variant achieves slightly lower MSE (0.361 vs. 0.279) and higher SRCC, suggesting that overall naturalness can be judged from acoustics alone when the test model is unseen. Compared with the random split (Table 6), the text advantage on Boundary MOS is amplified under Leave-One-Model-Out, confirming that transcription input helps the scorer generalize to unseen editing systems by providing explicit cues about the edit region.

Table 8: Ablation: effect of text transcription input under Leave-One-Model-Out. Best per dimension in **bold**.

Dimension	w/ Text				w/o Text			
	MSE↓	LCC↑	SRCC↑	KTAU↑	MSE↓	LCC↑	SRCC↑	KTAU↑
MOS	**0.279**	0.657	0.476	0.367	0.361	**0.658**	**0.494**	**0.383**
Boundary MOS	**0.271**	**0.684**	0.498	**0.398**	0.412	0.674	**0.500**	0.397
Emotion Con.	**0.839**	**0.723**	0.663	0.488	0.908	0.715	**0.667**	**0.491**
Prosody Con.	**0.877**	**0.625**	**0.513**	**0.381**	0.799	0.614	0.460	0.346
Environ. Con.	**0.453**	**0.721**	**0.671**	**0.498**	0.481	0.701	0.665	0.489