SEMOEDIT: PROBING AND HARNESSING THE EDITABILITY OF PRE-TRAINED SPEECH FLOWS

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing training-based speech emotion editing methods often require substantial task-specific training and can be unstable. This motivates us to investigate whether the pretrained generative dynamics of large-scale text-to-speech (TTS) models can be directly manipulated for training-free emotion editing. To answer this question, we probe the editability of pretrained flow-matching and hybrid TTS models by constructing a controlled test set and systematically diagnosing editing effects along the generative trajectory. Our analysis reveals that pretrained TTS models are substantially editable in emotion, but such editability is architecture- and trajectory-dependent and can be disrupted by early flow-matching steps, while cross-speaker emotion transport carries additional acoustic attributes beyond emotion. To address these limitations, we propose **SEmoEdit**, the first training-free framework that formulates emotion editing as dynamic velocity transport between source and target emotions, enabling robust, flow-based speech emotion editing directly within pretrained TTS models. SEmoEdit unifies three core operations: emotion replacement, emotion erasure, and continuous emotion interpolation, requiring neither parameter updates nor task-specific optimization. To systematically evaluate these capabilities, we introduce **SEmoEditBench**, a dataset comprising 600 editing cases, and conduct extensive experiments across state-of-the-art (SOTA) models and backbones. Our results show that SEmoEdit is highly effective and broadly applicable, outperforming existing training-based and activation-steering methods. Ultimately, this work reveals that pretrained speech flows possess rich, latent emotion-editing capabilities, providing useful guidance for real applications. **Code** and **benchmark** are provided in the supplementary material. **Audio samples** are available at `https://semoedit.pages.dev/`.

1 INTRODUCTION

Flow-matching TTS formulates speech generation as transport along a learned velocity field, enabling efficient non-autoregressive synthesis (Guo et al., 2024; Eskimez et al., 2024; Chen et al., 2025). Large-scale hybrid TTS models further integrate autoregressive generation with flow matching, achieving strong zero-shot voice cloning and expressive synthesis capabilities (Du et al., 2024; Zhou et al., 2026b; Hu et al., 2026). While controllable speech synthesis has been extensively studied (Xie et al., 2025a), these advances raise a natural question beyond generation: *can the capabilities encoded in pre-trained TTS models be directly harnessed to edit the emotion of an existing utterance while preserving its linguistic content and speaker identity?* In this paper, we define three speech emotion editing tasks as proxies for exploring this question: (1) *emotion replacement*, replacing the source emotion with a target emotion; (2) *emotion erasure*, neutralizing the source emotion; and (3) *emotion interpolation*, continuously traversing between source and target emotions.

Existing methods achieve related capabilities primarily through emotion conditioning or activation steering. Conditional generation approaches such as PromptTTS (Guo et al., 2023), EmoSphere++ (Cho et al., 2025), and IndexTTS2 (Zhou et al., 2026b) learn label-, prompt-, or reference-based representations to enable emotional speech generation. Dedicated emotion editing methods, such as dots.tts.edit (Wang et al., 2026a) and Bagpiper-Edit (Gong et al., 2026), further support emotion modification by localizing and regenerating selected spans of the acoustic representation. More recently, activation-steering-based methods, e.g., EmoSteer-TTS (Xie et al., 2025b) and CoCoEmo

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

(Wang et al., 2026b), derive emotion directions from activation differences between neutral and emotional speech and apply vector arithmetic to control emotion without fine-tuning the base model.

However, these approaches face several limitations. (1) They often require substantial resources beyond the pre-trained model. Dedicated editing methods usually depend on large-scale edit-specific datasets for training, while methods such as EmoSteer-TTS rely on paired emotional corpora, auxiliary emotion recognition models, or model-specific probing to derive steering vectors. (2) Although training-free, activation-steering methods are largely static, applying fixed activation directions with global coefficients at selected layers, tokens, or sampling steps, leading to unstable or unsuccessful results. (3) They lack a unified mechanism across editing tasks: different objectives or model architectures often require distinct input/output/model designs, curated datasets, or task-specific training.

To address these challenges and achieve robust training-free emotion editing, we need to answer three questions: *(1) Can pretrained speech flows be directly edited without training? (2) When along the generative trajectory should the editing be applied? (3) What acoustic attributes are actually carried by a source-to-target velocity transport?* Our study yields three observations: First, pretrained speech flows possess substantial direct editability. Second, editability is strongly non-uniform along the trajectory. Third, conditional speech velocity fields are not attribute-factorized.

Guided by these observations, we propose **SEmoEdit**, a training-free framework that recasts speech emotion editing as dynamic velocity transport between source and target emotions. Instead of injecting a fixed global activation direction, SEmoEdit dynamically estimates the source-to-target velocity transport process throughout sampling. The same formulation supports emotion replacement, erasure, and interpolation by controlling the editing strength of the emotional transport, without parameter updates or task-specific optimization. To systematically evaluate our framework, we introduce **SEmoEditBench**, a benchmark containing 600 speech emotion editing cases. We evaluate SEmoEdit on three different TTS backbones (Du et al., 2024; Chen et al., 2025; Zhou et al., 2026b) and compare it directly with leading training-based (Yan et al., 2025b; Wang et al., 2026a) and activation-steering approaches (Xie et al., 2025b; Wang et al., 2026b). Our results consistently demonstrate highly effective emotion editing across diverse models and tasks, revealing that pretrained TTS systems harbor latent emotion-editing capabilities that can be unlocked via inference-time velocity control. Our in-depth analysis of the editability also provides guidance for future research and applications. The contributions of this paper are summarized as follows:

- We present the first systematic study of pretrained speech flows' editability, asking and answering whether, when, and under what conditions an utterance can be directly edited without sacrificing intelligibility, naturalness, speaker identity, or emotional authenticity.
- We propose SEmoEdit, the first training-free framework to formulate speech emotion editing as dynamic velocity transport. By directly manipulating the velocity field, SEmoEdit seamlessly unifies emotion replacement, erasure, and continuous interpolation.
- We introduce SEmoEditBench and evaluate across SOTA models and backbones. Extensive comparisons demonstrate our framework's superiority over existing approaches, proving that robust emotion editing can be unlocked directly within pretrained TTS systems.

2 Analyzing the Editability of Pre-trained Speech Flows

2.1 Preliminaries

Conditional Flow Matching (CFM). Let $\mathbf{x}_1 \in \mathbb{R}^{D\times L}$ be a mel spectrogram and $\mathbf{c}$ its synthesis condition (e.g., text, speaker, paralinguistics). CFM (Lipman et al., 2023) learns a velocity field transporting a prior $p_0(\mathbf{x})$ (Gaussian or semantic tokens) to the speech distribution $p_1(\mathbf{x} \mid \mathbf{c})$. For data $\mathbf{x}_1 \sim p_1(\mathbf{x} \mid \mathbf{c})$ and noise $\mathbf{x}_0 \sim p_0(\mathbf{x})$, the optimal-transport path and conditional velocity are:

$$\mathbf{x}_t = (1-t)\mathbf{x}_0 + t\mathbf{x}_1, \qquad \boldsymbol{u}_t(\mathbf{x}_t \mid \mathbf{x}_1) = \mathbf{x}_1 - \mathbf{x}_0, \qquad t \in [0,1]. \tag{1}$$

A neural velocity field $\boldsymbol{v}_\theta(\mathbf{x}_t, t; \mathbf{c})$ is trained with $t \sim \mathcal{U}[0,1]$ via:

$$\mathcal{L}_{\mathrm{CFM}}(\theta) = \mathbb{E}_{t,\mathbf{x}_1,\mathbf{x}_0}\left[\left\|\boldsymbol{v}_\theta(\mathbf{x}_t, t; \mathbf{c}) - \boldsymbol{u}_t(\mathbf{x}_t \mid \mathbf{x}_1)\right\|_2^2\right]. \tag{2}$$

At inference, a sample is generated by solving the ODE from $\mathbf{x}(0) \sim p_0$:

$$\frac{\mathrm{d}\mathbf{x}(t)}{\mathrm{d}t} = \boldsymbol{v}_\theta(\mathbf{x}(t), t; \mathbf{c}), \qquad \mathbf{x}(1) \sim p_1(\mathbf{x} \mid \mathbf{c}). \tag{3}$$

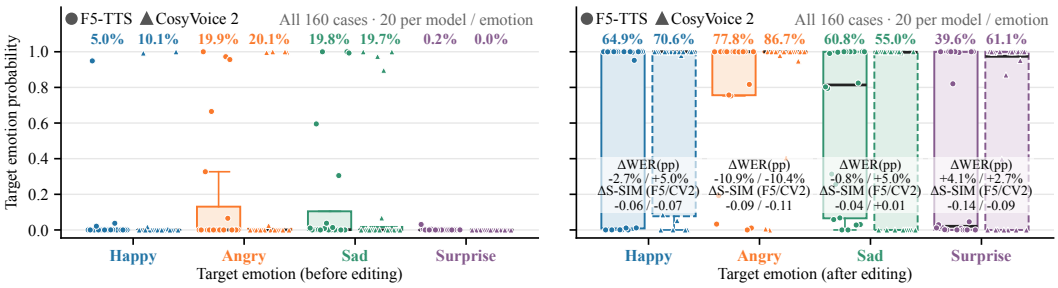


Figure 1: Target emotion2vec (Ma et al., 2024) probability, change in transcription error (ΔWER, in percentage points; pp), and change in speaker similarity (ΔS-SIM) before and after editing.

This formulation underpins modern flow-matching (Mehta et al., 2024; Eskimez et al., 2024; Chen et al., 2025) and hybrid TTS models (Du et al., 2024; Hu et al., 2026).

The Emotion Editing Problem. We formulate ideal speech emotion editing as transforming a source utterance $\mathbf{x}^{\text{src}} \sim p_1(\mathbf{x} \mid \boldsymbol{y}, \boldsymbol{s}, e_{\text{src}})$, conditioned on linguistic content $\boldsymbol{y}$, speaker $\boldsymbol{s}$, and emotion e_{src}, into a target emotion e_{tgt}:

$$\mathbf{x}^{\text{edit}} = \mathcal{E}(\mathbf{x}^{\text{src}}; e_{\text{src}} \to e_{\text{tgt}}), \qquad \mathbf{x}^{\text{edit}} \sim p_1(\mathbf{x} \mid \boldsymbol{y}, \boldsymbol{s}, e_{\text{tgt}}). \tag{4}$$

A perfect edit preserves content $\mathcal{C}(\cdot)$ and speaker $\mathcal{S}(\cdot)$ while exclusively updating the emotion $\mathcal{A}(\cdot)$:

$$\mathcal{C}(\mathbf{x}^{\text{edit}}) = \mathcal{C}(\mathbf{x}^{\text{src}}), \qquad \mathcal{S}(\mathbf{x}^{\text{edit}}) = \mathcal{S}(\mathbf{x}^{\text{src}}), \qquad \mathcal{A}(\mathbf{x}^{\text{edit}}) = e_{\text{tgt}}. \tag{5}$$

Since pre-trained TTS models learn generation under entangled conditions rather than factorized editing operators, we investigate *whether* their learned velocity fields support such editing, *when* the editing signal is effective along the trajectory, and *what* attributes it actually carries.

2.2 The Editability of Flow-Matching and Hybrid TTS Models

Q1: Can pre-trained TTS models be directly edited without additional training?

To test if pre-trained velocity fields contain a usable editing signal, we probe frozen F5-TTS and CosyVoice 2 using 80 parallel neutral-to-emotion pairs (matched text and speaker) from ESD (Zhou et al., 2022), covering 20 speakers and four emotions (Happy, Angry, Sad, Surprise).

We synthesize a source mel $\mathbf{x}^{\text{src}}$, length-match it to the target condition, and construct a diagnostic probe inspired by inversion-free image editing (Kulikov et al., 2025; Xu et al., 2024). The probe uses the conditional velocity difference under shared noise as a local editing signal:

$$\Delta \boldsymbol{v}_\theta(t) = \boldsymbol{v}_\theta\left(\overline{\mathbf{x}}_t^{\text{probe}}, t; \mathbf{c}^{\text{tgt}}\right) - \boldsymbol{v}_\theta(\overline{\mathbf{x}}_t^{\text{src}}, t; \mathbf{c}^{\text{src}}).$$

Accumulating this signal along the flow trajectory effectively shifts $\mathbf{x}^{\text{src}}$ toward the target emotion (complete transport formulation in Sec. 3). As Fig. 1 shows, this substantially increases target-emotion probability across all emotions, albeit with mixed Word Error Rate (WER) changes and moderate speaker similarity (S-SIM) degradation (details in Appendix B.1).

Observation 1: Pre-trained speech velocity fields contain a directly usable source-to-target editing signal, enabling emotion modification without model training.

Q2: When along the generative trajectory should editing be applied?

Building on Q1, we examine whether the editing signal's effectiveness varies along the generative trajectory. Full-step transport successfully transfers emotion but introduces temporal distortions, notably ***onset drift*** (mean 308 ms for F5-TTS, 245 ms for CosyVoice 2). As Fig. 2 shows, successful emotion transfer and zero Word Error Rate (WER) do not guarantee temporal preservation.

To locate where this distortion arises, we delay transport by k integration steps and measure the retained emotion gain, onset drift, and ΔWER. Fig. 2 reveals highly model-dependent, non-uniform

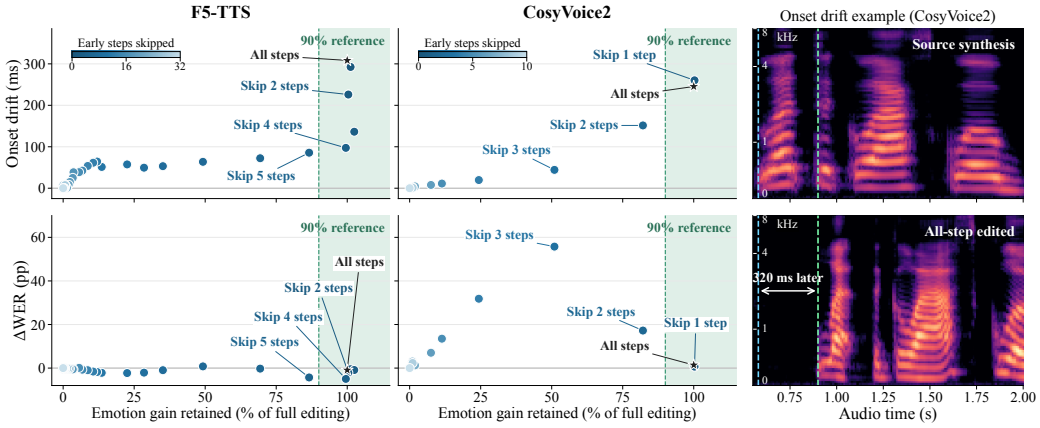


Figure 2: The first two columns show onset drift (top) and WER change from the source in percentage points (bottom) versus retained emotion gain relative to full-step editing. The green region denotes $\geq$90% gain retention. The third column demonstrates the onset drift problem.

trajectory behavior. For F5-TTS, skipping $k = 4$ steps retains 99.5% of the emotion gain while slashing onset drift to 97 ms and improving transcription.

Conversely, for CosyVoice 2, skipping $k = 1$ step preserves emotion but fails to reduce drift, whereas skipping $k = 3$ steps reduces drift at the expense of emotion strength and WER. Further early-stop and block-wise ablation analyses also confirm this temporal non-uniformity, which are detailed in Appendix B.2.

Observation 2: Editability is trajectory- and architecture-dependent: early transport may disrupt timing, while later steps are crucial for effective editing and content recovery.

Q3: What acoustic attributes are carried by source-to-target velocity transport?

Building on Q1, we examine if the editing signal isolates emotion when source and target speakers differ. We edit 80 Neutral utterances (from 20 ESD speakers, fixed text) toward different same-language speakers across four target emotions (Happy, Angry, Sad, Surprise). We track changes in speaker similarity (S-SIM), target-emotion probability gain (pp), and F0 median/range shifts (semitones).

Table 1: Cross-speaker emotion editing induces concurrent attribute changes.

Measure	F5-TTS	CosyVoice 2
S-SIM source ↓	−0.403	−0.378
S-SIM target ↑	0.331	0.342
Emotion gain (pp) ↑	45.14	52.11
F0 median gain (semitones) ↑	4.565	7.075
F0 range gain (semitones) ↑	1.433	2.769
Joint (/80)*	70	73

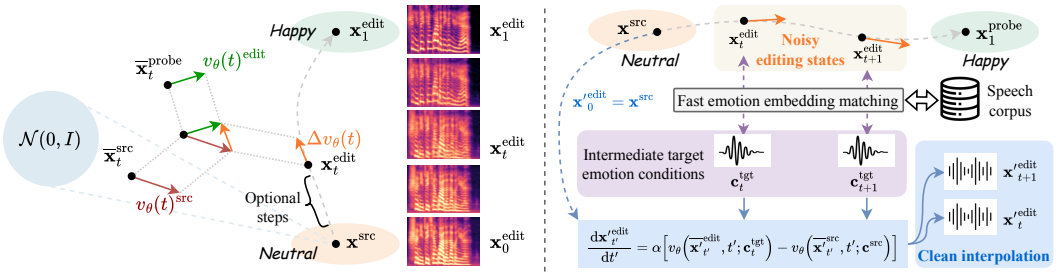
* Joint denotes the number of cases where target-speaker similarity increases, source-speaker similarity decreases, and target-emotion probability increases.

As Table 1 shows, velocity transport is not attribute-factorized. Speaker identity and emotion shift jointly (mean emotion gains of 45.14 pp for F5-TTS and 52.11 pp for CosyVoice 2), while F0 median and range also move toward the target reference. This confirms that cross-speaker transport inevitably entangles speaker and pitch characteristics with emotion. Detailed experimental settings and results are provided in Appendix B.3.

Observation 3: Velocity transport is not emotion-specific, but carries speaker identity and other attributes, such as pitch characteristics, alongside emotion.

3 SEMOEDIT

The observations in Sec. 2 directly inform the design of SEmoEdit. Based on these findings, SEmoEdit performs state-dependent velocity transport between speaker-aligned source and target emotion conditions while allowing the transport interval to adapt to the TTS backbone.



(a) Training-free emotion editing by dynamic velocity transport. (b) Training-free emotion interpolation by emotion bridging

Figure 3: Framework of SEmoEdit for dynamic speech emotion editing.

3.1 Dynamic Velocity Transport

Coupled emotion queries. Given a source acoustic representation $\mathbf{x}^{\mathrm{src}}$ and source emotion condition $\mathbf{c}^{\mathrm{src}}$, we construct an editing trajectory $\{\mathbf{x}_t^{\mathrm{edit}}\}_{t\in[0,1]}$ directly in the acoustic space. Here, t follows the CFM convention in Eq. 1, from the noise endpoint to the data endpoint, while $\mathbf{x}_0^{\mathrm{edit}} = \mathbf{x}^{\mathrm{src}}$ serves as the boundary condition of the editing dynamics rather than a sample from $p_0 = \mathcal{N}(\mathbf{0}, \mathbf{I})$. At each time t, we sample $\boldsymbol{\epsilon}_t \sim p_0$ with the same shape as $\mathbf{x}^{\mathrm{src}}$ and construct a noisy source query

$$\overline{\mathbf{x}}_t^{\mathrm{src}} = (1-t)\boldsymbol{\epsilon}_t + t\mathbf{x}^{\mathrm{src}}. \tag{6}$$

We then add the same noise perturbation used for the source query to the current editing state, yielding the coupled target-side query

$$\overline{\mathbf{x}}_t^{\mathrm{edit}} = \mathbf{x}_t^{\mathrm{edit}} + (\overline{\mathbf{x}}_t^{\mathrm{src}} - \mathbf{x}^{\mathrm{src}}). \tag{7}$$

Thus, the source and target-side queries share the same noise perturbation while being anchored at $\mathbf{x}^{\mathrm{src}}$ and $\mathbf{x}_t^{\mathrm{edit}}$, respectively. At initialization, they coincide: $\overline{\mathbf{x}}_0^{\mathrm{edit}} = \overline{\mathbf{x}}_0^{\mathrm{src}} = \boldsymbol{\epsilon}_0$.

Following Observation 1, we define the instantaneous editing signal as the velocity difference

$$\Delta\boldsymbol{v}_\theta(t) = \boldsymbol{v}_\theta\big(\overline{\mathbf{x}}_t^{\mathrm{edit}}, t; \mathbf{c}^{\mathrm{tgt}}\big) - \boldsymbol{v}_\theta(\overline{\mathbf{x}}_t^{\mathrm{src}}, t; \mathbf{c}^{\mathrm{src}}). \tag{8}$$

Unlike a fixed activation-steering direction (Xie et al., 2025b; Wang et al., 2026b), $\Delta\boldsymbol{v}_\theta(t)$ is recomputed from the evolving editing state at every step, making the transport more robust and stable.

Speaker-aligned emotion conditions. Observation 3 shows that cross-speaker velocity transport also carries speaker identity. We therefore align the target-reference timbre with the source speaker before computing Eq. 8: $\mathbf{c}^{\mathrm{tgt}} = \mathcal{E}_{\mathrm{vc}}\left(\mathbf{c}^{\mathrm{tgt}'}; \mathbf{c}^{\mathrm{src}}\right)$, where $\mathcal{E}_{\mathrm{vc}}$ converts the target reference to the source-speaker timbre while preserving its emotion, making $\mathbf{c}^{\mathrm{src}}$ and $\mathbf{c}^{\mathrm{tgt}}$ mainly differ in emotion, reducing speaker-dependent components in the velocity difference.

Trajectory-aware transport. Observation 2 shows that early transport can disrupt temporal structure. We therefore introduce an optional transport start time $\tau \in [0,1]$ and define $g_\tau(t) = \mathbb{I}[t \geq \tau]$, where $\tau = 0$ recovers full-step transport and $\tau > 0$ skips the early part of the trajectory. SEmoEdit then evolves the source utterance according to

$$\frac{\mathrm{d}\mathbf{x}_t^{\mathrm{edit}}}{\mathrm{d}t} = \alpha\, g_\tau(t)\, \mathbb{E}_{\boldsymbol{\epsilon}_t}\left[\Delta\boldsymbol{v}_\theta(t) \mid \mathbf{x}^{\mathrm{src}}\right], \qquad \mathbf{x}_0^{\mathrm{edit}} = \mathbf{x}^{\mathrm{src}}, \qquad \mathbf{x}^{(\alpha)} = \mathbf{x}_1^{\mathrm{edit}}. \tag{9}$$

The strength $\alpha \in \mathbb{R}$ controls the signed transport magnitude: $\alpha = 0$ recovers the source utterance, $0 < \alpha < 1$ interpolates toward the target emotion, $\alpha = 1$ performs the full source-to-target transport.

In practice, the expectation in Eq. 9 is approximated using n_{avg} coupled noise samples. For editing steps $0 = t_0 < \cdots < t_N = 1$, the update becomes

$$\mathbf{x}_{t_{i+1}}^{\mathrm{edit}} = \mathbf{x}_{t_i}^{\mathrm{edit}} + \alpha\, g_\tau(t_i)(t_{i+1} - t_i)\frac{1}{n_{\mathrm{avg}}}\sum_{j=1}^{n_{\mathrm{avg}}} \Delta\boldsymbol{v}_\theta^{(j)}(t_i). \tag{10}$$

Thus, SEmoEdit requires only forward evaluations of the frozen TTS velocity field, with no parameter updates or task-specific optimization. Fig. 3(a) illustrates the evolving editing states. Appendix A provides the complete pseudocode.

3.2 Enabling Stable Interpolation.

Although the trajectory in Eq. 10 provides intermediate editing states, directly decoding these states does not necessarily produce clean speech. An intermediate state simultaneously contains source and target acoustic patterns, which may interfere with each other and manifest as audible noise or other artifacts. In other words, the editing state can represent a meaningful intermediate emotion without itself being a well-formed speech sample.

We address this issue with **emotion bridging**. As shown in Fig. 3(b), for an intermediate editing state $\mathbf{x}_{t_i}^{\mathrm{edit}}$, we first obtain a clean target reference $\mathbf{c}_t^{\mathrm{tgt}}$ whose emotional representation (emotion2vec embedding) matches that state and align its timbre with the source speaker. $\mathbf{c}_t^{\mathrm{tgt}}$ serves as an emotion bridge: rather than using the noisy intermediate state as the output, we apply SEmoEdit again from the original source $\mathbf{x}^{\mathrm{src}}$ toward the bridged condition, thereby synthesizing a clean utterance $\mathbf{x}'^{\mathrm{edit}}_t$ at the corresponding emotion intensity. The coefficient α controls the emotion strength.

4 Experiments

4.1 SEmoEditBench

Existing speech editing benchmarks (Zhang et al., 2026; Yan et al., 2025a; Ma et al., 2026b) mainly pair each source utterance with a single instruction, whereas our method requires source–target emotion condition pairs whose velocity difference defines the editing direction. We therefore construct **SEmoEditBench**, a 600-case paired benchmark with emotion labels and editing instructions for evaluating emotion transfer and non-target attribute preservation.

Tasks. SEmoEditBench covers *emotion replacement*, *emotion erasure*, and *intensity control*, each under same-dataset same-speaker, same-dataset cross-speaker, and cross-dataset cross-speaker settings. It contains 320 replacement, 152 erasure, and 128 intensity cases from ESD, IEMOCAP, RAVDESS, and CREMA-D (Zhou et al., 2022; Busso et al., 2008; Livingstone & Russo, 2018; Cao et al., 2014). Details are in Appendix C.2.

Metrics. We evaluate our method across four dimensions (details in Appendix C.2): **1) Effectiveness:** Target emotion probability (TEP) and source emotion suppression (SES) for replacement, neutral probability (NP) and SES for erasure, plus emotion2vec similarity (E-SIM) and directional editing score (DES) when paired targets exist. **2) Intensity Control:** On a 90-case subset, we evaluate EIC-Emb, our newly proposed metric measuring the monotonic movement of emotion2vec embeddings toward targets across five strengths $\alpha \in \{0, 0.25, 0.5, 0.75, 1\}$. **3) Preservation & Quality:** Assessed via ΔWER (ΔCER for Chinese), S-SIM, and UTMOS (Radford et al., 2023; Baba et al., 2024). **4) Subjective Evaluation:** Speaker (SS-MOS), emotion (ES-MOS), and naturalness (N-MOS) evaluated on 20 sampled cases. All metrics are averaged across cases.

Models and Hyperparameters. We compare SEmoEdit (applied to three backbones, i.e., F5-TTS, CosyVoice 2, and IndexTTS 2 (Zhou et al., 2026b)) with representative training-based (Yan et al., 2025b; Wang et al., 2026a; Ma et al., 2026a) and activation-steering methods (Xie et al., 2025b; Wang et al., 2026b). For emotion replacement and erasure, we set $\alpha = 1$, while for intensity control we use $\alpha \in \{0, 0.25, 0.5, 0.75, 1\}$. We set $n_{\mathrm{avg}} = 1$ and $\tau = 0$ for all the main experiments, applying transport over the full generative trajectory. IndexTTS 2 is used to convert target references to the source speaker's timbre for SEmoEdit. A speech corpus with 10K samples is created for emotion bridging, and their emotion embeddings are pre-computed for fast matching.

4.2 Main Results

Comparison with training-based methods. As shown in Table 2, audio-conditioned SEmoEdit consistently outperforms all training-based baselines in editing success without task-specific training. Even its weakest backbone surpasses the strongest baseline on every emotion-related metric for replacement (e.g., TEP 0.498 vs. 0.434; DES 0.753 vs. 0.713) and erasure (NP 0.573 vs. 0.355; DES 0.873 vs. 0.766), while the best variants reach 0.691/0.920 TEP/DES for replacement and 0.704/0.938 NP/DES for erasure. These gains retain near-zero or negative ΔWER and additionally support continuous intensity control. Overall, pretrained generative dynamics provide stronger emotion-editing capability than dedicated editors trained on large-scale task-specific data.

Table 2: Main results on the same-dataset same-speaker setting of SEmoEditBench. The top three results for each metric are highlighted in **bold** with dark, medium, and light pink backgrounds representing the first , second , and third best performances, respectively.

Category	Method	Backbone	Metrics						
Emotion replacement			TEP ↑	SES ↑	E-SIM ↑	DES ↑	ΔWER ↓	S-SIM ↑	UTMOS ↑
Training-based	Step-Audio-EditX	–	0.222	0.445	0.521	0.550	**-0.048**	0.567	**3.091**
	dots.tts.edit	–	0.434	0.600	0.669	0.713	-0.035	0.272	2.381
	Auk	–	0.259	0.372	0.539	0.541	0.073	0.631	2.639
Activation Steering	CoCoEmo	CosyVoice 2	0.082	0.192	0.446	0.399	-0.015	**0.715**	**3.061**
		IndexTTS2	0.035	0.127	0.402	0.269	-0.024	**0.776**	2.659
	EmoSteer-TTS	F5-TTS	0.081	0.345	0.448	0.477	0.608	0.623	2.545
		CosyVoice 2	0.030	0.371	0.382	0.367	-0.015	0.478	2.515
		IndexTTS2	0.025	0.054	0.377	0.226	**-0.037**	**0.788**	2.666
SEmoEdit	**Audio condition**	F5-TTS	**0.498**	**0.684**	**0.704**	**0.753**	-0.026	0.372	2.177
		CosyVoice 2	**0.554**	**0.694**	**0.776**	**0.806**	-0.025	0.379	2.984
		IndexTTS2	**0.691**	**0.767**	**0.902**	**0.920**	-0.004	0.423	2.578
	Text condition†	CosyVoice 2	0.274	0.484	0.566	0.567	**-0.059**	0.610	**3.050**
		IndexTTS2	0.354	0.506	0.629	0.634	0.013	0.594	2.572
Emotion erasure			NP ↑	SES ↑	E-SIM ↑	DES ↑	ΔWER ↓	S-SIM ↑	UTMOS ↑
Training-based	Step-Audio-EditX	–	0.235	0.388	0.566	0.655	**-0.064**	0.577	**3.220**
	dots.tts.edit	–	0.355	0.549	0.674	0.766	**-0.065**	0.304	2.685
	Auk	–	0.229	0.311	0.528	0.582	0.014	**0.676**	**2.859**
Activation Steering	EmoSteer-TTS	F5-TTS	0.202	0.393	0.616	0.736	0.030	0.598	2.593
		CosyVoice 2	0.078	0.174	0.505	0.596	-0.012	0.548	2.490
		IndexTTS2	0.038	0.070	0.377	0.298	-0.041	**0.808**	2.836
SEmoEdit	**Audio condition**	F5-TTS	**0.573**	**0.700**	**0.821**	**0.873**	**-0.058**	0.359	2.597
		CosyVoice 2	**0.704**	**0.754**	**0.907**	**0.938**	-0.050	0.399	**3.144**
		IndexTTS2	**0.682**	**0.741**	**0.910**	**0.934**	-0.030	0.424	2.774
	Text condition†	CosyVoice 2	0.216	0.432	0.582	0.685	-0.052	**0.607**	**3.069**
		IndexTTS2	0.104	0.309	0.471	0.504	-0.050	**0.687**	2.698
Emotion intensity control			EIC-Emb ↑		$\Delta\text{WER}_{\alpha=1}$ ↓		$\text{S-SIM}_{\alpha=1}$ ↑	$\text{UTMOS}_{\alpha=1}$ ↑	
Activation Steering	CoCoEmo	CosyVoice 2	0.041		0.044		**0.621**	**3.234**	
		IndexTTS2	0.017		0.333		**0.703**	2.294	
	EmoSteer-TTS	F5-TTS	0.013		0.139		0.465	2.309	
		CosyVoice 2	-0.009		0.133		0.362	**2.575**	
		IndexTTS2	-0.005		0.400		**0.683**	2.355	
SEmoEdit	**Audio condition**	F5-TTS	**0.175**		**0.039**		0.307	1.967	
		CosyVoice 2	**0.172**		**0.017**		0.362	2.499	
		IndexTTS2	**0.192**		0.411		0.336	2.355	
	Text condition†	CosyVoice 2	0.144		**-0.017**		0.553	**3.080**	
		IndexTTS2	0.108		0.394		0.472	2.458	

† The text condition is an emotion instruction, e.g., *Speak with a happy tone.*.

Comparison with activation steering methods. Audio-conditioned SEmoEdit consistently outperforms fixed activation steering across all three backbones. For replacement and erasure, every SEmoEdit variant surpasses all steering baselines on all four emotion-related metrics. The advantage is also clear for intensity control, where EIC-Emb reaches 0.175, 0.172, and 0.192 on F5-TTS, CosyVoice 2, and IndexTTS2, compared with 0.013, 0.041, and 0.017 for their strongest steering counterparts. While SEmoEdit yields lower S-SIM, these consistent gains demonstrate that state-dependent velocity transport provides a more effective editing signal than fixed activation directions.

Human Evaluation. Table 3 (15 human evaluators) confirms SEmoEdit's stronger perceived editing capability. SEmoEdit achieves the best emotion-replacement score (ES-MOS 4.22 with IndexTTS2), the top three erasure scores, and higher intensity-control scores (EIC-MOS 2.80–3.65) than CoCoEmo (1.65) and EmoSteer-TTS (1.30). Although the baselines sometimes obtain higher SS-MOS or N-MOS, their low editing scores suggest little emotion changes. SEmoEdit therefore provides a better trade-off between edit effectiveness and output quality.

Generalization to out-of-distribution cases. The left panel of Fig. 4 shows that SEmoEdit remains effective in the challenging cross-dataset, cross-speaker setting without retraining, outperforming most ID baselines in Table 2. For intensity control, IndexTTS2 retains 95% of its ID-average score (0.190 vs. 0.199), whereas F5-TTS and CosyVoice 2 decline to some extent. Replacement and erasure also show TEP/NP drops. Overall, dynamic velocity transport generalizes across corpora and speakers without retraining, but continuous-control robustness depends on the backbone and cross-domain categorical emotion transfer remains a key limitation. Full results are in Appendix D.1.

Table 3: Subjective evaluation results. The first , second , and third place are highlighted.

Category	Method	Backbone	Emotion replacement			Emotion erasure			Intensity control
			SS-MOS ↑	ES-MOS ↑	N-MOS ↑	SS-MOS ↑	ES-MOS ↑	N-MOS ↑	EIC-MOS ↑
Training-based	Step-Audio-EditX	–	**3.67**	2.88	3.75	**4.20**	2.35	4.00	–
	Auk	–	3.38	2.80	3.52	**4.00**	2.95	**4.05**	–
	dots.tts.edit	–	3.52	**3.27**	3.77	3.55	3.15	3.65	–
Activation Steering	CoCoEmo	IndexTTS2	**4.42**	1.88	**3.98**	–	–	–	1.65
	EmoSteer-TTS	IndexTTS2	**4.35**	1.75	**4.12**	**4.50**	1.80	**4.25**	1.30
SEmoEdit	Ours	F5-TTS	2.88	**3.23**	3.25	2.85	**3.65**	3.80	**2.80**
		CosyVoice 2	3.02	3.20	3.42	3.45	**4.30**	**4.15**	**3.25**
		IndexTTS2	3.45	**4.22**	**3.88**	3.55	**4.40**	**4.15**	**3.65**

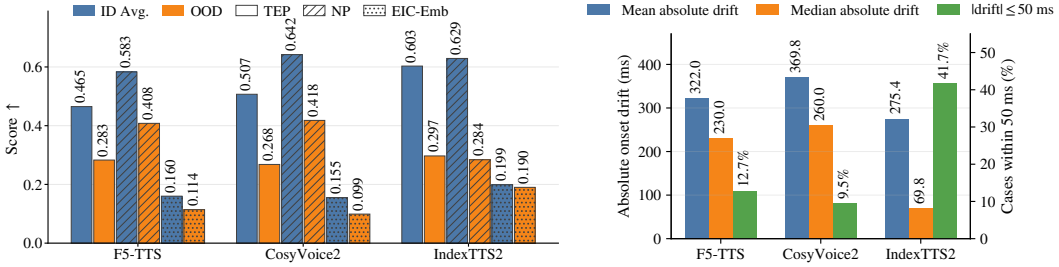


Figure 4: **Left:** SEmoEdit ID/OOD generalization across backbones. **Right:** Onset drift over 600 cases; lower mean and median absolute drift are better, while a higher fraction within 50 ms is better.

4.3 Ablation Study

Analysis of onset drifting. We further evaluate onset drift across all three backbones using 600 SEmoEditBench cases. As Fig. 4 shows, IndexTTS2 best preserves timing with a 69.8-ms median drift (41.7% ≤50 ms). Although long tail cases inflates IndexTTS2's mean, CosyVoice 2 yields the highest mean and median, confirming its stronger emotion-temporal coupling. This contrast reflects architecture designs: F5-TTS establishes timing jointly along fixed text length, with no explicit duration control; CosyVoice 2 relies on unaligned autoregressive semantic tokens, leaving its flow prior to impose a target-specific temporal scaffold; IndexTTS2 instead enforces a duration-controlled frame layout before flow matching, substantially reducing onset drift.

Analysis of noise. Fig. 5 shows that increasing n_{avg} from 1 to 8 has little effect on CosyVoice 2 and IndexTTS2. F5-TTS is more sensitive, but gains remain inconsistent (at $n_{\text{avg}} = 8$: TEP $+0.033$, ΔWER $+0.036$, UTMOS -0.234). Since additional samples provide no consistent benefit while increasing velocity-query cost up to $8\times$, we use $n_{\text{avg}} = 1$ throughout all experiments.

Table 4: Objective metrics before and after emotion bridging, averaged over 90 intensity editing cases.

Backbone	Variant	WER/CER ↓	S-SIM ↑	UTMOS ↑
F5-TTS	Raw	0.342	0.418	2.151
	Bridged	**0.205**	**0.431**	**2.411**
CosyVoice 2	Raw	0.360	0.349	2.134
	Bridged	**0.201**	**0.457**	**2.780**
IndexTTS2	Raw	0.823	0.379	1.918
	Bridged	**0.603**	**0.440**	**2.652**

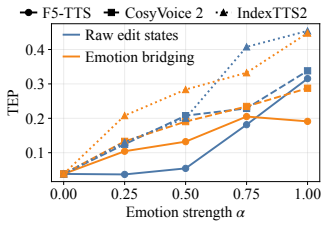


Figure 6: Mean TEP trajectories before and after bridging.

Analysis of the emotion bridging. Table 4 shows emotion bridging consistently improves acoustic stability across all backbones. It reduces WER/CER, increases UTMOS, and improves S-SIM, notably for CosyVoice 2. This suggests the second pass removes intermediate acoustic artifacts, yielding clearer, speaker-consistent speech. As Fig. 6 shows, bridging maintains highly stable intensity separation across most levels, exhibiting only minor fluctuations at high strengths. Thus, it effectively acts as a fidelity regularizer while largely preserving the interpolation range, intelligibility, timbre, and naturalness.

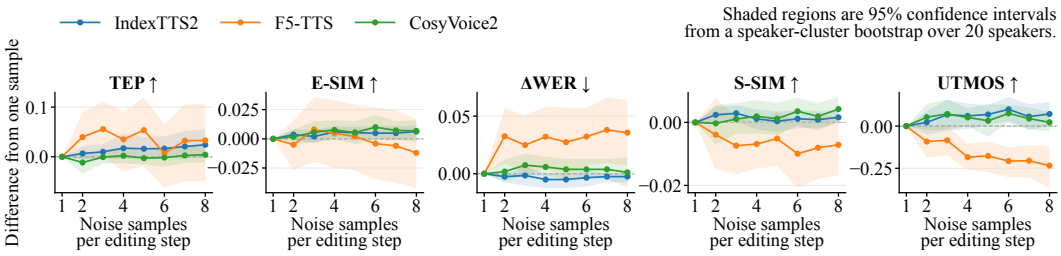


Figure 5: Sensitivity to the number of coupled noise samples per editing step.

Audio condition vs. text condition. As shown in Table 2, replacing audio conditions with text instructions sharply reduces replacement TEP from 0.554/0.691 to 0.274/0.354 on CosyVoice 2/IndexTTS2, with similar degradation in erasure and intensity control. This indicates that current instruction-conditioned TTS models cannot reliably induce the target-emotion velocity difference required by SEmoEdit, despite better speaker preservation.

4.4 Takeaways for Real Applications and Future Directions

Based on our experiments and findings, we want to highlight several key insights and potential future directions: (1) Flow-based editing techniques can be applied to acoustic flows, but the temporal nature of audio requires careful consideration. (2) The choice of backbone TTS model significantly affects the editing performance, with models that have explicit duration control (e.g., IndexTTS2) showing better stability. (3) Noise averaging does not provide consistent benefits. (4) Emotion bridging indicates that multi-pass editing strategies may be beneficial for complex tasks. (5) Text-conditioned training-free editing currently lags behind, highlighting the need for improved instruction-conditioned TTS models that can better capture the desired emotional transformations.

5 Related Work

Emotion-Conditioned Speech Synthesis generates new speech from prompts or labels. These methods offer prompt-driven control (PromptTTS (Guo et al., 2023), ControlSpeech (Ji et al., 2025), EmoVoice (Yang et al., 2025)), continuous intensity tuning (EmoSphere++ (Cho et al., 2025)), word-level alignment (WeSCon (Wang et al., 2026c)), or timbre disentanglement (IndexTTS2 (Zhou et al., 2026b)). However, they cannot edit existing audio. **Training-Based Emotion Editing** modifies source speech via mask-based inpainting (Emo-CampNet (Wang et al., 2024)), instruction-driven LMs (Step-Audio-EditX (Yan et al., 2025b), SpeechEdit (Pei et al., 2026)), caption rewriting (Bagpiper-Edit (Gong et al., 2026)), transcript-grounded span editing (dots.tts.edit (Wang et al., 2026a)), or voice-level tuning (VoiceDesigner (Hai et al., 2026)). Despite their flexibility, these rely on expensive annotations, dedicated training, or post-training. **Inference-Time Representation Steering** manipulates representations without retraining via activation steering (EmoSteer-TTS (Xie et al., 2025b), CoCoEmo (Wang et al., 2026b)), lightweight interventions (EmoShift (Zhou et al., 2026a)), or modifying SAE features (Du et al., 2026). Unlike these methods that primarily regenerate speech or steer via fixed offsets, SEmoEdit dynamically modulates velocity transport to edit existing audio, unifying emotion replacement, erasure, and interpolation.

6 Conclusion

This work presents the first systematic study of training-free emotion editability in pre-trained flow-matching and hybrid TTS models. Driven by our in-depth analysis and three key observations that reveal the potential and limitations of repurposing speech flows, we introduce the SEmoEdit framework. Extensive comparison and evaluations on SEmoEditBench demonstrate that SEmoEdit serves as a highly practical alternative to training-based approaches, achieving competitive, and often superior, performance without requiring parameter updates. We hope these insights inspire broader exploration of inference-time speech editing within foundational audio models.

AI USAGE STATEMENT

In this work, we used generative AI tools for text polishing, surveying related work, assisting with partial code writing for experiments, and generating data visualizations. We have not used generative AI tools for formulating the core scientific ideas, deriving foundational theoretical concepts, or drafting the primary contributions from scratch, and the generation of fully automated research manuscripts is not applicable to this work. Additionally, we used generative AI tools for structural formatting and grammatical refinement of the text. We have reviewed all AI-assisted work. Specifically, all AI-assisted literature summaries were manually cross-checked against the original published papers to ensure factual accuracy and avoid misattribution; all LLM-generated code and experimental scripts were rigorously tested, debugged, and verified by the authors to guarantee functional correctness; and all generated plots were strictly validated against our raw empirical data. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work presents advancements in zero-shot speech emotion editing and synthesis and we acknowledge that the capabilities developed in this research may raise potential concerns regarding security, privacy, and misinformation.

First, our method is capable of generating high-fidelity speech that mimics the acoustic identity of unseen individuals. While this capability drives positive applications such as personalized accessibility tools, audiobook generation, and creative post-production, it also poses security risks. The technology could potentially be misused for malicious impersonation, bypassing voice biometric authentication, or generating deceptive audio (deepfakes) without the target speaker's consent.

Second, the ability to directly modify the emotion of an *existing* speech utterance introduces unique ethical challenges. Altering the affective state of a real recording can fundamentally misrepresent the speaker's original intent, attitude, or psychological state. In malicious hands, emotion editing could be exploited to manipulate public perception, spread misinformation, or defame individuals by falsely portraying them as inappropriately angry, distressed, or joyful in sensitive contexts.

To mitigate these risks, we emphasize that our research is intended solely for benign applications in human-computer interaction and the creative arts. We strongly advocate for the concurrent development of robust audio provenance, watermarking, and deepfake detection technologies. Furthermore, we emphasize that any real-world deployment of voice cloning and emotion editing technologies must be governed by strict ethical guidelines, ensuring that an individual's voice is neither synthesized nor emotionally altered without their explicit and informed consent.

REPRODUCIBILITY STATEMENT

To support reproducibility, Sec. 3.1 defines the SEmoEdit update, and Appendix A provides pseudocode for its discrete implementation. Section 4.1 identifies the evaluated pre-trained backbones and the editing hyperparameters; Appendix B specifies the backbone-specific conditioning, solver schedules, guidance scales, and random seeds used in the editability experiments. Appendix C.1 describes the construction, splits, and reference-selection protocol of SEmoEditBench, while Appendix C.2 defines the evaluation models, preprocessing, and metrics. Complete results for the in-distribution and cross-dataset settings are reported in Appendix D.1. SEmoEdit uses frozen pre-trained components and requires no parameter updates or task-specific training.

The supplementary material includes our implementation code, SEmoEditBench dataset, and scripts for running the experiments.

REFERENCES

Kaito Baba, Wataru Nakata, Yuki Saito, and Hiroshi Saruwatari. The t05 system for the VoiceMOS Challenge 2024: Transfer learning from deep image classifier to naturalness MOS prediction of

high-quality synthetic speech. In *IEEE Spoken Language Technology Workshop (SLT)*, pp. 818–824, 2024. doi: 10.1109/SLT61566.2024.10832315.

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

Houwei Cao, David G Cooper, Michael K Keutmann, Ruben C Gur, Ani Nenkova, and Ragini Verma. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing*, 5(4):377–390, 2014.

Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, JianZhao JianZhao, Kai Yu, and Xie Chen. 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)*, pp. 6255–6271, 2025.

Deok-Hyeon Cho, Hyung-Seok Oh, Seung-Bin Kim, and Seong-Whan Lee. Emosphere++: Emotion-controllable zero-shot text-to-speech via emotion-adaptive spherical vector. *IEEE Transactions on Affective Computing*, 16(3):2365–2380, 2025.

Hongfei Du, Jiacheng Shi, Sidi Lu, Gang Zhou, and Ashley Gao. Sparse autoencoders for interpretable emotion control in text-to-speech. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=Hbt2ryJgz2.

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

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

Xun Gong, Jinchuan Tian, Haoran Wang, William Chen, Shinji Watanabe, and Yanmin Qian. Bagpiper-edit: Zero-shot open-ended audio editing via rich-caption. *arXiv preprint arXiv:2606.21227*, 2026.

Yiwei Guo, Chenpeng Du, Ziyang Ma, Xie Chen, and Kai Yu. Voiceflow: Efficient text-to-speech with rectified flow matching. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 11121–11125. IEEE, 2024.

Zhifang Guo, Yichong Leng, Yihan Wu, Sheng Zhao, and Xu Tan. Prompttts: Controllable text-to-speech with text descriptions. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Jiarui Hai, Karan Thakkar, Ke Chen, Yunyun Wang, Jiaqi Su, Rithesh Kumar, Mounya Elhilali, and Zeyu Jin. Voicedesigner: Text-to-voice generation and editing via unified diffusion modeling and data augmentation. *arXiv preprint arXiv:2608.13613*, 2026.

Hangrui Hu, Xinfa Zhu, Ting He, Dake Guo, Bin Zhang, Xiong Wang, Zhifang Guo, Ziyue Jiang, Hongkun Hao, Zishan Guo, et al. Qwen3-tts technical report. *arXiv preprint arXiv:2601.15621*, 2026.

Shengpeng Ji, Qian Chen, Wen Wang, Jialong Zuo, Minghui Fang, Ziyue Jiang, Hai Huang, Zehan Wang, Xize Cheng, Siqi Zheng, et al. Controlspeech: Towards simultaneous and independent zero-shot speaker cloning and zero-shot language style control. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6966–6981, 2025.

Vladimir Kulikov, Matan Kleiner, Inbar Huberman-Spiegelglas, and Tomer Michaeli. Flowedit: Inversion-free text-based editing using pre-trained flow models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19721–19730. IEEE, 2025.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Steven R Livingstone and Frank A Russo. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5):e0196391, 2018.

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

Ziyang Ma, Zhikang Niu, Wenming Tu, Tianrui Wang, Ruiqi Yan, Junxi Liu, Yanru Huo, Nickk Huang, Yang Liu, Qicong Xie, et al. Auk technical report: An open-source foundational model for speech generation and editing. *arXiv preprint arXiv:2609.08936*, 2026a.

Ziyang Ma, Ruiqi Yan, Ruiyang Xu, Jie Fang, Zhikang Niu, Yi-Wen Chao, Wenming Tu, Tianrui Wang, Qi Chen, Wenxi Chen, et al. Mmae: A massive multitask audio editing benchmark. *arXiv preprint arXiv:2606.07229*, 2026b.

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

Shivam Mehta, Ruibo Tu, Jonas Beskow, Éva Székely, and Gustav Eje Henter. Matcha-tts: A fast tts architecture with conditional flow matching. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 11341–11345. IEEE, 2024.

Hanchen Pei, Shujie Liu, Yanqing Liu, Jianwei Yu, Yuanhang Qian, Gongping Huang, Sheng Zhao, and Yan Lu. A unified neural codec language model for selective editable text to speech generation. *arXiv preprint arXiv:2601.12480*, 2026.

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

Hankun Wang, Bohan Li, Shi Lian, Xiaoyu Gu, Jing Peng, Da Zheng, Colin Zhang, and Kai Yu. dots. tts. edit: Precisely controlled speech editing with a continuous autoregressive model. *arXiv preprint arXiv:2608.02673*, 2026a.

Siyi Wang, Shihong Tan, Siyi Liu, Hong Jia, Gongping Huang, James Bailey, and Ting Dang. Cocoemo: Composable and controllable human-like emotional TTS via activation steering. In *Forty-third International Conference on Machine Learning*, 2026b. URL https://openreview.net/forum?id=wW2LIHSCfw.

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

Tianrui Wang, Haoyu Wang, Meng Ge, Cheng Gong, Chunyu Qiang, Ziyang Ma, Zikang Huang, Guanrou Yang, Xiaobao Wang, Eng-Siong Chng, et al. Word-level emotional expression control in zero-shot text-to-speech synthesis. *Advances in Neural Information Processing Systems*, 38: 147377–147405, 2026c.

Tianxin Xie, Yan Rong, Pengfei Zhang, Wenwu Wang, and Li Liu. 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*, pp. 764–791, 2025a.

Tianxin Xie, Shan Yang, Chenxing Li, Dong Yu, and Li Liu. Emosteer-tts: Fine-grained and training-free emotion-controllable text-to-speech via activation steering. *arXiv preprint arXiv:2508.03543*, 2025b.

Sihan Xu, Yidong Huang, Jiayi Pan, Ziqiao Ma, and Joyce Chai. Inversion-free image editing with language-guided diffusion models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9454–9461. IEEE, 2024.

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

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

Guanrou Yang, Chen Yang, Qian Chen, Ziyang Ma, Wenxi Chen, Wen Wang, Tianrui Wang, Yifan Yang, Zhikang Niu, Wenrui Liu, et al. Emovoice: Llm-based emotional text-to-speech model with freestyle text prompting. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 10748–10757, 2025.

Hanlin Zhang, Daxin Tan, Dehua Tao, Xiao Chen, Haochen Tan, and Linqi Song. Speechedit-itbench: A bilingual multi-attribute benchmark for instruction-guided speech editing. *arXiv preprint arXiv:2606.01804*, 2026.

Kun Zhou, Berrak Sisman, Rui Liu, and Haizhou Li. Emotional voice conversion: Theory, databases and esd. *Speech communication*, 137:1–18, 2022.

Li Zhou, Hao Jiang, Junjie Li, Tianrui Wang, and Haizhou Li. Emoshift: Lightweight activation steering for enhanced emotion-aware speech synthesis. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 17262–17266. IEEE, 2026a.

Siyi Zhou, Yiquan Zhou, Yi He, Xun Zhou, Jinchao Wang, Wei Deng, and Jingchen Shu. Indextts2: A breakthrough in emotionally expressive and duration-controlled auto-regressive zero-shot text-to-speech. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 35139–35148, 2026b.

A SEmoEdit Algorithm

Algorithm 1 summarizes the discrete dynamic velocity transport in Eq. 10. The target reference is first converted to the source-speaker timbre; when the two references are already speaker-aligned, this conversion is the identity operation. For backbones that prepend reference frames, each velocity evaluation retains its branch-specific prefix, and the difference below is computed only over the shared generated region.

Algorithm 1 SEmoEdit via dynamic velocity transport

Require: Frozen velocity field $\boldsymbol{v}_\theta$; source representation $\mathbf{x}^{\mathrm{src}}$; source and target-reference conditions $\mathbf{c}^{\mathrm{src}}, \mathbf{c}^{\mathrm{tgt}'}$; speaker alignment operator $\mathcal{E}_{\mathrm{vc}}$; time grid $0 = t_0 < \cdots < t_N = 1$; strength α; start time τ; noise samples per step n_{avg}

Ensure: Edited representation $\mathbf{x}^{(\alpha)}$

$\mathbf{c}^{\mathrm{tgt}} \leftarrow \mathcal{E}_{\mathrm{vc}}(\mathbf{c}^{\mathrm{tgt}'}; \mathbf{c}^{\mathrm{src}})$
$\mathbf{x}^{\mathrm{edit}}_{t_0} \leftarrow \mathbf{x}^{\mathrm{src}}$
for $i = 0, \ldots, N-1$ **do**
  $\Delta \boldsymbol{v}_i \leftarrow \mathbf{0}$
  for $j = 1, \ldots, n_{\mathrm{avg}}$ **do**
    Sample $\boldsymbol{\epsilon}_{i,j} \sim p_0$ with the same shape as $\mathbf{x}^{\mathrm{src}}$
    $\overline{\mathbf{x}}^{\mathrm{src}}_{i,j} \leftarrow (1 - t_i)\boldsymbol{\epsilon}_{i,j} + t_i \mathbf{x}^{\mathrm{src}}$
    $\overline{\mathbf{x}}^{\mathrm{edit}}_{i,j} \leftarrow \mathbf{x}^{\mathrm{edit}}_{t_i} + \overline{\mathbf{x}}^{\mathrm{src}}_{i,j} - \mathbf{x}^{\mathrm{src}}$
    $\Delta \boldsymbol{v}_i \leftarrow \Delta \boldsymbol{v}_i + \boldsymbol{v}_\theta(\overline{\mathbf{x}}^{\mathrm{edit}}_{i,j}, t_i; \mathbf{c}^{\mathrm{tgt}}) - \boldsymbol{v}_\theta(\overline{\mathbf{x}}^{\mathrm{src}}_{i,j}, t_i; \mathbf{c}^{\mathrm{src}})$
  end for
  $\mathbf{x}^{\mathrm{edit}}_{t_{i+1}} \leftarrow \mathbf{x}^{\mathrm{edit}}_{t_i} + \alpha\, \mathbb{I}[t_i \geq \tau](t_{i+1} - t_i)\Delta \boldsymbol{v}_i / n_{\mathrm{avg}}$
end for
return $\mathbf{x}^{(\alpha)} \leftarrow \mathbf{x}^{\mathrm{edit}}_{t_N}$

B Details of the Editability Diagnostic

B.1 Experiment Details for Q1

Experimental setup. We construct 80 neutral-to-emotion editing cases from the Emotional Speech Database (ESD), covering all 20 speakers (ten Mandarin and ten English) and four target emotions: Happy, Angry, Sad, and Surprise. For each speaker, four distinct parallel text groups are assigned to the four target emotions. Each case contains a neutral reference R_s and a target-emotion reference R_t spoken by the same speaker with the same text, together with a separate synthesis transcript that remains fixed during editing. Both models are evaluated on the same cases, giving 20 cases per emotion and 40 per language. We use frozen F5-TTS v1 Base and CosyVoice 2-0.5B with their supplied pre-trained vocoders; no additional training is performed.

F5-TTS uses 32 Euler integration steps with its supplied time schedule and sway-sampling coefficient -1, while CosyVoice 2 uses 10 Euler steps with its supplied cosine schedule and 32-bit floating-point computation. Classifier-free guidance uses scales $\gamma = 2$ for F5-TTS and $\gamma = 0.7$ for CosyVoice 2. Both editors use strength $\alpha = 1$, one noise sample per step, and the full interval $[0, 1]$. Case selection uses seed 42, and generation uses seed $42 + j$ for zero-indexed case j.

Editing procedure. Following Sec. 3.1, $\mathbf{x}^{\mathrm{src}}$ is the length-aligned source-generated mel segment, $\mathbf{x}^{\mathrm{edit}}_t$ is the evolving edited segment, and $\mathbf{c}^{\mathrm{src}}, \mathbf{c}^{\mathrm{tgt}}$ are constructed from R_s, R_t. Each model input contains a reference prefix followed by the generated segment. We initialize $\mathbf{x}^{\mathrm{edit}}_{t_0} = \mathbf{x}^{\mathrm{src}}$ and use evaluation times $0 = t_0 < \cdots < t_K = 1$, with $K = 32$ for F5-TTS and $K = 10$ for CosyVoice 2. At each step, the two branches share a fresh noise sample $\boldsymbol{\epsilon}_{t_i} \sim p_0$. The Euler discretization is

$$\begin{aligned} \overline{\mathbf{x}}^{\mathrm{src}}_{t_i} &= (1 - t_i)\boldsymbol{\epsilon}_{t_i} + t_i \mathbf{x}^{\mathrm{src}}, \qquad \overline{\mathbf{x}}^{\mathrm{edit}}_{t_i} = \mathbf{x}^{\mathrm{edit}}_{t_i} + \overline{\mathbf{x}}^{\mathrm{src}}_{t_i} - \mathbf{x}^{\mathrm{src}}, \\ \mathbf{x}^{\mathrm{edit}}_{t_{i+1}} &= \mathbf{x}^{\mathrm{edit}}_{t_i} + \alpha(t_{i+1} - t_i)\left[\boldsymbol{v}_\theta\big(\overline{\mathbf{x}}^{\mathrm{edit}}_{t_i}, t_i; \mathbf{c}^{\mathrm{tgt}}\big) - \boldsymbol{v}_\theta\big(\overline{\mathbf{x}}^{\mathrm{src}}_{t_i}, t_i; \mathbf{c}^{\mathrm{src}}\big)\right]. \end{aligned} \tag{11}$$

Each velocity evaluation retains its branch-specific reference prefix, which is removed before subtracting the equal-length generated-region velocities. Only the generated segment is updated, and the final state $\mathbf{x}^{\mathrm{edit}} = \mathbf{x}^{\mathrm{edit}}_{t_K}$ is decoded with the model's vocoder.

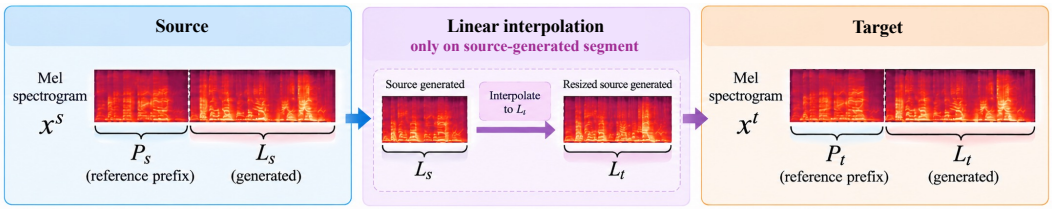


Figure 7: F5-TTS conditioning. The source-generated mel segment is linearly interpolated to the length of the target-generated mel segment.

F5-TTS conditioning. The source and target branches use the mel spectrograms of R_s and R_t, respectively, while each branch's text tokens contain its reference transcript followed by the shared synthesis transcript. We synthesize the two branches separately and linearly interpolate only the source-generated mel segment to the target-generated length, as shown in Fig. 7. Both reference mel conditions and their state prefixes retain their original lengths; the acoustic conditions are zero-padded over the generated region.

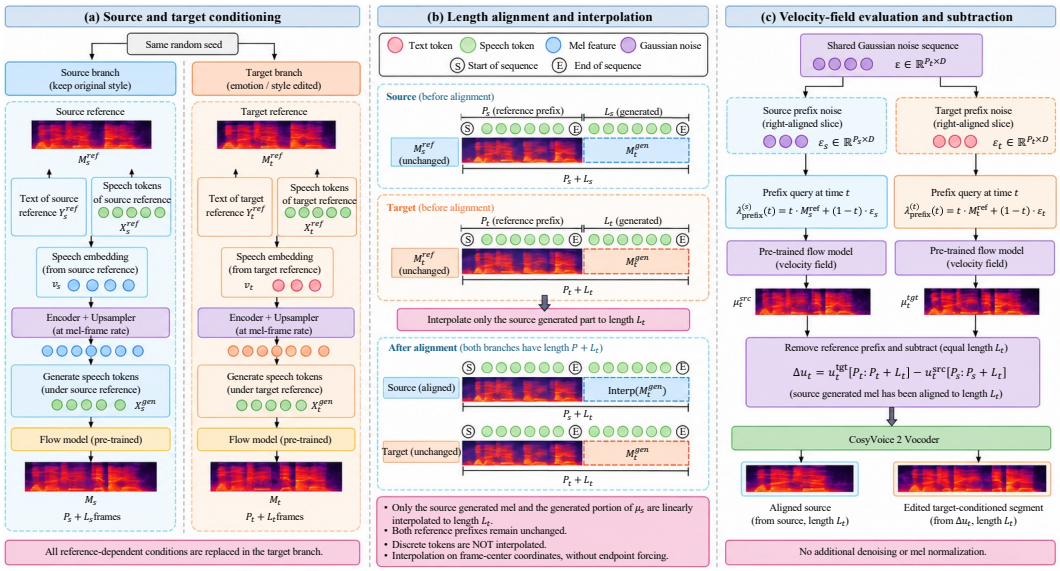


Figure 8: CosyVoice 2 conditioning. Each branch uses its own reference mel and speaker embedding, together with the generated mel segment under that reference. The source branch's generated mel is linearly interpolated to the target length, while the target branch remains unchanged. The two branches share a Gaussian noise sequence of length $\max(P_s, P_t)$, which is right-aligned to each reference prefix.

CosyVoice 2 conditioning. As shown in Fig. 8, source and target sequences are synthesized separately using their respective references and the same random seed. Each branch uses the text and discrete speech tokens of its own reference, together with the speech tokens generated under that reference; these are encoded and upsampled to a continuous representation μ at the mel-frame rate. The flow model additionally receives the corresponding speaker embedding and reference mel, so the target branch replaces all reference-dependent conditions. Let P_s, P_t denote the source and target prefix lengths and L_s, L_t their generated lengths. We interpolate only the source-generated mel and the generated portion of μ_s to L_t, leaving the target representation, both reference prefixes, and all discrete tokens unchanged. Interpolation uses frame-center coordinates without endpoint forcing. The resulting branch lengths are $P_s + L_t$ and $P_t + L_t$; after evaluating the full velocity fields, we remove each reference prefix and subtract the equal-length generated portions. Prefix noises are right-aligned slices of a shared Gaussian sequence of length $\max(P_s, P_t)$. Aligned-source and

edited segments are decoded with the same pre-trained vocoder and random state, without additional denoising or mel normalization.

Evaluation. We evaluate waveforms decoded from $\mathbf{x}^{\mathrm{src}}$ and $\mathbf{x}^{\mathrm{edit}}$. Let $\mathbf{x}^{\mathrm{src,unaligned}}$ denote the source mel before length interpolation; for F5-TTS it equals $\mathbf{x}^{\mathrm{src}}$. Target-emotion probability is computed with the pre-trained emotion2vec-plus-large[1] classifier by applying a softmax over Neutral, Happy, Angry, Sad, and Surprise. WER is computed against the synthesis transcript using Whisper-large-v3[2]. Text is lowercased, stripped of Unicode punctuation, whitespace-normalized, and segmented with Jieba[3] for Mandarin. All evaluation audio is converted to mono and resampled to 16 kHz without additional loudness normalization. Metrics are averaged over cases.

Probability and WER compare decoded $\mathbf{x}^{\mathrm{src}}$ and $\mathbf{x}^{\mathrm{edit}}$. Speaker similarity (S-SIM) instead uses $\mathbf{x}^{\mathrm{src,unaligned}}$ as the baseline to include the complete pipeline and compares each waveform with an emotion-matched reference:

$$\begin{aligned} S_{\mathrm{before}} &= \cos\big(E(\mathcal{V}(\mathbf{x}^{\mathrm{src,unaligned}})), E(R_s)\big)\,, \\ S_{\mathrm{after}} &= \cos\big(E(\mathcal{V}(\mathbf{x}^{\mathrm{edit}})), E(R_t)\big)\,, \qquad \Delta S = S_{\mathrm{after}} - S_{\mathrm{before}}. \end{aligned} \tag{12}$$

Here, E is an ECAPA-TDNN[4] speaker encoder and $\mathcal{V}$ is the model's waveform decoder. Because the speaker encoder can remain sensitive to emotion, emotion-matched references reduce the mismatch caused by comparing emotional speech only against a neutral reference. Accordingly, ΔS should be interpreted as reference-based speaker-similarity change rather than a direct measure of identity loss.

Table 5: Target-emotion probability before and after editing. Values are percentages and Δ is in percentage points (pp). CosyVoice uses the aligned-source baseline.

Model	Emotion	Before (%, ↑)	After (%, ↑)	Δ (pp, ↑)
F5-TTS	Happy	5.0456	64.8676	+59.8219
	Angry	19.9294	77.8099	+57.8805
	Sad	19.8356	60.8144	+40.9788
	Surprise	0.2028	39.6302	+39.4274
	Mean (80 cases)	11.2534	60.7805	+49.5272
CosyVoice 2	Happy	10.0841	70.5735	+60.4894
	Angry	20.0920	86.6830	+66.5910
	Sad	19.6899	54.9719	+35.2820
	Surprise	2.4802×10^{-5}	61.0794	+61.0794
	Mean (80 cases)	12.4665	68.3269	+55.8604
Both models	Mean (160 observations)	11.8599	64.5537	+52.6938

Table 6: Per-case-averaged WER before and after editing. Positive changes indicate degradation. CosyVoice uses the aligned-source baseline; its whole-pipeline comparison is reported in the text.

Model	Emotion	Before (%, ↓)	After (%, ↓)	Δ (pp, ↓)
F5-TTS	Happy	18.0595	15.3095	-2.7500
	Angry	20.3591	9.4702	-10.8889
	Sad	9.7123	8.9385	-0.7738
	Surprise	12.4702	16.5575	+4.0873
	Mean (80 cases)	15.1503	12.5689	-2.5813
CosyVoice 2	Happy	8.8095	13.8452	+5.0357
	Angry	20.6627	10.2619	-10.4008
	Sad	4.7500	9.7202	+4.9702
	Surprise	6.0813	8.8075	+2.7262
	Mean (80 cases)	10.0759	10.6587	+0.5828
Both models	Mean (160 observations)	12.6131	11.6138	-0.9993

Results. Tables 5, 6, and 7 provide the complete per-emotion results and model averages. Both models exhibit pronounced target-emotion probability gains: 49.53 pp for F5-TTS and 55.86 pp

[1] https://huggingface.co/emotion2vec/emotion2vec_plus_large
[2] https://huggingface.co/openai/whisper-large-v3
[3] https://github.com/fxsjy/jieba
[4] https://huggingface.co/speechbrain/spkrec-ecapa-voxceleb

Table 7: Emotion-matched S-SIM defined in Eq. 12. Before uses the neutral reference; after uses the target-emotion reference. CosyVoice before is the unedited generated speech before length alignment, so changes include alignment. Δ is an absolute cosine difference.

Model	Emotion	Before ↑	After ↑	Δ ↑
F5-TTS	Happy	0.6522	0.5966	-0.0556
	Angry	0.6629	0.5695	-0.0934
	Sad	0.6591	0.6186	-0.0405
	Surprise	0.6878	0.5499	-0.1379
	Mean (80 cases)	0.6655	0.5836	-0.0819
CosyVoice 2	Happy	0.6454	0.5716	-0.0738
	Angry	0.6501	0.5373	-0.1127
	Sad	0.6362	0.6444	+0.0082
	Surprise	0.6519	0.5589	-0.0930
	Mean (80 cases)	0.6459	0.5781	-0.0678
Both models	Mean (160 observations)	0.6557	0.5808	-0.0748

Table 8: Early-stop editing results. Editing begins at the first editing step and terminates after the indicated prefix. Results are averaged over 80 cases and three editing-noise repeats.

Model	Editing prefix	Emotion gain (pp)	Onset drift (ms)	ΔWER (pp)
F5-TTS	First 4 of 32 steps	0.015	0.38	+0.21
F5-TTS	All 32 steps	53.78	308.33	-0.92
CosyVoice 2	First 3 of 10 steps	3.43	22.21	+2.83
CosyVoice 2	First 6 of 10 steps	30.37	87.83	+88.88
CosyVoice 2	All 10 steps	56.09	245.42	+1.40

for CosyVoice 2. WER decreases by 2.58 pp for F5-TTS but increases by 0.58 pp for CosyVoice 2; the per-emotion results show that these averages mask heterogeneous changes across targets. Mean emotion-matched S-SIM decreases by 0.0819 for F5-TTS and 0.0678 for CosyVoice 2. Percentile 95% confidence intervals for these S-SIM differences are $[-0.1013, -0.0605]$ and $[-0.0833, -0.0520]$, obtained from 10,000 bootstrap samples of the 20 speaker clusters with seed 42. Overall, the results support substantial training-free emotional editability, with mixed effects on transcription accuracy and a moderate reduction in reference-based speaker similarity.

For completeness, CosyVoice WER is 7.7287% before alignment, 10.0759% after alignment, and 10.6587% after editing. The whole-pipeline increase is therefore 2.93 pp, including 2.35 pp from alignment. Its corresponding target-emotion probabilities are 3.6337%, 12.4665%, and 68.3269%. This distinction separates alignment effects from the editing effects reported in Tables 5 and 6.

B.2 Experiment Details for Q2

Experimental setup. We follow the diagnostic setup in Appendix B.1 unless otherwise stated. All trajectory interventions use the same 80 cases and model-specific inference settings. Each intervention is evaluated with three editing-noise repeats while keeping the source synthesis fixed. The reported t denotes the model's internal integration step; the native solver grids are nonuniform.

Trajectory-specific measures. We reuse the target-emotion probability change and ΔWER defined in Appendix B.1. For delayed-start experiments, we additionally report *retained emotion gain*, defined as the mean emotion gain relative to full-step editing. To quantify temporal disruption, we measure *onset drift* as the absolute difference between the detected source and edited speech onsets. Onset is detected using 20-ms RMS windows with a 10-ms hop and a threshold of 5% of peak RMS.

Delayed-start analysis. We first test whether the earliest editing updates are necessary. Specifically, we hold the editing state unchanged for the first k integration steps and apply the editing update at every remaining step. We sweep all possible starting steps on the native schedules of F5-TTS (32 steps) and CosyVoice 2 (10 steps), including full-step editing and no editing. Noise is shared across different start schedules at each corresponding integration step.

Table 9 reports representative operating points discussed in the main text, while Fig. 9 shows the complete scans. F5-TTS exhibits a short early interval in which onset drift can be substantially

Table 9: Selected delayed starts, averaged over cases and editing-noise repeats. Gain retention is relative to full-step editing; onset drift and ΔWER are relative to the decoded, length-aligned source.

Model	Editing start	Gain retained (%)	Onset drift (ms)	ΔWER (pp)
F5-TTS	Full-step	100.0	308.33	−0.92
	Skip 4 steps	99.5	97.33	−4.95
CosyVoice 2	Full-step	100.0	245.42	+1.40
	Skip 1 step	100.3	260.29	+0.67
	Skip 3 steps	51.0	44.25	+55.75

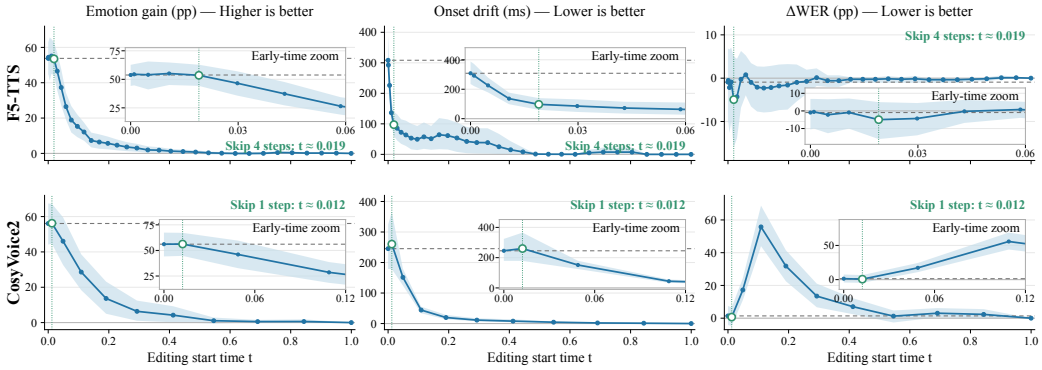


Figure 9: Complete delayed-start scans for F5-TTS and CosyVoice 2. Columns show emotion gain (pp), onset drift (ms), and ΔWER (pp), measured relative to the decoded, length-aligned source. Bands indicate pointwise 95% speaker-bootstrap intervals. The horizontal axis is the ODE step.

reduced while retaining all of the full-step emotion gain. CosyVoice 2 shows no similarly favorable delayed start: preserving the emotion gain provides little timing benefit, whereas stronger delay rapidly sacrifices editing effectiveness and linguistic fidelity.

Early-stop analysis. The delayed-start experiment tests whether early updates are *necessary*. Editing now begins at the first integration step and terminates after a selected prefix, after which the current state is frozen and decoded without subsequent editing updates. Table 10 reports representative results, and Fig. 10 shows the complete scan.

Early prefixes produce little emotion change compared with full-step editing. Moreover, intermediate states can have substantially worse transcription than the final state, showing that later updates may not only complete the emotion transfer but also recover intermediate content errors.

Table 10: Early-stop analysis. Editing starts at the first step and stops after the indicated prefix; values are means over 80 cases and three editing-noise repeats.

Model	Prefix	t	Emotion gain (pp)	Onset drift (ms)	ΔWER (pp)
F5-TTS	First 4 / 32	0.019	0.015	0.38	+0.21
F5-TTS	First 14 / 32	0.227	6.96	38.54	+1.31
F5-TTS	First 28 / 32	0.805	50.15	275.04	+6.03
F5-TTS	All 32 / 32	1.000	53.78	308.33	−0.92
CosyVoice 2	First 3 / 10	0.109	3.43	22.21	+2.83
CosyVoice 2	First 6 / 10	0.412	30.37	87.83	+88.88
CosyVoice 2	First 9 / 10	0.844	57.06	240.38	+1.84
CosyVoice 2	All 10 / 10	1.000	56.09	245.42	+1.40

Block-wise and strength analysis. The start- and stop-time scans show that editing effects vary strongly along the trajectory, but do not establish whether individual stages contribute independently. We therefore divide the trajectory into five stages and perform two complementary interventions. "Only block j" applies editing updates only within block j, whereas "skip block j" removes that block from an otherwise complete editing trajectory.

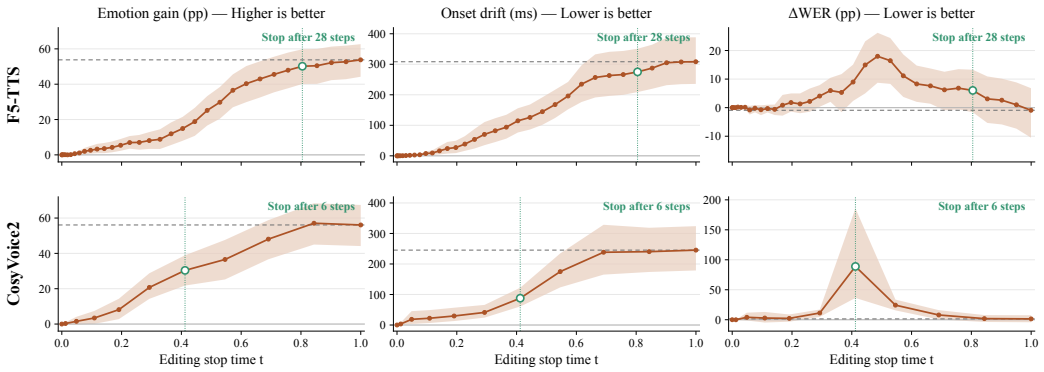


Figure 10: Early-stop scans for F5-TTS and CosyVoice 2. Columns report emotion gain, onset drift, and ΔWER. Bands show 95% speaker-bootstrap intervals; dashed lines indicate full-step means.

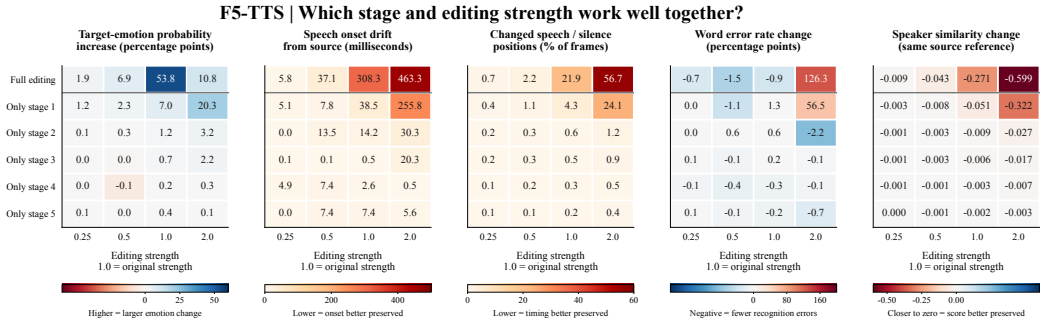


Figure 11: F5-TTS block and editing-strength ablations.

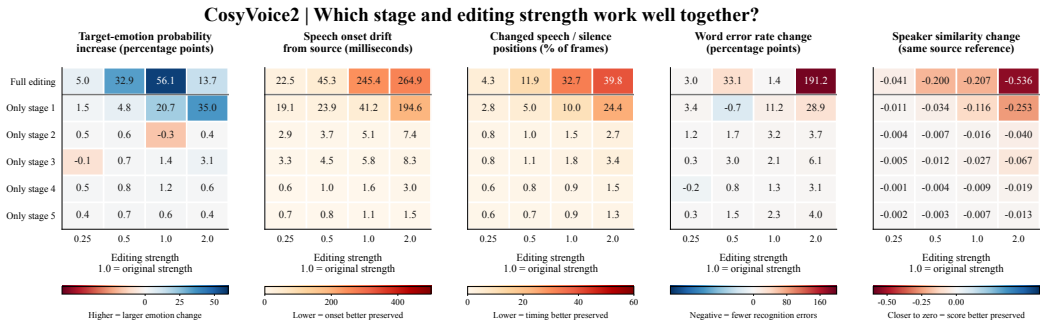


Figure 12: CosyVoice 2 block and editing-strength ablations.

Table 11: Block ablations. Values are emotion gain (pp), onset drift (ms), and ΔWER (pp).

Model	Block	Only block			Skip block		
		Gain	Drift	ΔWER	Gain	Drift	ΔWER
F5-TTS	1	6.96	38.54	1.31	4.66	53.71	−0.97
	2	1.17	14.17	0.60	36.00	205.38	13.88
	3	0.72	0.50	0.25	40.27	253.75	8.10
	4	0.23	2.62	−0.25	49.26	267.79	4.95
	5	0.36	7.38	−0.24	50.15	275.04	6.03
CosyVoice 2	1	20.73	41.21	11.23	6.37	11.25	13.49
	2	−0.27	5.12	3.20	55.67	243.58	3.36
	3	1.37	5.83	2.06	47.67	216.46	10.74
	4	1.24	1.58	1.35	56.85	238.75	1.39
	5	0.59	1.08	2.32	57.06	240.38	1.84

As shown in Table 11, no individual block reproduces the emotion gain of full-step editing. Conversely, removing a block from an otherwise complete trajectory can substantially change emotion gain, onset drift, or transcription accuracy. The contribution of a stage therefore depends on the state produced by preceding updates, supporting a state-dependent trajectory interpretation rather than an additive assignment of independent roles to individual stages.

We further examine how editing strength interacts with trajectory position. Figs. 11 and 12 (rows denote the full trajectory or individual blocks; columns denote editing strengths. Cell values are means with metric-specific color scales) show that increasing the strength does not consistently improve emotion transfer. In some stages, stronger updates even weaken the emotion change while substantially increasing onset drift or transcription errors. Here, stages are defined by solver-time boundaries and therefore contain different numbers of integration steps, while the strength directly scales the editing update at each step. These results show that the appropriate editing strength depends on where along the trajectory it is applied.

B.3 Experiment Details for Q3

Experimental Setup. We reuse the source cases, full-trajectory editor, and emotion evaluator from Appendix B.1. Each Neutral source is paired with the next speaker in dataset-ID order within the same language, with wraparound, requesting Happy, Angry, Sad, or Surprise editing targets (80 edits per model). Source and target references share the reference transcript; the synthesis text is unchanged. Editing uses seed 42. We keep the source mel spectrogram already length-aligned in Q1 fixed; the newly generated target mel segment is linearly interpolated to this fixed length and decoded for pitch comparison. CosyVoice 2 retains the complete target condition, including target-generated speech tokens and flow-conditioning features in the original paper.

Identity and pitch measures. For identity evaluation, each speaker has ten independent real enrollment recordings: two texts, each in five emotions, disjoint from all conditioning-reference and synthesis texts. We unit-normalize their ECAPA-TDNN embeddings, average them, and normalize the mean to obtain a speaker centroid. We measure identity changes relative to both the source and target centroids. For each centroid, we compute the cosine similarity before and after editing and report the edited-minus-source difference.

For pitch, we estimate F0 using the probabilistic YIN (pYIN) algorithm (Mauch & Dixon, 2014) from 16-kHz mono audio, with a 1024-sample window, a 160-sample hop, and a 50-600 Hz search range. Finite estimates are converted to semitones as $12 \log_2(F0/100\,\mathrm{Hz})$. For the median or 90th-minus-10th percentile range f, we report $|f(b) - f(t)| - |f(a) - f(t)|$, where b, a, and t are the aligned source, edited output, and aligned target synthesis. Positive values mean closer target pitch statistics. Both models have 79/80 valid pitch comparisons, covering all 20 source speakers.

Results. Table 1 reports the mean results across source speakers. We estimate uncertainty using 95% confidence intervals from 10,000 source-speaker bootstrap resamples (seed 42), conditional on the fixed speaker pairs and generation seed and without correction for multiple comparisons. Overall, speaker identity, emotion, and pitch all shift toward the target.

Table 12: SEmoEditBench task composition. Each task is evaluated under same-dataset same-speaker, same-dataset cross-speaker, and cross-dataset cross-speaker settings.

Task	Setting	Corpus	Cases	Construction
Replacement	Same dataset, same speaker	ESD, IEMOCAP	120	80 ESD and 40 IEMOCAP cases; sources and targets balanced across Neutral, Happy, Sad, Angry, and Surprise; Chinese and English in ESD.
Replacement	Same dataset, cross speaker	ESD, IEMOCAP	120	Same source cases as the same-speaker split, but the target reference comes from a different speaker of the same corpus.
Replacement	Cross dataset, cross speaker	ESD, IEMOCAP	80	Sources from one corpus and target references from the other; all English.
Erasure	Same dataset, same speaker	ESD, IEMOCAP	56	32 ESD and 24 IEMOCAP cases; source emotions Happy, Sad, Angry, and Surprise, all mapped to Neutral.
Erasure	Same dataset, cross speaker	ESD, IEMOCAP	56	Same source cases, target reference from a different speaker of the same corpus.
Erasure	Cross dataset, cross speaker	ESD, IEMOCAP	40	Sources from one corpus and target references from the other; all English.
Intensity	Same dataset, same speaker	RAVDESS, CREMA-D	44	20 RAVDESS and 24 CREMA-D cases; Neutral sources paired with same-speaker same-text emotional targets at five intensities.
Intensity	Same dataset, cross speaker	RAVDESS, CREMA-D	44	Same source cases, target reference from a different speaker of the same corpus.
Intensity	Cross dataset, cross speaker	RAVDESS, CREMA-D	40	Sources from one corpus and target references from the other; all English.

C Additional Details of SEmoEditBench

C.1 Task Construction

Table 12 summarizes the nine benchmark splits. The same-dataset same-speaker split uses a paired target recording from the same speaker reading the same text, isolating emotion from speaker and content. The same-dataset cross-speaker split keeps the source corpus but uses a different speaker as the target reference. The cross-dataset cross-speaker split is drawn from a corpus different from the source corpus, with a different speaker. Every system receives the source waveform, its ground-truth transcript, the requested target emotion (and target intensity for intensity control), and a target reference utterance with different linguistic content. The paired target waveform is reserved exclusively for evaluation and must not be used to generate the edit. Sampling is deterministic and speaker-aware, and source or paired-target waveforms are never reused as target references.

IEMOCAP sources are restricted to utterances between 2 and 10 seconds with categorical annotation agreement of at least 0.6. For ESD, IEMOCAP, and RAVDESS, the target reference matches the source speaker but uses different text. Because CREMA-D does not provide suitable same-speaker references for this protocol, its references are synthesized with IndexTTS2 (Zhou et al., 2026b): the neutral source provides speaker identity, a separate recording provides the requested emotion and intensity, and a synthesis transcript distinct from the source provides the reference content. The benchmark fixes all manifests, case identifiers, and references across evaluated systems. All intensity split summaries are averaged over the same 90-case common subset (Neutral source; Angry, Happy, Sad, or Surprise target; 30 cases per split) so that systems are directly comparable.

C.2 Metric Definitions

We evaluate each edit along three axes: editing success, preservation of non-target attributes, and perceptual quality. All objective metrics are computed per case and macro-averaged within each split using identical model configurations across systems; expected case count, failed case count, and failure rate are reported alongside the metric averages. Table 13 lists the symbols used throughout the metric definitions. Table 14 provides a compact overview of all metrics, their optimization directions, and task-level applicability; the subsections below define each metric in detail.

Table 13: Notation used in the metric definitions.

Symbol	Meaning
x_s, x_e, x_t	Source, edited, and paired-target waveforms, respectively
$e_{\text{src}}, e_{\text{tgt}}$	Source and target emotion labels
y	Manifest transcript (ground-truth text)
$P(e \mid x)$	emotion2vec+ Large posterior probability for emotion e given waveform x
$f_E(x)$	emotion2vec+ Large embedding of waveform x
$A(x)$	ASR transcript of waveform x produced by Whisper-large-v3
$f_S(x)$	ECAPA-TDNN speaker embedding of waveform x
$Q(x)$	UTMOSv2 quality prediction for waveform x
$x_e^{(\alpha)}$	Edited output at intensity strength α
$\alpha_1 < \cdots < \alpha_5$	Five intensity strengths $\{0, 0.25, 0.5, 0.75, 1\}$
z_s, z_t, z_i	emotion2vec embeddings $f_E(x_s)$, $f_E(x_t)$, $f_E(x_e^{(\alpha_i)})$
a_i	Relative progress of output i along the source–target direction: $(z_i - z_s)^\top (z_t - z_s)/\|z_t - z_s\|^2$

Table 14: Metrics used in SEmoEditBench. Upward and downward arrows indicate whether higher or lower values are preferred.

Metric	Definition	Applicability
Editing success		
Target emotion probability (TEP) ↑	$P(e_{\text{tgt}} \mid x_e)$	Replacement
Neutral probability (NP) ↑	$P(\text{Neutral} \mid x_e)$	Erasure
Source emotion suppression (SES) ↑	$P(e_{\text{src}} \mid x_s) - P(e_{\text{src}} \mid x_e)$	Replacement, erasure
Embedding-based effective intensity control (EIC-Emb) ↑	Mean over all pairs $i < j$ of $\text{sgn}(a_j - a_i) \cdot \min(1, \max(0, \lvert a_j - a_i \rvert / \tau)), \tau = 1$	Intensity control (paired target)
Emotion similarity (E-SIM) ↑	$\cos(f_E(x_e), f_E(x_t))$	Splits with paired targets
Directional editing score (DES) ↑	$\cos(f_E(x_e) - f_E(x_s), f_E(x_t) - f_E(x_s))$	Splits with paired targets
Preservation		
Relative word error rate (ΔWER) ↓	$\text{WER}(x_e) - \text{WER}(x_s)$ against the same transcript; effectively ΔCER for Chinese	All tasks
Speaker similarity (S-SIM) ↑	ECAPA-TDNN cosine verification score between x_s and x_e	All tasks
Quality		
UTMOS ↑	UTMOSv2 prediction for x_e	All tasks
Subjective evaluation		
Speaker-similarity MOS (SS-MOS) ↑	Perceived speaker similarity between the source and edited utterances	20 sampled case groups
Emotion similarity (ES-MOS) ↑	Perceived emotion similarity between the edited utterance and requested target emotion	20 sampled case groups
Naturalness MOS (N-MOS) ↑	Perceived naturalness of the edited utterance	20 sampled case groups

For DES, a zero edited displacement receives a score of zero, whereas a zero target displacement is invalid. E-SIM and DES are reported for all three replacement and erasure splits, each of which provides an evaluation-only paired target recording. EIC-Emb uses the corresponding paired target recordings for all three intensity splits, never the generation-time target reference.

C.2.1 Editing Success

Editing-success metrics measure whether the edited waveform expresses the requested target emotion (or, for erasure, neutral) while moving away from the source emotion. They are computed from emotion2vec+ Large posteriors and embeddings.

Target emotion probability (TEP). TEP measures the degree to which the edited waveform is recognized as the requested target emotion:

$$\text{TEP} = P(e_{\text{tgt}} \mid x_e). \tag{13}$$

Measurement. The emotion2vec+ Large model produces a posterior distribution over emotion categories from the edited waveform; the probability mass assigned to e_{tgt} is reported. *Data source.* Edited waveform x_e. *Evaluation criteria.* Higher values indicate stronger target-emotion expression; reported for the replacement task.

Neutral probability (NP). For emotion erasure, the requested output is emotionally neutral. NP measures the probability that the edited waveform is classified as neutral:

$$\text{NP} = P(\text{neutral} \mid x_e). \tag{14}$$

Measurement. Same as TEP but extracting the posterior mass for the neutral class. *Data source.* Edited waveform x_e. *Evaluation criteria.* Higher values indicate more complete erasure; reported for the erasure task.

Source emotion suppression (SES). SES quantifies the reduction of the source emotion from the source to the edited waveform:

$$\mathrm{SES} = P(e_{\mathrm{src}} \mid x_s) - P(e_{\mathrm{src}} \mid x_e). \tag{15}$$

Measurement. The emotion2vec posterior for the source emotion e_{src} is computed for both the source and edited waveforms, and their difference is taken. *Data sources.* Source waveform x_s and edited waveform x_e. *Evaluation criteria.* Higher values indicate stronger suppression of the source emotion; reported for both replacement and erasure.

Embedding-based effective intensity control (EIC-Emb). EIC-Emb is the primary metric used to evaluate intensity control in the main benchmark. For each case, the system produces five edited outputs $x_e^{(\alpha_i)}$ at increasing strengths $\alpha_1 < \cdots < \alpha_5$. Each output embedding is projected onto the source–target direction via the relative progress

$$a_i = \frac{(z_i - z_s)^\top (z_t - z_s)}{\|z_t - z_s\|^2}, \tag{16}$$

where $a_i = 0$ means no progress beyond the source, $a_i = 1$ reaches the target's projection, negative values move away from the target, and values above 1 overshoot. EIC-Emb then applies a pairwise effective-control transform: for every pair $i < j$, the signed progress difference $d = a_j - a_i$ is mapped to

$$\mathrm{sgn}(d) \cdot \min\left(1, \max\left(0, \frac{|d|}{\tau}\right)\right), \qquad \tau = 1, \tag{17}$$

and the mean over all $\binom{5}{2} = 10$ pairs is reported. *Measurement.* emotion2vec+ Large embeddings are computed for the source, the paired target, and all five edited outputs; the relative progress values a_i are derived and the pairwise formula above is applied. *Data sources.* Source waveform x_s, paired-target waveform x_t (evaluation-only), and the five edited outputs $x_e^{(\alpha_i)}$. *Evaluation criteria.* Higher values indicate more effective and monotonic control; a score of 1 means every stronger edit produces strictly more progress toward the target with differences of at least τ, 0 means no consistent increase, and negative values indicate decreasing progress. EIC-Emb is reported for the intensity-control task on cases that have a paired target recording. Because a_i is a signed projection rather than a cosine similarity, EIC-Emb preserves the magnitude of the emotional displacement and is anchored to the actual paired target recording.

Emotion similarity (E-SIM). E-SIM measures the cosine similarity between the edited waveform's embedding and the paired target's embedding:

$$\text{E-SIM} = \cos(f_E(x_e), f_E(x_t)), \qquad \cos(u, v) = \frac{u^\top v}{\|u\|_2 \|v\|_2}. \tag{18}$$

Measurement. emotion2vec+ Large embeddings are computed for both waveforms and their cosine similarity is reported. *Data sources.* Edited waveform x_e and paired-target waveform x_t. *Evaluation criteria.* Higher values indicate that the edit matches the emotional expression of the target; reported for splits with paired targets.

Directional editing score (DES). DES measures whether the edit moves in the same direction as the target relative to the source:

$$d_e = f_E(x_e) - f_E(x_s), \qquad d_t = f_E(x_t) - f_E(x_s), \qquad \mathrm{DES} = \cos(d_e, d_t). \tag{19}$$

A zero edited displacement $d_e = \mathbf{0}$ receives a score of zero, whereas a zero target displacement $d_t = \mathbf{0}$ is invalid. *Measurement.* emotion2vec+ Large embeddings are computed for the source, edited, and paired-target waveforms; the displacement vectors are formed and their cosine similarity is reported. *Data sources.* Source x_s, edited x_e, and paired-target x_t waveforms. *Evaluation criteria.* Higher values indicate that the edit travels along the correct source-to-target emotional direction; reported for splits with paired targets.

C.2.2 PRESERVATION

Preservation metrics verify that the edit modifies only the intended affective expression while keeping the linguistic content and speaker identity intact.

Relative word error rate (ΔWER). ΔWER measures the change in transcription error from the source to the edited waveform against the same ground-truth transcript y:

$$\Delta\mathrm{WER} = \mathrm{WER}\left(y, A(x_e)\right) - \mathrm{WER}\left(y, A(x_s)\right). \tag{20}$$

Measurement. Both the source and edited waveforms are transcribed by Whisper-large-v3; WER is computed against the manifest transcript y after text normalization. For Chinese, tokenization is character-level, making this effectively a character error rate (ΔCER). *Data sources.* Source waveform x_s, edited waveform x_e, and ground-truth transcript y. *Evaluation criteria.* Lower values indicate better content preservation; $\Delta\mathrm{WER} = 0$ means the edit introduced no transcription errors relative to the source. Reported for all tasks.

Speaker similarity (S-SIM). S-SIM measures the cosine similarity between the speaker embeddings of the source and edited waveforms:

$$\text{S-SIM} = \cos(f_S(x_s), f_S(x_e)). \tag{21}$$

Measurement. A pretrained ECAPA-TDNN speaker verifier extracts embeddings from both waveforms, and their cosine similarity is reported. *Data sources.* Source waveform x_s and edited waveform x_e. *Evaluation criteria.* Higher values indicate better preservation of the source speaker identity; reported for all tasks.

C.2.3 Quality

UTMOS. UTMOS estimates the perceptual naturalness of the edited waveform:

$$\mathrm{UTMOS} = Q(x_e). \tag{22}$$

Measurement. The pretrained UTMOSv2[5] predictor takes the edited waveform as input and outputs a quality score. *Data source.* Edited waveform x_e. *Evaluation criteria.* Higher values indicate better predicted audio quality; reported for all tasks.

C.2.4 Subjective Evaluation

In addition to objective metrics, we uniformly sample 20 case groups from the benchmark and collect human judgments on a five-point scale (1–5) for three attributes.

Speaker-similarity MOS (SS-MOS). *Definition.* Perceived speaker similarity between the source and edited utterances. *Measurement.* Raters compare the source and edited waveforms and score how well the speaker identity is preserved. *Evaluation criteria.* Higher scores indicate better speaker preservation.

Emotion similarity MOS (ES-MOS). *Definition.* Perceived emotion similarity between the edited utterance and the requested target emotion. *Measurement.* Raters are given the edited waveform together with the target emotion label (and the target reference when available) and score how closely the edit matches the intended emotion. *Evaluation criteria.* Higher scores indicate closer emotional match to the target.

Naturalness MOS (N-MOS). *Definition.* Perceived naturalness of the edited utterance. *Measurement.* Raters listen to the edited waveform and score its overall naturalness, including prosody, absence of artifacts, and fluency. *Evaluation criteria.* Higher scores indicate more natural-sounding speech.

D Experimental Results

D.1 Complete In- and Out-of-Distribution Results

Tables 15–17 report all objective metrics for ID-SS (same dataset and speaker), ID-CS (same dataset, cross speaker), and OOD (cross dataset and speaker). Unsupported methods are omitted, and intensity preservation and quality are measured at $\alpha = 1$. Rankings are computed independently within each setting. Pink, green, and blue cells denote ID-SS, ID-CS, and OOD, respectively; dark, medium, and light shades mark first, second, and third place, with displayed ties sharing a rank.

[5] https://github.com/sarulab-speech/UTMOSv2

Table 15: Complete emotion-replacement results under both in-distribution settings and the out-of-distribution setting.

Category	Method	Backbone	Setting	TEP ↑	SES ↑	E-SIM ↑	DES ↑	ΔWER ↓	S-SIM ↑	UTMOS ↑
Training-based	Step-Audio-EditX	–	ID-SS	0.222	0.445	0.521	0.550	**-0.048**	0.567	**3.091**
			ID-CS	0.196	0.420	0.519	0.543	-0.021	0.561	**3.053**
			OOD	0.211	0.452	0.526	0.575	-0.026	0.519	**3.142**
	Auk	–	ID-SS	0.259	0.372	0.539	0.541	0.073	0.631	2.639
			ID-CS	0.259	0.372	0.539	0.541	0.073	0.631	2.670
			OOD	**0.312**	0.353	**0.586**	0.580	0.103	0.580	2.753
	dots.tts.edit	–	ID-SS	0.434	0.600	0.669	0.713	**-0.035**	0.272	2.381
			ID-CS	**0.434**	0.600	**0.669**	**0.713**	-0.035	0.272	2.394
			OOD	**0.357**	0.483	**0.638**	**0.691**	**-0.051**	0.215	2.512
Activation-steering	CoCoEmo	CosyVoice 2	ID-SS	0.082	0.192	0.446	0.399	-0.015	**0.715**	**3.061**
			ID-CS	0.133	0.189	0.447	0.364	-0.014	**0.717**	**3.046**
			OOD	0.136	0.217	0.431	0.470	-0.039	**0.667**	**3.169**
	CoCoEmo	IndexTTS2	ID-SS	0.035	0.127	0.402	0.269	-0.024	**0.776**	2.659
			ID-CS	0.034	0.101	0.396	0.298	-0.019	**0.782**	2.658
			OOD	0.046	0.125	0.398	0.334	-0.024	**0.735**	2.723
	EmoSteer-TTS	F5-TTS	ID-SS	0.081	0.345	0.448	0.477	0.608	0.623	2.545
			ID-CS	0.081	0.345	0.448	0.477	0.608	0.623	2.507
			OOD	0.117	0.449	0.475	0.562	0.904	0.550	2.503
	EmoSteer-TTS	CosyVoice 2	ID-SS	0.030	0.371	0.382	0.367	-0.015	0.478	2.515
			ID-CS	0.030	0.371	0.382	0.367	-0.015	0.478	2.523
			OOD	0.042	0.396	0.384	0.460	-0.029	0.425	2.530
	EmoSteer-TTS	IndexTTS2	ID-SS	0.025	0.054	0.377	0.226	**-0.037**	**0.788**	2.666
			ID-CS	0.025	0.054	0.377	0.226	**-0.037**	**0.788**	2.687
			OOD	0.037	0.069	0.361	0.240	**-0.060**	**0.758**	2.707
SEmoEdit	**Ours**	F5-TTS	ID-SS	**0.498**	**0.684**	**0.704**	**0.753**	-0.026	0.372	2.177
			ID-CS	0.432	**0.692**	0.646	0.700	**-0.049**	0.434	2.494
			OOD	0.283	**0.544**	0.563	**0.629**	**-0.057**	0.395	2.466
	Ours	CosyVoice 2	ID-SS	**0.554**	**0.694**	**0.776**	**0.806**	-0.025	0.379	**2.984**
			ID-CS	**0.460**	**0.683**	**0.713**	**0.752**	**-0.045**	0.438	**2.925**
			OOD	0.268	**0.539**	0.550	**0.640**	-0.047	0.398	**3.206**
	Ours	IndexTTS2	ID-SS	**0.691**	**0.767**	**0.902**	**0.920**	-0.004	0.423	2.578
			ID-CS	**0.516**	**0.671**	**0.768**	**0.775**	0.000	0.468	2.686
			OOD	**0.297**	**0.547**	**0.576**	0.585	-0.046	0.436	2.930

Table 16: Complete emotion-erasure results under both in-distribution settings and the out-of-distribution setting. CoCoEmo is omitted because it does not support emotion erasure.

Category	Method	Backbone	Setting	NP ↑	SES ↑	E-SIM ↑	DES ↑	ΔWER ↓	S-SIM ↑	UTMOS ↑
Training-based	Step-Audio-EditX	–	ID-SS	0.235	0.388	0.566	0.655	**-0.064**	0.577	**3.220**
			ID-CS	0.347	0.488	0.606	0.688	-0.028	0.586	**3.133**
			OOD	0.269	0.465	**0.746**	**0.818**	**-0.070**	**0.561**	**3.197**
	Auk	–	ID-SS	0.229	0.311	0.528	0.582	0.014	**0.676**	**2.859**
			ID-CS	0.229	0.311	0.528	0.582	0.014	**0.676**	**2.881**
			OOD	0.272	0.371	0.591	0.585	0.033	**0.651**	2.959
	dots.tts.edit	–	ID-SS	0.355	0.549	0.674	0.766	**-0.065**	0.304	2.685
			ID-CS	0.355	0.549	0.674	0.766	**-0.065**	0.304	2.675
			OOD	**0.350**	**0.501**	0.745	0.790	**-0.070**	0.254	2.752
Activation-steering	EmoSteer-TTS	F5-TTS	ID-SS	0.202	0.393	0.616	0.736	0.030	**0.598**	2.593
			ID-CS	0.202	0.393	0.616	0.736	0.030	**0.598**	2.575
			OOD	0.258	0.451	0.731	**0.812**	0.030	0.549	2.643
	EmoSteer-TTS	CosyVoice 2	ID-SS	0.078	0.174	0.505	0.596	-0.012	0.548	2.490
			ID-CS	0.078	0.174	0.505	0.596	-0.012	0.548	2.491
			OOD	0.084	0.167	0.605	0.682	-0.007	0.510	2.519
	EmoSteer-TTS	IndexTTS2	ID-SS	0.038	0.070	0.377	0.298	-0.041	**0.808**	2.836
			ID-CS	0.038	0.070	0.377	0.298	-0.041	**0.808**	2.845
			OOD	0.053	0.099	0.452	0.317	-0.045	**0.798**	2.853
SEmoEdit	**Ours**	F5-TTS	ID-SS	**0.573**	**0.700**	**0.821**	**0.873**	**-0.058**	0.359	2.597
			ID-CS	**0.592**	**0.670**	**0.802**	**0.857**	**-0.081**	0.409	2.706
			OOD	**0.408**	**0.557**	**0.758**	**0.820**	**-0.098**	0.399	2.665
	Ours	CosyVoice 2	ID-SS	**0.704**	**0.754**	**0.907**	**0.938**	-0.050	0.399	**3.144**
			ID-CS	**0.580**	**0.687**	**0.839**	**0.854**	**-0.047**	0.443	**3.166**
			OOD	**0.418**	**0.532**	**0.772**	0.777	**-0.078**	0.431	**3.303**
	Ours	IndexTTS2	ID-SS	**0.682**	**0.741**	**0.910**	**0.934**	-0.030	0.424	2.774
			ID-CS	**0.576**	**0.634**	**0.798**	**0.806**	-0.032	0.470	2.783
			OOD	0.284	0.385	0.692	0.729	-0.064	0.460	**3.050**

Table 17: Complete emotion-intensity-control results under both in-distribution settings and the out-of-distribution setting. Training-based baselines are omitted because they do not support continuous intensity control.

Category	Method	Backbone	Setting	EIC-Emb ↑	ΔWER$_{\alpha=1}$ ↓	S-SIM$_{\alpha=1}$ ↑	UTMOS$_{\alpha=1}$ ↑
Activation-steering	CoCoEmo	CosyVoice 2	ID-SS	0.041	**0.044**	**0.621**	**3.234**
			ID-CS	0.023	**0.006**	**0.648**	**3.206**
			OOD	0.057	**0.044**	**0.595**	**2.990**
	CoCoEmo	IndexTTS2	ID-SS	0.017	0.333	**0.703**	2.294
			ID-CS	0.007	0.378	**0.699**	2.344
			OOD	0.026	0.356	**0.688**	2.314
	EmoSteer-TTS	F5-TTS	ID-SS	0.013	0.139	0.465	2.309
			ID-CS	0.013	0.139	0.465	2.250
			OOD	0.013	0.139	0.465	2.243
	EmoSteer-TTS	CosyVoice 2	ID-SS	-0.009	0.133	0.362	**2.575**
			ID-CS	-0.009	0.133	0.362	**2.655**
			OOD	-0.009	0.133	0.362	2.542
	EmoSteer-TTS	IndexTTS2	ID-SS	-0.005	0.400	**0.683**	2.355
			ID-CS	-0.005	0.400	**0.683**	2.335
			OOD	-0.005	0.400	**0.683**	2.313
SEmoEdit	**Ours**	F5-TTS	ID-SS	**0.175**	**0.039**	0.307	1.967
			ID-CS	**0.146**	**0.011**	0.342	2.094
			OOD	**0.114**	**0.011**	0.271	2.083
	Ours	CosyVoice 2	ID-SS	**0.172**	**0.017**	0.362	**2.499**
			ID-CS	**0.138**	**0.000**	0.371	**2.700**
			OOD	**0.099**	**0.006**	0.309	**2.870**
	Ours	IndexTTS2	ID-SS	**0.192**	0.411	0.336	2.355
			ID-CS	**0.206**	0.400	0.344	2.289
			OOD	**0.190**	0.428	0.282	**2.687**